INTUITIVE BIOSTATISTICS

Harvey Motulsky, M.D.

Department of Pharmacology
University of California, San Diego
and
President, GraphPad Software, Inc.

New York *Oxford*
OXFORD UNIVERSITY PRESS
1995

CONTENTS IN BRIEF

1

Introduction to Statistics

There is something fascinating about science. One gets such a wholesale return of conjecture out of a trifling investment of fact.
Mark Twain (Life on the Mississippi, 1850)

This is a book for "consumers" of statistics. The goals are to teach you enough statistics to

1. Understand the statistical portions of most articles in medical journals.
2. Avoid being bamboozled by statistical nonsense.
3. Do simple statistical calculations yourself, especially those that help you interpret published literature.
4. Use a simple statistics computer program to analyze data.
5. Be able to refer to a more advanced statistics text or communicate with a statistical consultant (without an interpreter).

Many statistical books read like cookbooks; they contain the recipes for many statistical tests, and their goal (often unstated) is to train "statistical chefs" able to whip up a P value on moment's notice. This book is based on the assumption that statistical tests are best calculated by computer programs or by experts. This book, therefore, will not teach you to be a chef, but rather to become an educated connoisseur or critic who can appreciate and criticize what the chef has created. But just as you must learn a bit about the differences between broiling, boiling, baking, and basting to become a connoisseur of fine food, you must learn a bit about probability distributions and null hypotheses to become an educated consumer of the biomedical literature. Hopefully this book will make it relatively painless.

WHY DO WE NEED STATISTICAL CALCULATIONS?

When analyzing data, your goal is simple: You wish to make the strongest possible conclusions from limited amounts of data. To do this, you need to overcome two problems:

- Important differences are often obscured by biological variability and/or experimental imprecision, making it difficult to distinguish real differences from random variation.
- The human brain excels at finding patterns and relationships, but tends to overgeneralize. For example, a 3-year-old girl recently told her buddy, "You can't become a

doctor; only girls can become doctors.'' To her this made sense, as the only three doctors she knew were women. This inclination to overgeneralize does not seem to go away as you get older, and scientists have the same urge. Statistical rigor prevents you from making this kind of error.

MANY KINDS OF DATA CAN BE ANALYZED WITHOUT STATISTICAL ANALYSIS

Statistical calculations are most helpful when you are looking for fairly small differences in the face of considerable biological variability and imprecise measurements. Basic scientists asking fundamental questions can often reduce biological variability by using inbred animals or cloned cells in controlled environments. Even so, there will still be scatter among replicate data points. If you only care about differences that are large compared with the scatter, the conclusions from such studies can be obvious without statistical analysis. In such experimental systems, effects small enough to require statistical analysis are often not interesting enough to pursue.

If you are lucky enough to be studying such a system, you may heed the following aphorisms:

> If you need statistics to analyze your experiment, then you've done the wrong experiment.
>
> If your data speak for themselves, don't interrupt!

Most scientists are not so lucky. In many areas of biology, and especially in clinical research, the investigator is faced with enormous biological variability, is not able to control all relevant variables, and is interested in small effects (say 20% change). With such data, it is difficult to distinguish the signal you are looking for from the noise created by biological variability and imprecise measurements. Statistical calculations are necessary to make sense out of such data.

STATISTICAL CALCULATIONS EXTRAPOLATE FROM SAMPLE TO POPULATION

Statistical calculations allow you to make general conclusions from limited amounts of data. You can extrapolate from your data to a more general case. Statisticians say that you extrapolate from a *sample* to a *population.* The distinction between sample and population is key to understanding much of statistics. Here are four different contexts where the terms are used.

- *Quality control.* The terms *sample* and *population* make the most sense in the context of quality control where the sample is randomly selected from the overall population. For example, a factory makes lots of items (the population), but randomly selects a few items to test (the sample). These results obtained from the sample are used to make inferences about the entire population.
- *Political polls.* A random sample of voters (the sample) is polled, and the results are used to make conclusions about the entire population of voters.

- *Clinical studies.* The sample of patients studied is rarely a random sample of the larger population. However, the patients included in the study are representative of other similar patients, and the extrapolation from sample to population is still useful. There is often room for disagreement about the precise definition of the population. Is the *population* all such patients that come to that particular medical center, or all that come to a big city teaching hospital, or all such patients in the country, or all such patients in the world? While the population may be defined rather vaguely, it still is clear we wish to use the sample data to make conclusions about a larger group.
- *Laboratory experiments.* Extending the terms *sample* and *population* to laboratory experiments is a bit awkward. The data from the experiment(s) you actually performed is the sample. If you were to repeat the experiment, you'd have a different sample. The data from all the experiments you could have performed is the population. From the sample data you want to make inferences about the ideal situation.

In biomedical research, we usually assume that the population is infinite, or at least very large compared with our sample. All the methods in this book are based on that assumption. If the population has a defined size, and you have sampled a substantial fraction of the population (>10% or so), then you need to use special methods that are not presented in this book.

WHAT STATISTICAL CALCULATIONS CAN DO

Statistical reasoning uses three general approaches:

Statistical Estimation

The simplest example is calculating the mean of a sample. Although the calculation is exact, the mean you calculate from a sample is only an estimate of the population mean. This is called a *point estimate.* How good is the estimate? As we will see in Chapter 5, it depends on the sample size and scatter. Statistical calculations combine these to generate an interval estimate (a range of values), known as a *confidence interval* for the population mean. If you assume that your sample is randomly selected from (or at least representative of) the entire population, then you can be 95% sure that the mean of the population lies somewhere within the 95% confidence interval, and you can be 99% sure that the mean lies within the 99% confidence interval. Similarly, it is possible to calculate confidence intervals for proportions, for the difference or ratio of two proportions or two means, and for many other values.

Statistical Hypothesis Testing

Statistical hypothesis testing helps you decide whether an observed difference is likely to be caused by chance. Various techniques can be used to answer this question: If there is no difference between two (or more) populations, what is the probability of randomly selecting samples with a difference as large or larger than actually observed? The answer is a probability termed the *P value.* If the P value is small, you conclude that the difference is statistically *significant* and unlikely to be due to chance.

Statistical Modeling

Statistical modeling tests how well experimental data fit a mathematical model constructed from physical, chemical, genetic, or physiological principles. The most common form of statistical modeling is linear regression. These calculations determine "the best" straight line through a particular set of data points. More sophisticated modeling methods can fit curves through data points.

WHAT STATISTICAL CALCULATIONS CANNOT DO

In theory, here is how you should apply statistical analysis to a simple experiment:

1. Define a population you are interested in.
2. Randomly select a sample of subjects to study.
3. Randomly select half the subjects to receive one treatment, and give the other half another treatment.
4. Measure a single variable in each subject.
5. From the data you have measured in the samples, use statistical techniques to make inferences about the distribution of the variable in the population and about the effect of the treatment.

When applying statistical analysis to real data, scientists confront several problems that limit the validity of statistical reasoning. For example, consider how you would design a study to test whether a new drug is effective in treating patients infected with the human immunodeficiency virus (HIV).

- The population you really care about is all patients in the world, now and in the future, who are infected with HIV. Because you can't access that population, you choose to study a more limited population: HIV patients aged 20 to 40 living in San Francisco who come to a university clinic. You may also exclude from the population patients who are too sick, who are taking other experimental drugs, who have taken experimental vaccines, or who are unable to cooperate with the experimental protocol. Even though the population you are working with is defined narrowly, you hope to extrapolate your findings to the wider population of HIV-infected patients.
- Randomly sampling patients from the defined population is not practical, so instead you simply attempt to enroll all patients who come to morning clinic during two particular months. This is termed a *convenience sample.* The validity of statistical calculations depends on the assumption that the results obtained from this convenience sample are similar to those you would have obtained had you randomly sampled subjects from the population.
- The variable you really want to measure is survival time, so you can ask whether the drug increases life span. But HIV kills slowly, so it will take a long time to accumulate enough data. As an alternative (or first step), you choose to measure the number of helper (CD4) lymphocytes. Patients infected with the HIV have low numbers of CD4 lymphocytes, so you can ask whether the drug increases CD4 cell number (or delays the reduction in CD4 cell count). To save time and expense, you have switched from an important variable (survival) to a proxy variable (CD4 cell count).

- Statistical calculations are based on the assumption that the measurements are made correctly. In our HIV example, statistical calculations would not be helpful if the antibody used to identify CD4 cells was not really selective for those cells.
- Statistical calculations are most often used to analyze one variable measured in a single experiment, or a series of similar experiments. But scientists usually draw general conclusions by combining evidence generated by different kinds of experiments. To assess the effectiveness of a drug to combat HIV, you might want to look at several measures of effectiveness: reduction in CD4 cell count, prolongation of life, increased quality of life, and reduction in medical costs. In addition to measuring how well the drug works, you also want to quantify the number and severity of side effects. Although your conclusion must be based on all these data, statistical methods are not very helpful in blending different kinds of data. You must use clinical or scientific judgment, as well as common sense.

In summary, statistical reasoning can not help you overcome these common problems:

- The population you really care about is more diverse than the population from which your data were sampled.
- You collect data from a "convenience sample" rather than a random sample.
- The measured variable is a proxy for another variable you really care about.
- Your measurements may be made or recorded incorrectly, and assays may not always measure exactly the right thing.
- You need to combine different kinds of measurements to reach an overall conclusion.

You must use scientific and clinical judgment, common sense, and sometimes a leap of faith to overcome these problems. Statistical calculations are an important part of data analysis, but interpreting data also requires a great deal of judgment. That's what makes research challenging. This is a book about statistics, so we will focus on the statistical analysis of data. Understanding the statistical calculations is only a small part of evaluating clinical and biological research.

WHY IS IT HARD TO LEARN STATISTICS?

Five factors make it difficult for many students to learn statistics:

- The terminology is deceptive. Statistics gives special meaning to many ordinary words. To understand statistics, you have to understand that the statistical meaning of terms such as *significant, error,* and *hypothesis* are distinct from the ordinary uses of these words. As you read this book, pay special attention to the statistical terms that sound like words you already know.
- Many people seem to believe that statistical calculations are magical and can reach conclusions that are much stronger than is actually possible. The phrase *statistically significant* is seductive and is often misinterpreted.
- Statistics requires mastering abstract concepts. It is not easy to think about theoretical concepts such as populations, probability distributions, and null hypotheses.
- Statistics is at the interface of mathematics and science. To really grasp the concepts of statistics, you need to be able to think about it from both angles. This book

emphasizes the scientific angle and avoids math. If you think like a mathematician, you may prefer a text that uses a mathematical approach.

- The derivation of many statistical tests involves difficult math. Unless you study more advanced books, you must take much of statistics on faith. However, you can learn to *use* statistical tests and interpret the results even if you don't fully understand how they work. This situation is common in science, as few scientists really understand all the tools they use. You can interpret results from a pH meter (measures acidity) or a scintillation counter (measures radioactivity), even if you don't understand *exactly* how they work. You only need to know enough about how the instruments work so that you can avoid using them in inappropriate situations. Similarly, you can calculate statistical tests and interpret the results even if you don't understand how the equations were derived, as long as you know enough to use the statistical tests appropriately.

ARRANGEMENT OF THIS BOOK

Parts I through V present the basic principles of statistics. To make it easier to learn, I have separated the chapters that explain confidence intervals from those that explain P values. In practice, the two approaches are used in parallel. Basic scientists who don't care to learn about clinical studies may skip Chapters 6 (survival curves) and 9 (case-control studies) without loss of continuity.

Part VI describes the design of clinical studies and discusses how to determine sample size. Basic scientists who don't care to learn about clinical studies can skip this entire part. However, Chapter 22 (sample size) is of interest to all. Part VII explains the most common statistical tests. Even if you use a computer program to calculate the tests, reading these chapters will help you understand how the tests work. The tests mentioned in this section are described in detail.

Part VIII gives an overview of more advanced statistical tests. These tests are not described in detail, but the chapters provide enough information so that you can be an intelligent consumer of papers that use these tests. The chapters in this section do not follow a logical sequence, so you can pick and choose the topics that interest you. The only exception is that you should read Chapter 31 (multiple regression) before Chapters 32 (logistic regression) or the parts of Chapter 33 (comparing survival curves) dealing with proportional hazards regression.

The statistical principles and tests discussed in this book are widely used, and I do not give detailed references. For more information, refer to the general textbook references listed in Appendix 1.

1

CONFIDENCE INTERVALS

Statistical analysis of data leads to two kinds of results: confidence intervals and P values. The two give complementary information and are often calculated in tandem. For the purposes of clarity and simplicity, this book presents confidence intervals first and then presents P values. Confidence intervals let you state a result with *margin of error.* This section explains what this means and how to calculate confidence intervals.

2

Confidence Interval of a Proportion

PROPORTIONS VERSUS MEASUREMENTS

The results of experiments can be expressed in different ways. In this chapter we will consider only results expressed as a proportion or fraction. Here are some examples: the proportion of patients who become infected after a procedure, the proportion of patients with myocardial infarction who develop heart failure, the proportion of students who pass a course, the proportion of voters who vote for a particular candidate. Later we will discuss other kinds of variables, including measurements and survival times.

THE BINOMIAL DISTRIBUTION: FROM POPULATION TO SAMPLE

If you flip a coin fairly, there is a 50% probability (or chance) that it will land on heads and a 50% probability that it will land on tails. This means that, in the long run, a coin will land on heads about as often as it lands on tails. But in any particular series of tosses, you may not see the coin land on heads exactly half the time. You may even see all heads or all tails.

Mathematicians have developed equations, known as the *binomial distribution,* to calculate the likelihood of observing any particular outcome when you know the proportion in the overall population. Using the binomial distribution, you can answer questions such as these:

- If you flip a coin 10 times, what is the probability of getting exactly 7 heads?
- If you flip a coin 10 times, what is the probability of getting 7 or more heads?
- If 5% of patients undergoing an operation get infected, what is the chance that 10 or more of the next 30 patients will be infected?
- If a couple's chance of passing a genetic disease to each child is 25%, what is the chance that their first three children will all be unaffected?
- If 40% of voters are Democrats, what is the chance that a random sample of 500 voters will include more than 45% Democrats?

Perhaps you've seen the equations that help you answer these kinds of questions, and recall that there are lots of factorials. If you're interested, the equation is presented at the end of this chapter.

The binomial distribution is not immediately useful when analyzing data because it works in the wrong direction. The theory starts with a known probability (i.e., 50% of coin flips are heads) and calculates the likelihood of any particular result in a sample. When analyzing data, we need to work in the opposite direction. We don't know the overall probability. That's what we are trying to find out. We do know the proportion observed in a single sample and wish to make inferences about the overall probability.

The binomial distribution can still be useful, but it must be turned backwards to generate confidence intervals. I show you how to do this at the end of the chapter. For now, accept the fact that it can be done and concentrate on interpreting the results.

THE CONFIDENCE INTERVAL OF A PROPORTION: FROM SAMPLE TO POPULATION

Let's start with an example. Out of 14 patients you have treated with a particular drug, three suffered from a particular side effect. The proportion is 3/14, which equals 0.2143. What can you say about the probability of complications in the entire population of patients who will be treated with this drug?

There are two issues to think about. First, you must think about whether the 14 patients are representative of the entire population of patients who will receive the drug. Perhaps these patients were selected in such a way as to make them more (or less) likely than other patients to develop the side effect. Statistical calculations can't help you answer that question, and we'll assume that the sample adequately represents the population. The second issue is random sampling, sometimes referred to as *margin of error.* Just by chance, your sample of 14 patients may have had an especially high or an especially low rate of side effects. The overall proportion of side effects in the population is unlikely to equal exactly 0.2143.

Here is a second example. You polled 100 randomly selected voters just before an election, and only 33 said they would vote for your candidate. What can you say about the proportion of *all* voters who will vote for your candidate? Again, there are two issues to deal with. First, you need to think about whether your sample is really representative of the population of voters, and whether people tell the pollsters the truth about how they will vote. Statistical calculations cannot help you grapple with those issues. We'll assume that the sample is perfectly representative of the population of voters and that every person will vote as they said they would on the poll. Second, you need to think about sampling error. Just by chance, your sample may contain a smaller or larger fraction of people voting for your candidate than does the overall population.

Since we only know the proportion in one sample, there is no way to be sure about the proportion in the population. The best we can do is calculate a range of values that bracket the true population proportion. How wide does this range of values have to be? In the overall population, the fraction of patients with side effects could be as low as 0.000001% (or lower) or as high as 99.99999% (or higher). Those values are exceedingly unlikely but not absolutely impossible. If you want to be 100% sure that your range includes the true population value, the range has to include these possibilities. Such a wide range is not helpful. To create a narrower and more useful range, you must accept the possibility that the interval will not include the true population value.

Scientists usually accept a 5% chance that the range will not include the true population value. The range or interval is called the *95% confidence interval,* abbreviated *95% CI.* You can be 95% sure that the 95% CI includes the true population value. It makes sense that the margin of error depends on the sample size, so that the confidence interval is wider in the first example (14 subjects) than in the second (100 subjects). Before continuing, you should think about these two examples and write down your intuitive estimate of the 95% CIs. Do it now, before reading the answer in the next paragraph.

Later in this chapter you'll learn how to calculate the confidence interval. But it is easier to use an appropriate computer program to calculate the 95% CIs instantly. All examples in this book were calculated with the simple program GraphPad InStat (see Appendix 2), but many other programs can perform these calculations. Here are the results. For the first example, the 95% CI extends from 0.05 to 0.51. For the second example, the 95% CI extends from 0.24 to 0.42. How good were your guesses? Many people tend to imagine that the interval is narrower than it actually is.

What does this mean? Assuming that our samples were randomly chosen from the entire populations, we can be 95% sure that the range of values includes the true population proportion. Note that there is no uncertainty about what we observed in the sample. We are absolutely sure that 21.4% of our subjects suffered from the side effect and that 33.0% of the people polled said they would vote for our candidate. Calculation of a confidence interval cannot overcome any mistakes that were made in tabulating those numbers. What we don't know is the proportion in the entire population. However, we can be 95% sure that it lies within the calculated interval.

The term *confidence interval,* abbreviated CI, refers to the range of values. The correct syntax is to express the CI as 5% to 51%, as 0.05 to 0.51, or as [0.05,0.51]. It is considered to be bad form to express the CI as 5%–51% or as 28% ± 23%. The two ends of the CI are called the *confidence limits.*

WHAT EXACTLY DOES IT MEAN TO SAY THAT YOU ARE "95% SURE"?

When you only have measured one sample, you don't know the value of the population proportion. It either lies within the 95% CI you calculated or it doesn't. There is no way for you to know. If you were to calculate a 95% CI from many samples, the population proportion will be included in the CI in 95% of the samples, but will be outside of the CI the other 5% of the time. More precisely, you can be 95% certain that the 95% CI calculated from your sample includes the population proportion.

Figure 2.1 illustrates the meaning of CIs. Here we assume that the proportion of voters in the overall population who will vote for your candidate equals 0.28 (shown as the horizontal dotted line). We created 50 samples, each with 100 subjects and calculated the 95% CI for each. Each 95% CI is shown as a vertical line extending from the lower confidence limit to the upper confidence limit. The value of the observed proportion in each sample is shown as a small hatch mark in the middle of each CI. The first line (on the left) corresponds to our example. The other 49 lines represent results that could have been obtained by random sampling from the same population.

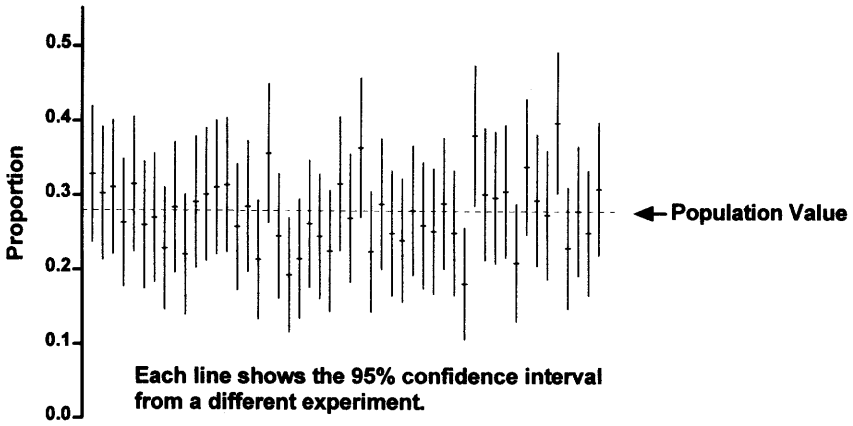


Figure 2.1. The meaning of a confidence interval. In this example, we know that the true proportion of "success" in the population equals 0.28. This is shown as a horizontal dotted line. Using a computer program that can generate random numbers, we randomly selected 50 samples from this population. Each vertical line shows the 95% CI for the proportion "success" calculated from that one sample. With most of the samples, the observed proportion is close to the true proportion, but in some samples there is a big discrepancy. In four of the samples (for example, the fifth from the right), the 95% CI does not include the true value. If you were to collect data from many samples, you'd expect to see this 5% of the time.

In most of the samples, the population value is near the middle of the 95% CI. In some of the samples, the population value is near one of the ends of the 95% CI. And in four of the samples, the population value lies outside the 95% CI. In the long run, 1 out of 20 95% CIs will not include the population value.

Figure 2.1 is useful for understanding CIs, but you cannot create such a figure when you analyze data. When you analyze data, you don't know the actual population value, as you only have results from a single sample. There is no way you can know whether the 95% CI you calculate includes the population value. All you know is that, in the long run, 95% of such intervals will contain the population value and 5% will not. Of course, every CI you calculate will contain the sample proportion you obtained. What you can't know for sure is whether the interval also contains the population proportion.

Note that the CI is not always symmetrical around the sample proportion. In the first example, it extends further to the right than the left. With larger sample sizes, the 95% CI becomes narrower and more symmetrical.

ASSUMPTIONS

The interpretation of the CI depends on the following assumptions:

- Random (or representative) sample
- Independent observations
- Accurate assessment
- Assessing an event you really care about

Random (or Representative) Sample

The 95% CI is based on the assumption that your sample was randomly selected from the population. In many cases, this assumption is not true. You can still interpret the CI as long as you assume that your sample is representative of the population.

This assumption would be violated if the 14 patients included in the sample were the first 14 patients to be given the drug. If so, they are likely to be sicker than future patients and perhaps more likely to get a side effect. The assumption would be violated in the election example if the sample was not randomly selected from the population of voters. In fact, this mistake was made in the Roosevelt-Landon U.S. Presidential Election. To find subjects, the pollsters relied heavily on phone books and automobile registration lists. In 1936, Republicans were far more likely than Democrats to own a phone or a car, and therefore the poll selected too many Republicans. The poll predicted that Landon would win by a large margin, but that didn't happen.

The reputable polling organizations no longer make this type of mistake and go to a lot of trouble to ensure that their samples are representative of the entire population. However, many so-called polls are performed in which television viewers are invited to call in their opinion. In the United States, this usually is a 900 phone number for which the caller pays for the privilege of being polled! Clearly, the self-selected "sample" tabulated by such "polls" is not representative of any population, so the data mean nothing. For example, in June 1994 the football star O.J. Simpson was arrested for allegedly murdering his ex-wife, and the events surrounding the arrest were given a tremendous amount of television coverage. One news show performed a telephone "poll" of its viewers asking whether the press was giving too much coverage. Clearly the results are meaningless, as people who really thought that there was too much coverage probably weren't watching the show.

Independent Observations

The 95% CI is only valid when all subjects are sampled from the same population and each has been selected independently of the others. Selecting one member of the population should not change the chance of selecting anyone else. This assumption would be violated if some patients were given the drug twice and included in the sample more than once (maybe we only had 12 patients but two were double counted). The assumption would also be violated if several of the patients came from one family. The propensity for some drug reactions is familial, so observations from two patients from one family are not independent observations. In the election sample, the assumption would be violated if the pollsters polled both husband and wife in each family, or if some voters were polled more than once or if half the subjects were sampled from one city and half from another.

Accurate Assessment

The 95% CI is only valid when the number of subjects in each category is tabulated correctly. This assumption would be violated in our first example if one of the patients actually had taken a different drug, or if one of the "drug side effects" was actually caused by something else. The assumption would be violated in the election example if the pollster recorded some of the opinions incorrectly.

Assessing an Event You Really Care About

The 95% CI allows you to extrapolate from the sample to the population for the event that you tabulated. But sometimes you really care about a different event. In our drug reaction example, our interpretation of the results must depend on the severity of the drug reactions. If we included all possible reactions, no matter how mild, the results won't tell us what we really care about: the proportion of patients who develop severe (life-threatening) drug reactions.

In the voting example, we assessed our sample's response on a poll on a particular date, so the 95% CI gives us the margin of error for how the population would respond on that poll on that date. We wish to extrapolate to election results in the future but can do so only by making an additional assumption—that people will vote as they said they would. This assumption was violated in a classic mistake in the polling prior to the Dewey versus Truman Presidential Election in the United States in 1948. Polls of many thousand voters showed that Dewey would win by a large margin. Because the CI was so narrow, the pollsters were very confident. Newspapers were so sure of the results that they prematurely printed the erroneous headline "Dewey Beats Truman." In fact, Truman won. Why was the poll wrong? The polls were performed in September and early October, and the election was held in November. Many voters changed their mind in the interim period. The 95% CI correctly presented the margin of error in September but was inappropriately used to predict voting results 2 months later.

OBTAINING THE CONFIDENCE INTERVAL OF A PROPORTION FROM A TABLE

The width of the CI depends on both the value of the proportion and the size of the sample. If you make the sample larger, the CI becomes narrower. If the sample proportion is close to 50%, the CI is wider than if the proportion is far from 50%.

The CI is determined from the binomial distribution using calculations described later in the chapter. The answers have been tabulated in Table 2.1 shown here, which is abridged from Table A5.1 in the Appendix. Find the column corresponding to the numerator and the row corresponding to the denominator, and read the 95% CI.

Consider another example. You have performed a procedure 15 times with a single adverse incident. Without any other knowledge of the procedure, what is the

Table 2.1. Confidence Interval of a Proportion

	Numerator				
Denominator	0	1	2	3	4
10	0.00 to 0.31	<0.01 to 0.45	0.03 to 0.56	0.07 to 0.65	0.12 to 0.74
11	0.00 to 0.28	<0.01 to 0.41	0.02 to 0.52	0.06 to 0.61	0.11 to 0.69
12	0.00 to 0.26	<0.01 to 0.38	0.02 to 0.48	0.05 to 0.57	0.10 to 0.65
13	0.00 to 0.25	<0.01 to 0.36	0.02 to 0.45	0.05 to 0.54	0.09 to 0.61
14	0.00 to 0.23	<0.01 to 0.34	0.02 to 0.43	0.05 to 0.51	0.08 to 0.58
15	0.00 to 0.22	<0.01 to 0.32	0.02 to 0.40	0.04 to 0.48	0.08 to 0.55

95% CI for the average rate of adverse incidents? The answer is the 95% CI of the observed proportion 1/15, which extends from <0.01 to 0.32. With 95% confidence, the true proportion of complications may be less than 1% or as high as 32%. Most people are surprised by how wide the CIs are.

THE SPECIAL CASES OF 0 AND 100 PERCENT*

The 95% CI allows for a 2.5% chance that the population proportion is higher than the upper confidence limit and a 2.5% chance that the population proportion is lower than the lower confidence limit. If you observed 0 successes out of N trials, then you know that there is no possibility that the population proportion is less than the observed proportion. You only need an upper confidence limit, as the lower limit must be exactly 0. A similar problem occurs when you observed N successes out of N trials. You only need a lower confidence limit, as the upper limit must be 100%. Because the uncertainty only goes in one direction, the "95%" confidence interval really gives you 97.5% confidence.

EXAMPLE

In order to better counsel the parents of premature babies, M.C. Allen et al. investigated the survival of premature infants.† They retrospectively studied all premature babies born at 22 to 25 weeks gestation at the Johns Hopkins Hospital during a 3-year period. The investigators separately tabulated deaths for infants by their gestational age. Of 29 infants born at 22 weeks gestation, none survived 6 months. Of 39 infants born at 25 weeks gestation, 31 survived for at least 6 months.

The investigators presented these data without CI, but you can calculate them. It only makes sense to calculate a CI when the sample is representative of a larger population about which you wish to make inferences. It is reasonable to think that these data from several years at one hospital are representative of data from other years at other hospitals, at least at big-city university hospitals in the United States. If you aren't willing to make that assumption, you shouldn't calculate a CI. But the data wouldn't be worth collecting if the investigators didn't think that the results would be similar in other hospitals in later years.

For the infants born at 25 weeks gestation, we want to determine the 95% CI of 31/39. These values are not on Table A5.1, so you'll need to calculate the CI by computer or by hand. If you use the InStat program, you'll find that the 95% CI ranges from 63% to 91%. (If you use Equation 2.1, which follows, you'll calculate an approximate interval: 67% to 92%). This means that if the true proportion of surviving infants was any less than 63%, there is less than a 2.5% chance of observing such a large proportion just by chance. It also means that if the true proportion were any greater than 91%, the chance observing such a small proportion just by chance is less

*This section is more advanced than the rest. You may skip it without loss of continuity.

†MC Allen, PK Donohue, AE Dusman. The limit of viability—Neonatal outcome of infants born at 22 to 25 weeks gestation. *N Engl J Med* 329:1597–1601, 1993.

than 2.5%. That leaves us with a 95% chance (100% − 2.5% − 2.5%) that the true proportion is between 63% and 91%.

For the infants born at 22 weeks gestation, we want to determine the CI of 0/29. You shouldn't use Equation 2.2, because the numerator is too small. Instead use a computer program, or Table A5.1 in the Appendix. The 95% CI extends from 0% to 11.9%. We can be 95% sure that the overall proportion of surviving infants is somewhere in this range. (Since we observed 0%, the CI really only goes in one direction, so we really can be 97.5% sure rather than 95% sure.) Even though no babies born at 22 weeks gestational age survived in our sample, our CI includes the possibility that the overall survival rate is as high as 11.9%. This means that if the overall survival rate in the population was any value greater than 11.9%, there would be less than a 2.5% chance of observing 0 survivors in a sample of 29.

These CIs only account for sampling variability. When you try to extrapolate these results to results you expect to see in your hospital, you also need to account for the different populations served by different hospitals and the different methods used to care for premature infants. The true CIs, which are impossible to calculate, are almost certainly wider than the ones you calculate.

CALCULATING THE CONFIDENCE INTERVAL OF A PROPORTION*

The calculations used to determine the exact CIs shown in the table are quite complex and should be left to computer programs. But it is easy to calculate an approximate CI as long as the numbers are not too small.

Equation 2.1 is a reasonable approximation for calculating the 95% CI of a proportion p assessed in a sample with N subjects. The confidence limits this equation calculates are not completely accurate, but it is a reasonable approximation if at least five subjects had each outcome (in other words, the numerator of the proportion is 5 or greater and the denominator is at least 5 greater than the numerator).

Approximate 95% CI of proportion:

$$\left(p - 1.96\sqrt{\frac{p(1-p)}{N}}\right) \text{ to } \left(p + 1.96\sqrt{\frac{p(1-p)}{N}}\right) \qquad (2.1)$$

Beware of the variable p. The p used here is not the same as a P value (which we will discuss extensively in later chapters). The use of the letter p for both purposes is potentially confusing. This book uses an upper case P for P values and a lower case p for proportions, but not all books follow this convention.

The approximation can be used with the election example given at the beginning of the chapter. The sample proportion is 33/100 or 33.0%. The number of subjects

*This section contains the equations you need to calculate statistics yourself. You may skip it without loss of continuity.

with each outcome is greater than 5 (33 said they would vote for one candidate, and 67 said they would vote for the other), so the approximation can be used. The 95% CI ranges from 0.24 to 0.42. Because the sample is large, these values match the values determined by the exact methods to two decimal places. With smaller samples, this approximate method is less accurate.

You may calculate CIs for any degree of confidence. By convention, 95% CIs are presented most commonly. If you want to be more confident that your interval contains the population value, you must make the interval wider. If you are willing to be less confident, then the interval can be narrower. To generate a 90% CI, substitute the number 1.65 for 1.96 in Equation 2.1. To generate a 99% CI, substitute 2.58. You'll learn where these numbers come from in Chapter 4.

THE BINOMIAL EQUATION*

Assume that in the overall population, the proportion of "successes" is p. In a sample of N subjects, what is the chance that observing exactly R successes? The answer is calculated by Equation 2.2:

$$\text{Probability of R successes in N trials} = \left(\frac{N!}{R!(N-R)!}\right) p^R (1-p)^{N-R}. \quad (2.2)$$

The exclamation point denotes factorial. For example, $3! = 3 \times 2 \times 1 = 6$. The term on the right $[p^R(1-p)^{N-R}]$ is the probability of obtaining a particular sequence of "successes" and "failures." That term is very small. The term on the left takes into account that there are many different sequences of successes and failures that lead to the same proportion success.

Equation 2.2 calculates the probability of observing exactly R successes in N trials. Most likely, you really want to know the chance of observing *at least* S successes in N trials. To do this, reevaluate Equation 2.2 for $R = S$, $R = S + 1$, $R = S + 2\ldots$ up to $R = N$ and sum all the resulting probabilities.

The word *success* is used quite generally to denote one of the possible outcomes. You could just as well use the word *failure* or *outcome A*. Equation 2.2 uses p and $(1 - p)$ symmetrically, so it doesn't matter which outcome you label *success* and which outcome you label *failure.*

HOW THE CONFIDENCE INTERVALS ARE DERIVED

You can calculate and interpret CIs without knowing how they are derived. But it is nice to have a feel for what's really going on. I'll try to give the flavor of the logic using our second example. Recall that we observed that 33 out of 100 subjects polled said they would vote a certain way. The 95% CI is 0.24 to 0.42.

*This section contains the equations you need to calculate statistics yourself. You may skip it without loss of continuity.

The binomial equation works from population to sample. It lets us answer this kind of question. Given a hypothetical proportion in the population, what is the chance of observing some particular proportion (or higher) in a sample of defined size? The equations are a bit messy, but it is not surprising that probability theory can answer those kinds of questions.

To find the upper 95% confidence limit (U), we need to pose this question: If the population proportion equals U, what is the probability that a sample proportion (N = 100) will equal 0.33 or less? We set the answer to 2.5% and then solve for U. To find the lower 95% confidence limit (L), we pose the opposite question. If the population proportion equals L, what is the probability that a sample proportion (N = 100) will equal 0.33 or more? We set the answer to 2.5% and then solve for L. Solving those equations for U and L is not easy. Fortunately, the answers have been tabulated: L = 0.24 and U = 0.42. We've set up the problem to allow for a 2.5% chance of being wrong in each direction. Subtracting these from 100% leaves you with 95% confidence that the population proportion is between L and U, between 0.24 and 0.42.

So the trick is to use mathematical theory that makes predictions about the samples from hypothetical populations, and twist it around so that it can make inferences about the population from the sample. It's tricky to understand and it's amazing that it works. Fortunately you can calculate statistical analyses and interpret statistical results without having a solid understanding of the mathematical theory.

SUMMARY

Many types of data can be reduced to either/or categories and can be expressed as a proportion of events or subjects that fall into each category. The distribution of such events is described by the binomial distribution.

The challenge of statistics is to start with an observation in one sample and make generalizations about the overall population. One way to do this is to express the results as a confidence interval. Having observed a proportion in your sample, you can calculate a range of values that you can be 95% sure contains the true population proportion. This is the 95% confidence interval. You can calculate the interval for any degree of confidence you want, but 95% is standard. If you want to be more confident, you must make the interval wider.

OBJECTIVES

1. You must be familiar with the following terms:
   - Binomial distribution
   - Cumulative binomal distribution
   - Confidence interval
   - Confidence limit
   - Sample
   - Population
   - Random sample
   - Independence

2. When you see a proportion reported in a scientific paper (or elsewhere), you should be able to
   - Define the population.
   - Calculate (or determine from a table) the 95% CI of the proportion.
   - Interpret the 95% CI in the context of the study.
   - State the assumptions that must be true for the interpretation to be valid.

PROBLEMS

1. Of the first 100 patients to undergo a new operation, 6 die. Can you calculate the 95% CI for the probability of dying with this procedure? If so, calculate the interval. If not, what information do you need? What assumptions must you make?
2. A new drug is tested in 100 patients and lowers blood pressure by an average of 6%. Can you calculate the 95% CI for the fractional lowering of blood pressure by this drug? If so, calculate the interval. If not, what information do you need? What is the CI for the fractional lowering of blood pressure? What assumptions must you make?
3. You use a hemocytometer to determine the viability of cells stained with Trypan Blue. You count 94 unstained cells (viable) and 6 stained cells (indicating that they are not viable). Can you calculate the 95% CI for the fraction of stained (dead) cells? If so, calculate the interval. If not, what information do you need? What assumptions must you make?
4. In 1989, 20 out of 125 second-year medical students in San Diego failed to pass the biostatistics course (until they came back for an oral exam). Can you calculate the 95% CI for the probability of passing the course? If so, calculate the interval. If not, what information do you need? What assumptions must you make?
5. Ross Perot won 19% of the vote in the 1992 Presidential Election in the United States. Can you calculate the 95% CI for the fraction of voters who voted for him? If so, calculate the interval. If not, what information do you need? What assumptions must you make?
6. In your city (population = 1 million) last year a rare disease had an incidence of 25/10,000 population. Can you calculate the 95% CI for the incidence rate? If so, calculate the interval. If not, what information do you need? What assumptions must you make?
7. Is it possible for the lower end of a CI for a proportion to be negative? To be zero?
8. Is it possible to calculate a 100% CI?

3

The Standard Deviation

The previous chapter dealt with data that can be expressed as a fraction or proportion. However, the results of many experiments are expressed as measurements, for example, blood pressure, enzyme activity, IQ score, blood hemoglobin, or oxygen saturation. Working with these kinds of variables is more difficult than proportions, because you must deal with variability within samples as well as differences between groups.

SOURCE OF VARIABILITY

When you measure a variable in a number of subjects, the results will rarely be the same in all subjects. Scatter comes from three sources:

- *Imprecision or experimental error.* Many statistics books (especially those designed for engineers) implicitly assume that most variability is due to imprecision. In medical studies, imprecision is often a small source of variability.
- *Biological variability.* People (and animals, even cells) are different from one another, and these differences are important! Moreover, people (and animals) vary over time due to circadian variations, aging, and alterations in activity, mood, and diet. In biological and clinical studies, much or most of the scatter is often due to biological variation.
- *Blunders.* Mistakes and glitches (lousy pipetting, mislabeled tubes, transposed digits, voltage spikes, etc.) also contribute to variability.

Many statisticians use the word *error* when referring to all of these sources of variability. Note that this use of the word *error* is quite different from the everyday use of the word to mean *mistake.* I prefer the term *scatter* to *error.*

Another term you should know is *bias.* Biased measurements are systematically wrong. Bias can be caused by any factor that consistently alters the results: the proverbial thumb on the scale, miscalibrated pipettes, bugs in computer programs, the placebo effect, and so on. As used in statistics, the word *bias* refers to anything that leads to systematic errors, not only the preconceived notions of the experimenter. Bias will not usually contribute to scatter, as it will increase (or decrease) all values. Statistical calculations cannot detect or correct bias (unless you do additional experiments), and statistical analysis of biased data will rarely be helpful. As computer programmers say: Garbage in, garbage out.

DISPLAYING VARIABILITY WITH HISTOGRAMS

To discuss the distribution of data, we will work with an artificial example. One hundred medical students worked as pairs, and each student measured the systolic blood pressure of his or her partner. It is difficult to make much sense from a list of 100 numbers, so instead we display the results on a histogram as in Figure 3.1. A histogram is a bar graph. Systolic blood pressure is shown on the horizontal axis, measured in mmHg (same as torr). The base of each bar spans a range of blood pressure values, and the height of each bar is the number of individuals who have blood pressures that fall within that range.

To create a histogram you must decide how wide to make each bar. To make Figure 3.1, we set the width of each bar to 5 mmHg. The tallest bar in the graph shows that 15 subjects had blood pressure values between 117.5 and 122.5 mmHg. If the bars are too wide, the histogram will show too little detail, as shown in the left panel of Figure 3.2. If the bars are too narrow, the histogram will show too much detail and it is hard to interpret (right half of Figure 3.2). Some histograms plot relative frequencies (proportions or percentages) on the Y axis rather than the actual number of observations.

Before continuing with our analysis of the numbers, we should think a bit about what the numbers really tell us (or what they would tell us if they weren't made up). Blood pressure is affected by many variables. In serious studies of blood pressure, investigators go to a great deal of trouble to make sure that the measurements are as consistent and unbiased as possible.

This study of students has a number of flaws. Most important,

- The recorded values may differ substantially from arterial pressure. Every measurement was made by a different inexperienced person. Measuring blood pressure with

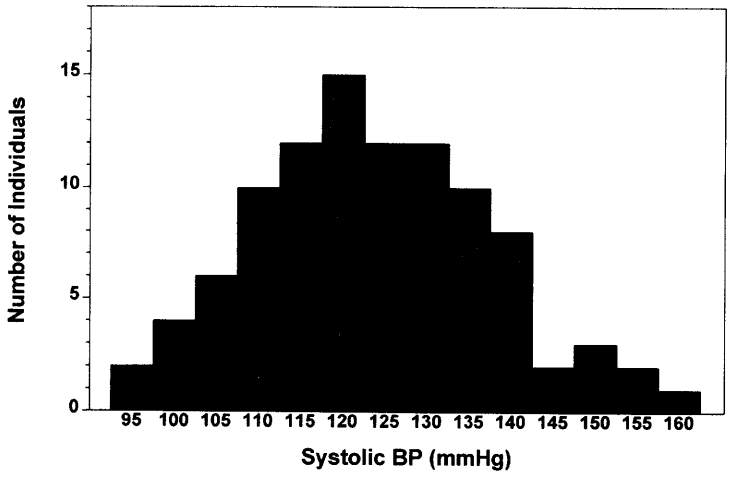


Figure 3.1. Histogram of blood pressures of 100 students. Each bar spans a width of 5 mmHg on the horizontal axis. The height of each bar represents the number of students whose blood pressure lies within that range.

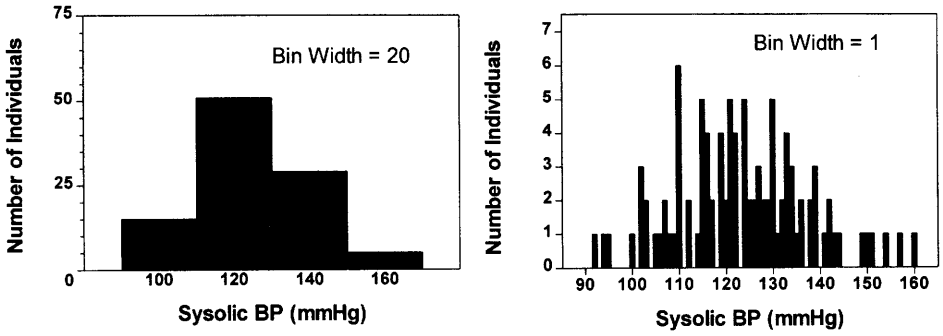


Figure 3.2. More histograms of blood pressure of 100 students. In the left panel, each bar spans a width of 20 mmHg. Because there are so few bars, the graph is not very informative. In the right panel, each bar has a width of 1 mmHg. Because there are so many bars, the graph contains too much detail, and it is hard to see the distribution of values. Histograms are generally most useful when they contain approximately 10 to 20 bars.

a cuff is somewhat subjective, as it requires noting the position of a bouncing column of mercury at the instant that a faint sound appears (systolic) or disappears (diastolic). It is also important to place the cuff correctly (at the level of the heart). Experience is required.

- The values cannot be compared to values for resting supine blood pressure measured in other studies. Having a pressure measured by a friend for the first time is stressful and may elevate the pressure.

I won't belabor the flaws of this example, except to repeat: Garbage in, garbage out! Statistical analysis of flawed data is unlikely to be useful. Designing good studies is difficult—much harder than statistical analysis.

THE MEAN AND MEDIAN

One way to describe the distribution of the data is to calculate the mean or average. You probably already know how to do that: Add up all the numbers and divide by the number of observations. For these 100 blood pressures, the mean is 123.4 mmHg. If you think visually, you should think of the mean as the "center of gravity" of the histogram. If you printed the bars of the histogram of Figure 3.1 on a sheet of wood or plastic, it would balance on a fulcrum placed at $X = 123.4$, the mean.

Another way to describe the central tendency of the data is to calculate the median. To do this you must rank the values in order and find the middle one. Since there are an even number of data points in this sample, the median lies between the 50th and 51st ranked value. The 50th ranked value is 122 and the 51st ranked value is 124. We take the average of those two values to obtain the median, which is 123 mmHg. You are probably already familiar with the term *percentile.* The median is the value at the 50th percentile because 50% of the values fall below the median.

In this example, the mean and median are almost the same. This is not always true. What would happen if the largest value (160 mmHg) were mistakenly entered

into the computer as 16,000 mmHg? The median would be unchanged at 123 mmHG. Changing that one value does not change the value of the 50th and 51st ranked values, so does not change the median. The mean, however, increases to 281 mmHg. In this case, there would be one value greater than the mean and 99 values below the mean. There are always just as many values below the median as above it. The number of data points above and below the mean will not be equal if the histogram is not symmetrical.

QUANTIFYING VARIABILITY WITH PERCENTILES

In addition to quantifying the middle of the distribution with the mean or median, you also want to quantify the scatter of the distribution. One approach is to report the lowest and highest values. For this example the data range from 92 to 160 mmHg. Another approach is to report the 25th and 75th percentiles. For this example, the 25th percentile is 115 mmHg. This means that 25% of the data points fall below 115. The 75th percentile is 133 mmHg. This means that 75% of the values lie below 133 mmHg. The difference between the 75th and 25th percentiles is called the *interquartile range,* which equals 18 mmHg.

The interquartile range can be depicted graphically as a box and whisker plot, as shown in Figure 3.3. The box extends from the 25th to 75th percentiles with a horizontal line at the median (50th percentile). The whiskers indicate the smallest and largest values. (Whiskers can be defined in other ways as well.)

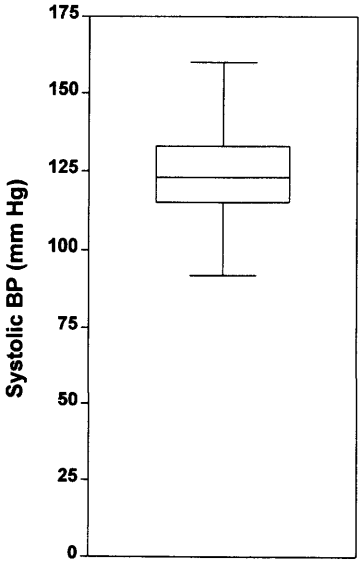


Figure 3.3. A box-and-whiskers plot. The horizontal line in the middle shows the median (50th percentile) of the sample. The top and bottom of the box show the 75th and 25th percentiles, respectively. In this graph, the top and bottom of the whiskers show the maximum and minimum values (other box-and-whiskers graphs define the whiskers in different ways). From this graph you can instantly see that the median is about 125 mmHg, and half the sample have systolic blood pressures between about 115 and 135 mmHg.

QUANTIFYING VARIABILITY WITH THE VARIANCE AND STANDARD DEVIATION

One problem with the interquartile range is that it tells you nothing about the values that lie below the 25th percentile or above the 75th percentile. We'd like a measure of scatter that somehow includes all the values. One way to quantify the scatter of all the values is to calculate the average of the deviations of the values from the mean. But it turns out that this is not helpful—the positive deviations balance out the negative deviations so the average deviation is always 0. Another thought would be to take the average of the absolute values of the deviations. This seems reasonable, but it turns out to be pretty much a statistical dead end that does not facilitate other inferences. Instead statisticians quantify scatter as the average of the square of the deviations of each value from the mean. This value is termed the *variance.* The variance of a population (abbreviated σ^2) is defined by Equation 3.1, where μ is the population mean, N is the number of data points, and Y_i is an individual value.

$$\text{Population variance} = \sigma^2 = \frac{\sum_{i=1}^{i=N}(Y_i - \mu)^2}{N}. \tag{3.1}$$

You may wonder why the deviations are squared. Why not just sum the absolute values of the deviations? This is difficult to explain intuitively. It just falls out of the math and makes statistical equations work. Basically, the idea is that one large deviation contributes more scatter to the data than do two deviations half that size. More practically, defining the variance in this way facilitates calculations of confidence intervals (CIs) and P values, as you will see in later chapters.

The variance is expressed in the units of the data squared. Thinking about the square of units is difficult, and so the variance is not an intuitive way to quantify scatter. The square root of the variance is called the *standard deviation,* abbreviated SD or σ (Equation 3.2):

$$\text{Population SD} = \sigma = \sqrt{\frac{\sum_{i=1}^{i=N}(Y_i - \mu)^2}{N}} \tag{3.2}$$

Caution! Don't use this equation until you read the next section.

The SD is expressed in the same units as the data. This makes it much easier to think about than the variance. Even so, you may not find it easy to think about the square root of the average of the squares of the deviations! You'll find it easier to think about SDs after you read the next chapter. We will discuss the interpretation of the SD later in this chapter and in the next chapter.

Returning to our blood pressure example, we encounter a problem when we try to use Equations 3.1 and 3.2 to figure out the variance and SD. The problem is that those equations apply only to the entire population, but we know only about a sample of 100 subjects. To use Equations 3.1 or 3.2, you need to know the value of μ, the population mean, but you don't know it. Read on to see how the equations need to be adjusted to deal with samples.

N OR N − 1? THE SAMPLE SD VERSUS THE POPULATION SD

Equations 3.1 and 3.2 assume that you have made measurements on the entire population. As we have already discussed, this is rarely the case. The whole point of statistical calculations is to make inferences about the entire population from measurements of a sample. This introduces an additional complexity to the calculation of the variance and SD. To calculate the SD using Equation 3.2, you need to calculate the deviation of each value from the population mean. But you don't know the population mean. All you know is the sample mean. Except for the rare cases where the sample mean happens to equal the population mean, values are always closer (on average) to their sample mean than to the overall population mean. The sum of the squares of the deviations from the sample mean is therefore smaller than the sum of squares of the deviations from the population mean, and Equation 3.2 gives too small a value for the SD. This problem is eliminated by reducing the denominator to N−1, rather than N. Calculate the variance and SD from a sample of data using Equation 3.3:

$$\text{Sample variance} = s^2 = \frac{\sum_{i=1}^{N}(Y_i - m)^2}{N-1}. \qquad (3.3)$$

$$\text{Sample SD} = s = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - m)^2}{N-1}}.$$

Note that we switched from Greek (μ, σ) to Roman (m, s) letters when switching from discussions of population mean and SD to discussions of sample mean and SD. This book hides a lot of the mathematical detail and sometimes glosses over the distinction. If you refer to more mathematical books, pay attention to the difference between Greek and Roman letters.

> If the difference between N and N−1 ever matters to you, then you are probably up to no good anyway—e.g. trying to substantiate a questionable hypothesis with marginal data.
>
> W.H. Press et al., *Numerical Recipes*

Here is another way to understand why the denominator is N−1 rather than N. When we calculate the sample mean m, we take the sum of all Y values and divide by the number of values N. Why divide by N? You learned to calculate a mean so long ago that you probably never thought about it. The mean is technically defined as the sum divided by degrees of freedom. The sample mean has N degrees of freedom because each of the N observations is free to assume any value. Knowing some of the values does not tell you anything about the remaining values. The sample variance is the mean of the square of the deviations of the values from the sample mean. This mean has only N−1 degrees of freedom. Why? It is because you must calculate the sample mean m before you can calculate the sample variance and SD. Once you know the sample mean and N−1 values, you can calculate the value of the remaining (Nth) value with certainty. The Nth value is absolutely determined from the sample mean and the other N−1 values. Only N−1 of the values are free to assume any value.

Therefore, we calculate the average of the squared deviations by dividing by N−1, and say that the sample variance has N−1 *degrees of freedom,* abbreviated df.

We'll discuss degrees of freedom again in later chapters. Many people find the concept of degrees of freedom to be quite confusing. Fortunately, being confused about degrees of freedom is not a big handicap! You can calculate statistical tests and interpret the results with only a vague understanding of degrees of freedom.

Returning to the blood pressure example, the sample SD is 14.0 mmHg. We'll discuss how to interpret this value soon. The sample variance is 196.8 $mmHg^2$. It is rare to see variances reported in the biomedical literature* but common to see SDs.

The term *sample SD* is a bit misleading. The sample SD is the best possible estimate of the SD of the entire population, as determined from one particular sample. In clinical and experimental science, one should routinely calculate the sample SD.

CALCULATING THE SAMPLE STANDARD DEVIATION WITH A CALCULATOR

Many calculators can calculate the SD. You have to first press a button to place the calculator into statistics or SD mode. Then enter each value and press the button labeled "enter" or "M+." After you have entered all values, press the appropriate button to calculate the SD. As we have seen, there are two ways to calculate the SD. Some calculators can compute either the sample or the population SD, depending on which button you press. The button for a sample SD is often labeled σ_{N-1}. Other calculators (especially those designed for use in business rather than science), compute only the population SD (denominator = N) rather than the sample SD (denominator = N−1).

To test your calculator, here is a simple example. Calculate the sample SD of these values: 120, 80, 90, 110, and 95. The mean is 99.0, and the sample SD is 16.0. If your calculator reports that the SD equals 14.3, then it is attempting to calculate the population SD. If you haven't collected data for the entire population, this is an invalid calculation that underestimates the SD. If your calculator computes the SD incorrectly, you may correct it with Equation 3.4:

$$\text{Sample SD} = \text{Invalid ``population'' SD} \cdot \sqrt{\frac{N}{N-1}} \qquad (3.4)$$

INTERPRETING THE STANDARD DEVIATION

We will discuss the interpretation of the SD in the next chapter. For now, you can use the following rule of thumb: Somewhat more than half of the observations in a population usually lie within 1 SD on either side of the mean. With our blood pressure example, we can say that most likely somewhat more than half of the observations will be within 14 mmHg of the mean of 123 mmHg. In other words, most of the

*However, it is common to see results analyzed with a method known as analysis of variance (ANOVA), which is explained in Chapter 30.

observations probably lie between 109 and 137 mmHg. This definition is unsatisfying because the words *somewhat* and *usually* are so vague. We'll interpret the SD more rigorously in the next chapter.

COEFFICIENT OF VARIATION (CV)*

The SD can be normalized to the coefficient of variation (CV) using Equation 3.5:

$$CV\% = 100 \cdot \frac{SD}{mean}. \tag{3.5}$$

Because the SD and mean are expressed in the same units, the CV is a unitless fraction. Often the CV is expressed as a percentage. If you change the units in which you express the data, you do not change the CV. In our example of 100 blood pressure measurements, the mean blood pressure is 123.4 mmHg and the SD is 14.0 mmHg. The CV is 11.4%. The CV is useful for comparing scatter of variables measured in different units.

SUMMARY

Many kinds of data are expressed as measurements. You can display the scatter of measurements in a sample on a histogram. The center of the distribution can be described by the mean or median. The spread or scatter of the data can be described by the range, the interquartile range, the variance, the SD, or the coefficient of variation.

OBJECTIVES

1. You must be familiar with the following terms:
   - Error
   - Bias
   - Histogram
   - Mean
   - Median
   - Standard deviation
   - Variance
   - Coefficient of variation
2. Given a list of measurements, you should be able to
   - Define the population the values came from.
   - Determine whether it is reasonable to think that the data are representative of that population.
   - Draw a histogram of the distribution of values, and a box-and-whiskers plot.
   - Calculate the mean, median, standard deviation, and coefficient of variation.

*This section is more advanced than the rest. You may skip it without loss of continuity.

PROBLEMS

1. Estimate the SD of the age of the students in a typical medical school class (just a rough estimate will suffice).
2. Estimate the SD of the number of hours slept each night by adults.
3. The following cholesterol levels were measured in 10 people (mg/dl):
   260, 150, 160, 200, 210, 240, 220, 225, 210, 240
   Calculate the mean, median, variance, standard deviation, and coefficient of variation. Make sure you include units.
4. Add an 11th value (931) to the numbers in Question 3 and recalculate all values.
5. Can a SD be 0? Can it be negative?

4

The Gaussian Distribution

PROBABILITY DISTRIBUTIONS

As you saw in the last chapter, a histogram plots the distribution of values in a sample. In many situations, the histogram is bell shaped. The first two panels in Figure 4.1 show bell-shaped distributions of two samples. The second sample is larger, and the histogram was created with narrower bins. This makes it appear more smooth.

The third panel in the figure plots the distribution of values in the entire population. This is not a histogram because the population is infinitely large. What you see instead is a probability distribution. Although probability distributions are similar to histograms, the Y axes are different. In a histogram, the Y axis shows the number of observations in each bin. In a probability distribution, the Y axis can't show the number of observations in each bin because the population is infinitely large and there aren't any bins. Instead we represent probability as the *area under the curve.* The area under the entire curve represents the entire population, and the proportion of area that falls between two X values is the probability of observing a value in that interval. The Y axis is termed the *probability density,* a term that is difficult term to define precisely. Fortunately, you can understand probability distributions without rigorously defining the scale of the Y axis.

THE GAUSSIAN DISTRIBUTION

The symmetrical bell-shaped distribution is called a *Gaussian distribution.* You expect variables to approximate a Gaussian distribution when the variation is caused by many independent factors. For example, in a laboratory experiment with membrane fragments, variation between experiments might be caused by imprecise weighing of reagents, imprecise pipetting, the variable presence of chunks in the membrane suspension, and other factors. When you combine all these sources of variation, you are likely to see measurements that distribute with an approximately Gaussian distribution. In a clinical study, you expect to see an approximately Gaussian distribution when the variation is due to many independent genetic and environmental factors. When the variation is largely due to one factor (such as variation in a single gene), you expect to see bimodal or skewed distributions rather than Gaussian distributions.

You learned how to calculate the mean and standard deviation (SD) of a sample in the last chapter. Figure 4.2 shows the mean and SD of an ideal Gaussian distribution.

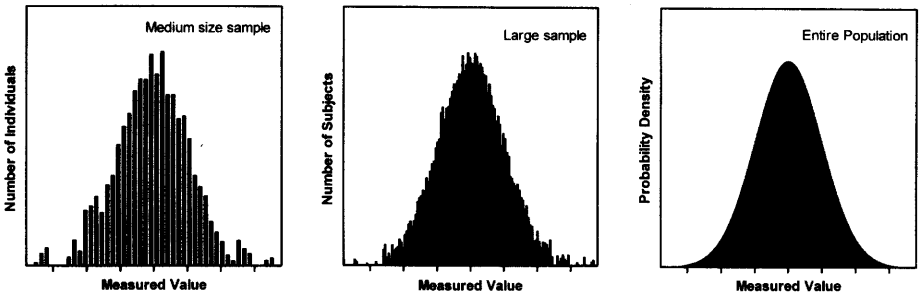


Figure 4.1. Histograms and probability distributions. The left panel shows a histogram for a medium-sized sample. The height of each bar denotes the number of subjects whose value lies within the span of its base. The middle panel shows a histogram for a larger sample. Each bar was made narrower. The right panel shows the distribution for the entire population. Since the population is infinite and there are no bars, the Y axis cannot denote numbers of subjects. Instead, it is called the *probability density.*

The mean, of course, is the center of the Gaussian distribution. The SD is a measure of the spread of the distribution. The area under the curve within 1 SD of the mean is shaded in Figure 4.3. The area under the entire curve represents the entire population. The shaded portion is about two thirds of the entire area. This means that about two thirds of a Gaussian population are within 1 SD of the mean.*

You can also see that the vast majority of the values lie within 2 SDs of the mean. More precisely, 95% of the values in a Gaussian distribution lie within 1.96 SDs of the mean. Areas under a portion of the Gaussian distribution are tabulated as the "z" distribution, where z is the number of standard deviations away from the mean.

For each value of z, the table shows four columns of probabilities. The first column shows the fraction of a Gaussian population that lies within z SDs of the mean. The last column shows the fraction that lies more than z SDs of the mean above the mean or more than z SDs below the mean. Thus the sum of the first column and the last column equals 100% for all values of z. Because the value in the last column includes components on both sides of the distribution, it is termed a *two-tailed probability.* This two-tailed probability (column four) is the sum of each tail, the probability that a value will be more than z SDs above the mean (the second column) and the probability that a value will be more than z SDs below the mean (the third column). The Gaussian distribution is symmetrical so the values in the second and third columns are identical and sum to the values in the fourth column. A longer version of this table is included in the Appendix as Table A5.2.

From Figure 4.3 we estimated that about two thirds of data in a Gaussian distribution lie within 1 SD of the mean. We can get a more exact value from this table. Look in the first column of the row for z = 1.0. The value is 68.27%. Look in the row for z = 2.0. Just over 95% of the population lies within 2 SDs of the mean. More precisely,

*I have deliberately avoided defining the scale of the Y axis or the units used to measure the area under the curve. We never care about area per se, but only about the fraction of the total area that lies under a defined portion of the curve. When you calculate this fraction the units cancel, so you can get by without understanding them precisely.

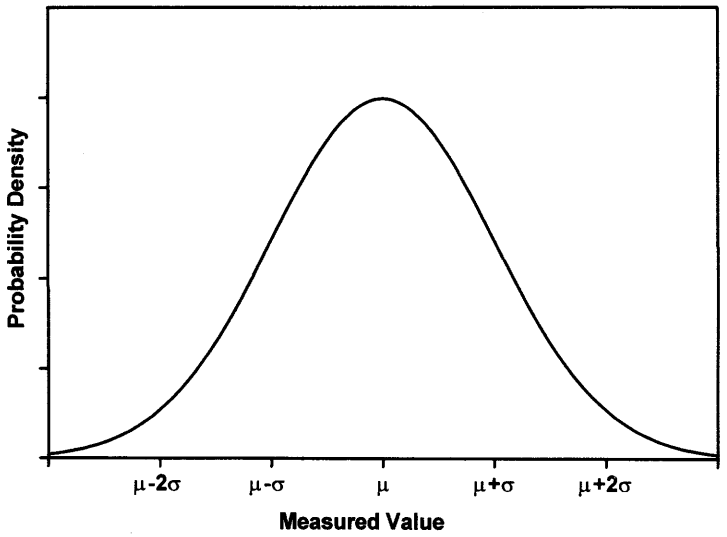


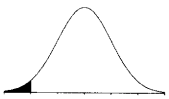
Figure 4.2. An ideal Gaussian distribution. The distribution is centered on the mean μ. The width of the distribution is determined by the standard deviation σ.

95% of the population lies within 1.96 SDs of the mean. You'll see that number 1.96 again, so you should remember what it means.

USING THE GAUSSIAN DISTRIBUTION TO MAKE INFERENCES ABOUT THE POPULATION

In the example of the previous chapter, we analyzed the blood pressure of 100 subjects. The sample mean was 123.4 mmHg and the sample SD was 14.0 mmHg. These values are also our best guesses for the mean and SD of the population.

Table 4.1. The Gaussian Distribution

	Fraction of the Population that is			
z	Within z SDs of the mean	More than z SDs above the mean	More than z SDs below the mean	More than z SDs above or below the mean
				
0.5	38.29%	30.85%	30.85%	61.71%
1.0	68.27%	15.87%	15.87%	31.73%
1.5	86.64%	6.68%	6.68%	13.36%
2.0	95.45%	2.28%	2.28%	4.55%
2.5	98.76%	0.62%	0.62%	1.24%
3.0	99.73%	0.13%	0.13%	0.27%
3.5	99.95%	0.02%	0.02%	0.05%

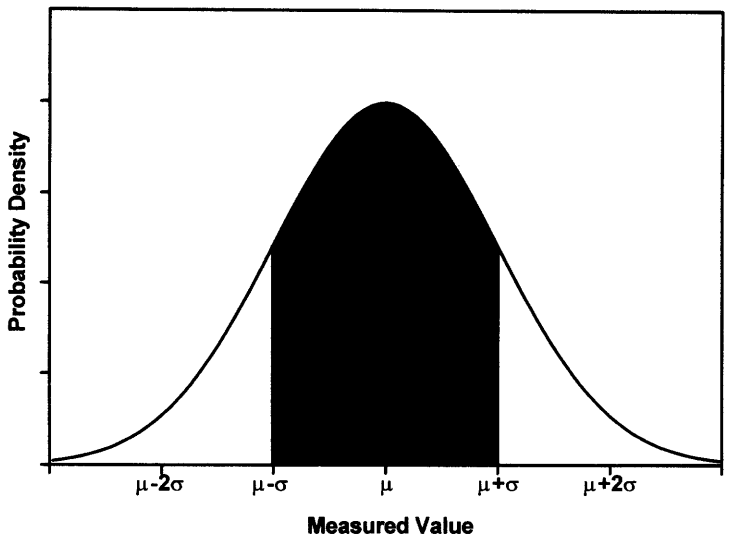


Figure 4.3. Interpreting probability distributions. The area under the entire curve represents the entire population. All values within 1 SD from the mean are shaded. The ratio of the shaded area to the total area is the fraction of the population whose value lies within 1 SD of the mean. You can see that the shaded area is about two thirds of the total area. If you measured it more exactly, you'd find that the shaded area equals 68.27% of the total area.

If you are willing to assume that the distribution of values in the population follows a Gaussian distribution, then you can plot your best guess for that distribution (Figure 4.4) and can make quantitative predictions about the population.

When reading research articles, you'll often see means and SDs without seeing the distribution of the values. When you read "mean = 123, SD = 14," you should have an intuitive sense of what that means if the population is Gaussian. If you have a talented right brain, you might be able to visualize Figure 4.4 in your head. If your left brain dominates, you should perform some very rough calculations in your head (two thirds of the values lie between 109 and 137). Either way, you should be able to interpret SDs in terms of the probable distribution of the data.

You can use the z table to answer questions such as this one: What fraction of the population has a systolic blood pressure greater than 140? Since the mean is 123.4 and the SD = 14.0, we are asking about deviations more than (140.0 − 123.4)/14 = 1.2 SDs from the mean. Setting z = 1.2, you can consult Table A5.2 in the Appendix and predict that 11.51% of the population has a blood pressure greater than 140.

More generally, you can calculate z for any value Y using Equation 4.1:

$$z = \frac{|Y - \text{mean}|}{SD}. \tag{4.1}$$

There are two problems with these calculations. One problem is that you don't know the mean and SD of the entire population. If your sample is small, this may introduce substantial error. The sample in this example had 100 subjects, so this is not a big problem. The second problem is that the population may not really be Gaussian.

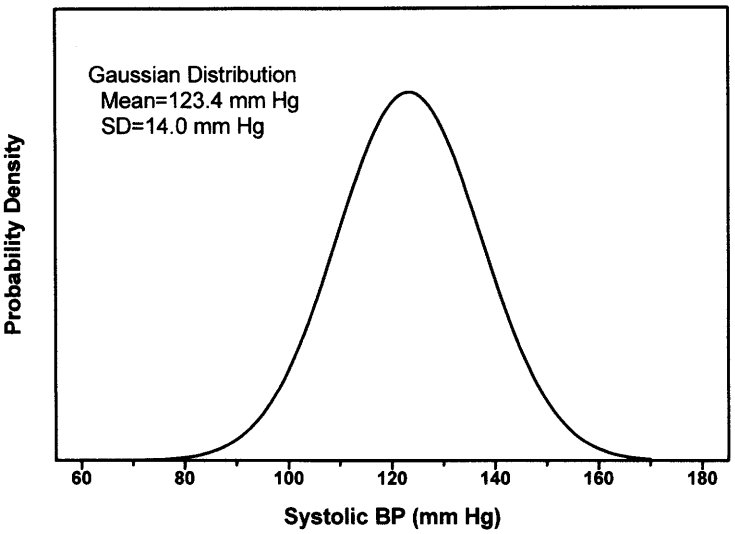


Figure 4.4. Estimating the distribution of values. If you knew only that the mean = 123.4 and the SD = 14 mmHg, what can you say about the distribution of values in the overall population? Before you can make any inferences, you need to make additional assumptions. If you assume that the population follows a Gaussian distribution (at least approximately), then this figure is your best guess.

This is a big problem. Even if the distribution only deviates a bit from a Gaussian distribution, the deviation will be most apparent in the tails of the distribution. But this example asks about the fraction of values in the tail. If the population is not really Gaussian, the calculation will give the wrong result.

THE PREDICTION INTERVAL*

You know that 95% of the values in a Gaussian population lie within 1.96 SDs of the population mean. What can you say about the distribution of values when you only know the mean and SD of one sample?

If your sample is large, then the mean and SD calculated from a sample are likely to be very close to the true population values and you expect 95% of the population to lie within 1.96 SDs of the sample mean. If your sample is small, you'll need to go more than 1.96 SDs in each direction to account for the likely discrepancies between the mean and SD you calculated from the sample and the true mean and SD in the overall population. To include 95% of the population, we need to create a range of values that goes more than 1.96 SDs in each direction of the sample mean. This range is called the *prediction interval.*

To calculate a prediction interval, you need to go K SDs from the mean, where K is obtained from Table 4.2 (the derivation of this table is beyond the scope of this book). For example, if your sample has five subjects (N = 5), then the prediction

*This section is more advanced than the rest. You may skip it without loss of continuity.

Table 4.2. 95% Prediction Interval

N	K	N	K
2	15.56	18	2.17
3	4.97	19	2.16
4	3.56	20	2.14
5	3.04	25	2.10
6	2.78	30	2.08
7	2.62	35	2.06
8	2.51	40	2.05
9	2.43	50	2.03
10	2.37	60	2.02
11	2.33	70	2.01
12	2.29	80	2.00
13	2.26	90	2.00
14	2.24	100	1.99
15	2.22	200	1.98
16	2.20	∞	1.96
17	2.18		

These values were calculated according to page 62 of G. J. Hahn and W. Q. Meeker, *Statistical Intervals*, John Wiley & Sons, 1991.

interval extends 3.04 SDs in each direction. With N = 10, the prediction interval extends 2.37 SDs in each direction.

Here is the precise definition of the 95% prediction interval: Each new observation has a 95% chance of being within the interval. On average, you expect that 95% of the entire population lies within the prediction interval, but there is a 50% chance that fewer than 95% of the values lie within the interval and a 50% chance that more than 95% of the values lie within the interval. The prediction interval is only valid when the population is distributed according to a Gaussian distribution.

NORMAL LIMITS AND THE "NORMAL" DISTRIBUTION

From our blood pressure data, what can we say about the "normal range" of blood pressures? Here is a simple, but flawed, approach that is often used:

> I am willing to assume that blood pressure in the population follows a Gaussian distribution. We know that 95% of the population lies within 1.96 SDs of the mean. We'll define the other 5% as abnormal. Using the sample mean and SD from our study, we can define the "normal range" as 123.4 $\pm$ (1.96 $\times$ 14.0) or 96 to 151 mm Hg.

There are a number of problems with this simple approach:

- It really doesn't make sense to define normal and abnormal just in terms of the distribution of values in the general population. What you really care about is consequences. We know that high blood pressure can cause atherosclerosis, heart disease, and strokes. So the questions we really want to answer are these: In which subjects is the blood pressure high enough to be dangerous? More precisely, we want to know which subjects have blood pressure high enough that the increased risk of heart

disease and stroke is high enough to justify the hassle, expense, and risk of treatment. Answering these questions is difficult. You need to follow large groups of people for many years. There is no way to even begin to approach these questions by analyzing only the distribution of blood pressure values collected at one time. You also need to assess blood pressure in a standard manner so that the results are comparable. In this "study" inexperienced students measured each other's pressures. The measurements are likely to be far from the true arterial pressure (it takes experience to measure blood pressure), and the arterial pressure is likely to be far from its resting value (some will find it stressful to have a friend take their blood pressure in public).

- Our definition of normal and abnormal really should depend on other factors such as age and sex. What is abnormal for a 25 year old may be normal for an 80 year old.
- In many cases, we don't really want a crisp boundary between "normal" and "abnormal." It often makes sense to label some values as "borderline."
- You don't need to decide based on one measurement. Before labeling someone as having abnormally high blood pressure, you should accurately measure the pressure several times in a controlled environment.
- Even if the population is approximately Gaussian, it is unlikely to follow a Gaussian distribution exactly. The deviations from Gaussian are likely to be most apparent in the tails of the distribution. There is also no reason to think that the population distribution is symmetrical around the mean.
- Why do you want to define 2.5% of the population to be abnormally high and 2.5% of the population to be abnormally low? What's so special about 2.5%? Why should there be just as many abnormally high values as abnormally low?

The problem is that the word *normal* has at least three meanings:

- Mathematical statisticians use the term as a synonym for a Gaussian distribution. However, there is nothing unusual about variables distributed in some other way.
- Scientists usually use the term to denote values that are commonly observed.
- Clinicians sometimes use *normal* to mean usual, and sometimes use *normal* to mean values that are not associated with the presence of disease.

You should be a bit confused at this point. Determining a "normal range" is a complicated issue. The important point is to realize that it is usually far too simple to define the "normal range" as the mean plus or minus two SDs.

> Everybody believes in the normal approximation, the experimenters because they think it is a mathematical theorem, the mathemeticians because they think it is an experimental fact.
>
> G. Lippman (1845–1921)

SUMMARY

Many variables distribute in a Gaussian distribution. This is expected when many independent facctors account for the variability, with no one factor being most important. Areas under portions of a Gaussian distribution are tabulated as the "z" distribution. The range of values that are 95% certain to contain the next value sampled from the population is known as the *prediction interval.* The term *normal limits* can

be defined in many ways. Using the Gaussian distribution to make rules about "normal" values is not always useful.

OBJECTIVES

1. You should be familiar with the following terms:
   - Gaussian distribution
   - Probability distribution
   - Normal
   - Prediction interval
2. Using tables, you should be able to determine the fraction of a Gaussian population that lies more (or less) than z SDs from the mean.
3. You should understand the complexities of "normal limits" and the problem of defining normal limits as being the mean plus or minus 2 SDs.

PROBLEMS

1. The level of an enzyme in blood was measured in 20 patients, and the results were reported as 79.6 ± 7.3 units/ml (mean ± SD). No other information is given about the distribution of the results. Estimate what the probability of distribution might have looked like.
2. The level of an enzyme in blood was measured in 20 patients and the results were reported as 9.6 ± 7.3 units/ml (mean ± SD). No other information is given about the distribution of the results. Estimate what the probability of distribution might have looked like.
3. The values of serum sodium in healthy adults approximates a Gaussian distribution with a mean of 141 mM and a SD of 3 mM. Assuming this distribution, what fraction of adults has sodium levels less than 137 mM? What fraction has sodium levels between 137 and 145 mM?
4. You measure plasma glaucoma levels in five patients and the values are 146, 157, 165, 131, 157 mg/dl. Calculate the mean and SD. What is the 95% prediction interval? What does it mean?
5. The Weschler IQ scale was created so that the mean is 100 and the standard deviation is 15. What fraction of the population has an IQ greater than 135?

5

The Confidence Interval of a Mean

INTERPRETING A CONFIDENCE INTERVAL OF A MEAN

You already learned how to interpret the CI of a proportion in Chapter 2. So interpreting the CI of a mean is easy. The 95% CI is a range of values, and you can be 95% sure that the CI includes the true population mean.

Example 5.1

This example is the same as the example followed in the last two chapters. You've measured the systolic blood pressure (BP) of all 100 students in the class. The sample mean is 123.4 mmHg, and the sample SD is 14.0 mmHg. You consider this class to be representative of other classes (different locations, different years), so you know that the sample mean may not equal the population mean exactly. But with such a large sample size, you expect the sample mean to be close to the population mean. The 95% CI for the population mean quantifies this. You'll learn how to calculate the CI later in the chapter. For now, accept that a computer does the calculations correctly and focus on the interpretation. For this example, the 95% CI ranges from 120.6 mmHg to 126.2 mmHg. You can be 95% sure that this range includes the overall population mean.

Example 5.2

This example continues the study of BPs. But now you randomly select a sample of five students, whose systolic BPs (rounded off to the nearest 5) are 120, 80, 90, 110, and 95 mmHg. The mean of this sample is 99.0 mmHg and the sample SD is 15.97 mmHg. But the sample mean is unlikely to be identical to the population mean. With such a large SD and such a small sample, it seems likely that the sample mean may be quite far from the population mean. The 95% CI of the population mean quantifies the uncertainty. In this example, the 95% CI ranges from 79.2 to 118.8 mmHg. You can be 95% sure that this wide range of values includes the true population mean.

What Does it Mean to be 95% Sure?

Note that there is no uncertainty about the sample mean. We are 100% sure that we have calculated the sample mean correctly. By definition, the CI is always centered

on the sample mean. We are not sure about the value of the population mean. However, we can be 95% sure that the calculated interval contains it.

What exactly does it mean to be "95% sure"? If you calculate a 95% CI from many independent samples, the population mean will be included in the CI in 95% of the samples, but will be outside of the CI in the other 5% of the samples. When you only have measured one sample, you don't know the value of the population mean. The population mean either lies within the 95% CI or it doesn't. You don't know, and there is no way to find out. But you can be 95% certain that the 95% CI includes the population mean.

The correct syntax is to express the CI as 79.2 to 118.8 mmHg, or as [79.2, 118.8]. It is considered bad form to express the CI as 99.0 ± 19.8 or 79.2 − 118.8 (the latter form would be confusing with negative numbers).

ASSUMPTIONS THAT MUST BE TRUE TO INTERPRET THE 95% CI OF A MEAN

The CI depends on these assumptions:

- Your sample is randomly selected from the population.

  In Example 5.1, we measured the BP of every student in the class. If we are only interested in that class, there would be no need to calculate a CI. We know the class mean exactly. It only makes sense to calculate a CI to make inferences about the mean of a larger population. For this example, the population might include students in future years and in other cities.

  In Example 5.2, we randomly sampled five students from the class, so the assumption is valid. If we let students volunteer for the study, we'd be measuring BP in students who are especially interested in their BP (perhaps they've been told it was high in the past). Such a sample is not representative of the entire population, and any statistical inferences are likely to be misleading.

  In clinical studies, it is not feasible to randomly select patients from the entire population of similar patients. Instead, patients are selected for the study because they happened to be at the right clinic at the right time. This is called a *convenience sample* rather than a *random sample.* For statistical calculations to be meaningful, we must assume that the convenience sample adequately represents the population, and the results are similar to what would have been observed had we used a true random sample.
- The population is distributed in a Gaussian manner, at least approximately. This assumption is not too important if your sample is large. Since Example 5.1 has 100 subjects, the Gaussian assumption matters very little (unless the population distribution is very bizarre). Example 5.2 has only five students. For such a small sample, the CI can only be interpreted if you assume that the overall population is approximately Gaussian.
- All subjects come from the same population, and each has been selected independently of the others. In other words, selecting one subject should not alter the chances of selecting any other. Our BP example would be invalid if there were really less than 100 students, but some were measured twice. It would also be invalid if some of the

students were siblings or twins (because BP is partly controlled by genetic factors, so two siblings are likely to have pressures more similar than two randomly selected subjects).

In many situations, these assumptions are not strictly true: The patients in your study may be more homogeneous than the entire population of patients. Measurements made in one lab will have a smaller SD than measurements made in other labs at other times. More generally, the population you really care about may be more diverse than the population that the data were sampled from. Furthermore, the population may not be Gaussian. If any assumption is violated, the CI will probably be too optimistic (too narrow). The true CI (taking into account any violation of the assumptions) is likely to be wider than the CI you calculate.

CALCULATING THE CONFIDENCE INTERVAL OF A MEAN

To calculate the CI, you need to combine the answers to four questions:

1. What is the sample mean? It is our best guess of the population mean.
2. How much variability? If the data are widely scattered (large SD), then the sample mean is likely to be further from the population mean than if the data are tight (small SD).
3. How many subjects? If the sample is large, you expect the sample mean to be close to the population mean and the CI will be very narrow. With tiny samples, the sample mean may be far from the population mean so the confidence interval will be wide.
4. How much confidence? Although CIs are typically calculated for 95% confidence, any value can be used. If you wish to have more confidence (i.e. 99% confidence) you must generate a wider interval. If you are willing to accept less confidence (i.e. 90% confidence), you can generate a narrower interval.

Equation 5.1 calculates the 95% CI of a mean from the sample mean (m), the sample standard deviation (s) and the sample size (N):

$$95\%\ \text{CI:}\ \left(m - t^* \cdot \frac{s}{\sqrt{N}}\right)\ \text{to}\ \left(m + t^* \cdot \frac{s}{\sqrt{N}}\right) \tag{5.1}$$

The CI is centered on the sample mean m. The width of the interval is proportional to the SD of the sample, s. If the sample is more scattered, the SD is higher and the CI will be wider. The width of the interval is inversely proportional to the square root of sample size, N. Everything else being equal, the CIs from larger samples will be narrower than CIs from small samples.

The width of the CI also depends on the value of a new variable, t*. As you expect, its value depends on how confident you want to be. If you want 99% confidence instead of 95% confidence, t* will have a larger value. If you only want 90% confidence, t* will have a smaller value. Additionally, the value of t* depends on the sample size. The value comes from the t distribution, which you'll learn more about later in the chapter. The value of t* is tabulated in Table 5.1, and more fully in Table A5.3 in the Appendix. Since this table has several uses, the rows do not directly show sample size,

Table 5.1. Critical Value of t for 95% Confidence

df	t*	df	t*
1	12.706	12	2.179
2	4.303	13	2.160
3	3.182	14	2.145
4	2.776	15	2.131
5	2.571	20	2.086
6	2.447	25	2.060
7	2.365	30	2.042
8	2.306	40	2.021
9	2.262	60	2.000
10	2.228	120	1.980
11	2.201	∞	1.960

but rather show degrees of freedom (df). When calculating the confidence interval of a mean, df = N−1 as explained previously on pages 27–28.

In Example 5.1, the sample mean is 123.4 mmHg and the SD is 14.0 mmHg. Because the sample has 100 subjects, there are 99 df. The table does not show t* for 99 df. Interpolating between the values shown for 60 df and 120 df, t* is about 1.99. The CI ranges from 120.6 to 126.2 mmHg.

Note that as the sample size increases, the value of t* approaches 1.96 (for 95% confidence). It is enough to remember that the value is approximately 2, so long as the sample has at least 20 or so subjects.

For Example 5.2, the mean of this sample is 99.0 mmHg with a SD of 15.97 mmHg. With only five subjects, there are only 4 df, and t* equals 2.776. The 95% CI of the population mean ranges from 79.2 to 118.8 mmHg. With a small sample, the confidence interval is wide.

To understand why Equation 5.1 works, you need to understand two things. First, you need to understand where the ratio $s/\sqrt{N}$ comes from. Next you need to understand where the value of t* comes from. Both of these topics are complicated, but I'll try to explain them briefly in the next two sections.

THE CENTRAL LIMIT THEOREM

Where does the ratio $s/\sqrt{N}$ in Equation 5.1 come from? To understand the answer, we first need to digress a bit.

If you collect many samples from one population, they will not all have the same mean. What will the distribution of means look like? Mathematicians approach this kind of question by manipulating equations and proving theorems. We'll answer the question by simulating data, an approach that is far more intuitive to many.

Figure 5.1 shows simulated data from throwing dice 2000 times. The X axis shows the average of the numbers that appear on top of the dice, and the Y axis shows the number of times each average appears. We are sampling from a "population" of numbers with a flat distribution—each die is equally likely to land with 1, 2, 3, 4, 5, or 6 on top. The left panel shows a histogram of 2000 throws of one die. The distribution

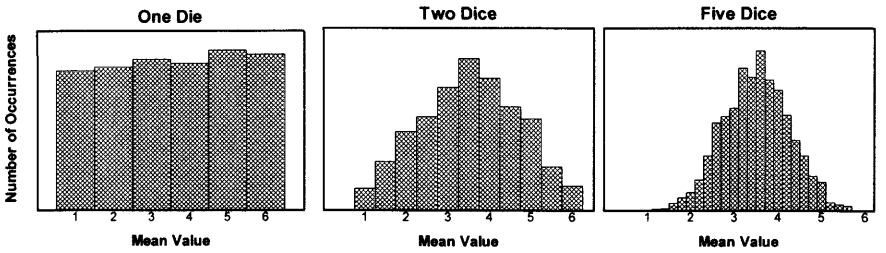


Figure 5.1. An illustration of the central limit theorem. Each panel shows the simulated results of throwing dice 2000 times. The X axis shows the mean of the values on top of the dice, and the Y axis shows how frequently each value appeared. If you only throw one die at a time (left panel) the distribution of values is flat. Each number is equally likely to appear. If you throw two dice at a time (middle panel), the distribution is not flat. You are more likely to get a mean of 3.5 (a sum of 7) than a mean of 1 or 6. The right panel shows the results with five dice. The histogram starts to look like a Gaussian distribution. The central limit theorem states this: No matter how the original population is distributed, the distribution of sample means will approximate a Gaussian distribution if the samples are large enough.

of this sample of 2000 observations closely resembles the overall "population," with some random variability. In the middle panel, two dice were thrown 2000 times, and we plot the distribution of the sample of 2000 mean values. This distribution differs substantially from the original population. If you've every played with dice, this comes as no surprise. There are six ways for a pair of dice to sum to seven (1&6, 2&5, 3&4, 4&3, 5&2, or 6&1), but there is only one way for the sum to equal 12 (6&6) or 2 (1&1). The figure shows average values, and shows that the average of 3.5 (two dice totaling seven) occurred about six times more often than did the average 1 or 6. The right panel shows the results when we threw five dice in each of 2000 trials. The distribution is approximately bell-shaped.

This example illustrates an important statistical theorem, known as the central limit theorem. The proof of this theorem is not accessible to nonmathematicians. But the theorem itself can be understood: No matter how the values in the population are distributed, the distribution of means from independently chosen samples will approximate a Gaussian distribution, so long as your samples are large enough.

The central limit theorem explains the computer generated data of Figure 5.1. Even though the "population" has a flat and chunky (only integers) distribution, the distribution of means approximates a bell-shaped distribution. As you increase the sample size, the distribution of sample means comes closer and closer to the ideal Gaussian distribution.

No matter how the population is distributed, the distribution of sample means will approximate a Gaussian distribution if the sample size is large enough. How large is that? Naturally, the answer is "it depends"! It depends on your definition of "approximately" and on the distribution of the population. Even if the population distribution is really weird, a sample of 100 values is large enough to invoke the central limit theorem. If the population distribution is approximately symmetrical and unimodal (it looks like a mountain and not like a mountain range), then you may invoke the central limit theorem even if you have only sampled a dozen or so values.

What is the standard deviation (SD) of the distribution of sample means? It is not the same as the SD of the population. Since sample means are distributed more compactly than the population, you expect the SD of the distribution of sample means to be smaller than the SD of the population. You also expect it to depend on the SD of the population and the sample size. If the population is very diverse (large SD), then the sample means will be spread out more than if the population is very compact (small SD). If you collect bigger samples, the sample means will be closer together so the SD of sample means will be smaller.

The central limit theorem proves that the SD of sample means equals the SD of the population divided by the square root of the sample size. This is the origin of the ratio $s/\sqrt{N}$ in Equation 5.1.

THE STANDARD ERROR OF THE MEAN

Because the phrase *standard deviation of sample means* is awkward, this value is given a shorter name, the *standard error of the mean,* abbreviated SEM. Often the SEM is referred to as the *standard error,* with the word *mean* missing but implied. The term is a bit misleading, as the standard error of the mean usually has nothing to do with standards or errors.

The central limit theorem calculates the SEM from the SD of the population and the sample size with Equation 5.2.

$$\text{SEM} = \frac{\text{SD}}{\sqrt{\text{N}}} \tag{5.2}$$

The SEM quantifies the precision of the sample mean. A small SEM indicates that the sample mean is likely to be quite close to the true population mean. A large SEM indicates that the sample mean is likely to be far from the true population mean.

Note that the SEM does *not* directly quantify scatter or variability in the population. Many people misunderstand this point. A small SEM can be due to a large sample size rather than due to tight data. With large sample sizes, the SEM is always tiny.

We can substitute Equation 5.2 into Equation 5.1 to calculate the 95% CI of a mean more simply.

$$95\%\ \text{CI: } (\text{m} - \text{t}^* \cdot \text{SEM}) \text{ to } (\text{m} + \text{t}^* \cdot \text{SEM}) \tag{5.3}$$

Since the value of t* is close to 2 for large sample sizes, you can remember that the 95% confidence interval of a mean extends approximately two standard errors on either side of the sample mean.

THE t DISTRIBUTION

Equations 5.1 and 5.3 include the variable t*. To understand those equations, therefore, you need to understand the t distribution.

Imagine this experiment. Start with a Gaussian population of known mean and SD. From the population, collect many samples of size N. For each sample, calculate the sample mean and SD and then calculate the ratio defined by Equation 5.4.

$$t = \frac{\text{Sample mean} - \text{Population mean}}{\text{Sample SD}/\sqrt{N}} = \frac{m - \mu}{\text{SEM}} \quad (5.4)$$

Since the sample mean is equally likely to be larger or smaller than the population mean, the value of t is equally likely to be positive or negative. So the t distribution is symmetrical around t = 0. With small samples, you are more likely to observe a large difference between the sample mean and the population mean, so the numerator of the t ratio is likely to be larger with small samples. But the denominator of the t ratio will also be larger with small samples, because the SEM is larger. So it seems reasonable that the distribution of the t ratio would be independent of sample size. But it turns out that that isn't true—including N in the equation does not entirely correct for differences in sample size, and the expected distribution of t varies depending on the size of the sample. Therefore, there is a family of t distributions for different sample sizes.

Figure 5.2 shows the distribution of t for samples of five subjects (df = 4). You'll see t distributions for other numbers of df in Chapter 23. The t distribution looks similar to the Gaussian distribution, but the t distribution is wider. As the sample size increases, the t distribution becomes more and more similar to the Gaussian distribution in accordance with the central limit theorem.

Since Figure 5.2 plots probability density, the area under the entire curve represents all values in the distribution. The tails of the distribution are shaded for all values of t greater than 2.776 or less than −2.776. Calculations prove that each of these tails represents 2.5% of the distribution. This means that the t ratio will be between −2.776 and 2.776 in 95% of samples (N = 5) randomly chosen from a Gaussian population. If the samples were of a different size, the cutoff would not be 2.776 but rather some other number as shown in Table 5.1.

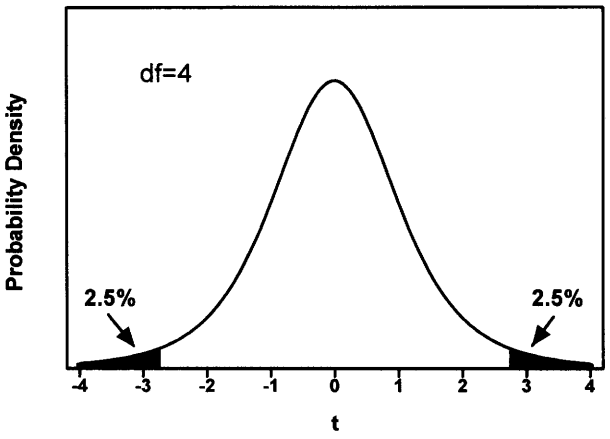


Figure 5.2. The t distribution for four degrees of freedom. Here are the steps you'd need to do to create this graph experimentally. Start with a population whose mean you know. Collected many samples of five subjects (df = 4) from this population. Calculate the t ratio for each using Equation 5.4. Of course, mathematicians can derive the distribution from first principles without the need for endless sampling. The largest and smallest 2.5% of the distribution are shaded. In 95% of the samples, t will be between −2.776 and 2.776.

So far, the logic has been in the wrong direction—we've started with a known population and are looking at the variability between samples. Data analysis goes the other way. When calculating CI, you don't know the population mean. All you know is the mean and SD of one particular sample. But we can use the known distribution of t to make inferences about the population mean. Let's compute the CI for example 5.2. The sample has five subjects, so you can be 95% sure that the t ratio will be between −2.776 and 2.776. Rearrange Equation 5.4 to solve for the population mean as a function of the sample mean (m), the sample standard deviation (s), the sample size (N), and t.

$$\mu = m - t \cdot \frac{s}{\sqrt{N}} = m - t \cdot SEM \tag{5.5}$$

You know the sample mean (m = 99.0 mmHg), the sample SD (s = 15.97 mmHg), and the sample size (N = 5). You also know that you can be 95% sure that t will be between −2.776 and 2.776. To calculate the confidence limits, first set t = 2.776 and calculate that μ = 79.2 mmHg. Then set t = −2.78 and calculate that μ = 118.8 mmHg. You've calculated the lower and upper limits of the 95% confidence interval. The 95% CI ranges from 79.2 to 118.8 mmHg.

In deriving the t distribution, statisticians started with a hypothetical population with known population mean and SD, and calculated the distribution of t in many samples. We then use the distribution backwards. Instead of making inferences about the distribution of samples from a population, you can make inferences about the population from a single sample. The ability to use probability distributions backwards is mathematically simple but logically profound, and of immense practical importance as it makes statistics useful to experimenters.

THE GAUSSIAN ASSUMPTION AND STATISTICAL INFERENCE

As noted earlier, the interpretation of the 95% CI depends on the assumption that the data were sampled from a Gaussian distribution. And, as you'll see in future chapters, this assumption is common to many statistical tests. Is this assumption reasonable for the blood pressure example? A few notes:

- The Gaussian distribution, by definition, extends infinitely in each direction. It allows for a very small proportion of BPs less than 0 (physically impossible) and a very small proportion of BPs greater than 300 (biologically impossible). Therefore, BPs cannot follow a Gaussian distribution exactly.
- The assumption doesn't need to be completely true to be useful. Just like the mathematics of geometry is based on mathematical abstractions such as a perfect rectangle, the mathematics of statistics is based upon the abstract Gaussian distribution. Geometry tells us that the area of a perfect rectangle equals its length times its width. No room is a perfect rectangle, but you can use that rule to figure out how much wallpaper you need. The calculation, based on an ideal model, is useful even if the walls are a bit warped. Similarly, statistical theory has devised methods for analyzing data obtained from Gaussian populations. Those methods are useful even when the distributions are not exactly Gaussian.
- You can look beyond our one sample. BP has been measured in lots of studies, and we can use information from these studies. While the distribution of BPs is often a

bit skewed (more high values than low) they tend to distribute according to a roughly Gaussian distribution. Since this has been observed in many studies with tens of thousands of subjects, it seems reasonable to assume that BP in our population is approximately Gaussian.

- You can think about the sources of scatter. The variability in BP is due to numerous genetic and environmental variables, as well as imprecise measurements. When scatter is due to the sum of numerous factors, you expect to observe a Gaussian distribution, at least approximately.*
- You can perform formal calculations to test whether a distribution of data is consistent with the Gaussian distribution. For more information, read about the Kolmogorov-Smirnov test in a more advanced book.

What should you do if the distribution of your data deviate substantially from a Gaussian distribution? There are three answers.

- You can mathematically transform the values to convert a nongaussian population into a Gaussian one. This is done by converting each value into its logarithm, reciprocal, or square root (or some other function). While this sounds a bit dubious, it can be a good thing to do. Often, there is a good biological or chemical justification for making the transformation. For example, it often makes sense both biologically and statistically to express acidity as pH rather than concentration of hydrogen ions, to express potency of a drug as log(EC_{50}) rather than EC_{50}†, and to express kidney function as the reciprocal of plasma creatinine concentration rather than the plasma creatinine concentration itself.
- You can rely on the central limit theorem and analyze large samples using statistical methods based on the Gaussian distribution, even if the populations are not Gaussian. You can rely on the central limit theorem when both of the following are true: (1) You are making inferences about the population means, and not about the details of the distribution itself. (2) Either the samples are very large or the population is approximately Gaussian.
- You can use statistical methods that are not based on the Gaussian distribution. For example, it is possible to calculate the 95% CI of a median without making any assumption about the distribution of the population. We'll discuss some of these methods, termed nonparameteric methods, later in the book.

CONFIDENCE INTERVAL OF A PROPORTION REVISITED*

You previously saw the CI of the proportion defined in Equation 2.1. This can be rewritten as shown in Equation 5.6.

Approximate 95% CI of proportion:

$$\left(p - z^* \cdot \sqrt{\frac{p(1-p)}{N}}\right) \text{ to } \left(p + z^* \cdot \sqrt{\frac{p(1-p)}{N}}\right) \tag{5.6}$$

*You also need to assume that the various factors all have nearly equal weight.

†The EC_{50} is the concentration of drug required to get half the maximal effect.

The value of z* comes from the Gaussian distribution. It equals 1.96 for 95% CI, because 95% of observations in a Gaussian population lie within 1.96 SD of the mean. You may substitute other numbers from the Gaussian distribution if you want to change the degree of confidence. For example, set z* = 2.58 to generate a 99% CI. This is the correct value, because 99% of a Gaussian distribution lie within 2.58 SD of the mean. The value of z* does not depend on the size of your sample.

This approximation works because the binomial distribution approximates the Gaussian distribution with large samples. It is a reasonable approximation when the numerator of the proportion is at least five, and the denominator is at least five larger than the numerator. If you have small samples or want to calculate the CI of a proportion more exactly, you must use tables or programs that are based on the binomial distribution. You should *not* replace z* in Equation 5.6 with t*.

ERROR BARS

Graphs of data often include error bars, and the text of articles often gives values plus or minus an error, i.e. 10.34 ± 2.3. The error bar or the plus/minus value usually denotes the SD or the SEM. Occasionally the error bars denote the range of the data, or the 95% CI of the mean. You must look at the methods section or figure legend to figure out what the authors are trying to say. If there is no explanation as to how the error bars were calculated, the error bar is nearly meaningless.

Figure 5.3 shows the appearance of several error bars. Note that some are capped while others are not. The difference is a matter of aesthetics, not statistics. Also note that some error bars extend above and below the data point, while others extend only above or only below. Usually this is done to prevent error bars from different data sets from overlapping. This is an artistic decision by the investigator. The true uncertainty always extends in both directions, even if the author only shows the error bar in one direction.

In theory, the choice of showing the SD or SEM should be based on statistical principles. Show the SD when you are interested in showing the scatter of the data. Show the SEM when you want to show how well you know the population mean. Sometimes people display the SEM for another reason: The SEM is the smallest measure of "error" and thus looks nicest.

Because the SD and SEM are used so frequently, anyone reading the biomedical literature must be able to convert back and forth without looking up the equations in a book. There are only a few equations in statistics that must be committed to memory, and these are two of them:

$$\text{SEM} = \text{SD} / \sqrt{\text{N}} \qquad (5.7)$$

$$\text{SD} = \text{SEM} \cdot \sqrt{\text{N}}$$

> The only role of the standard error . . . is to distort and conceal the data. The reader wants to know the actual span of the data; but the investigator displays an estimated zone for the mean.
>
> A. R. Feinstein, *Clinical Biostatistics*

*This section is more advanced than the rest. You may skip it without loss of continuity.

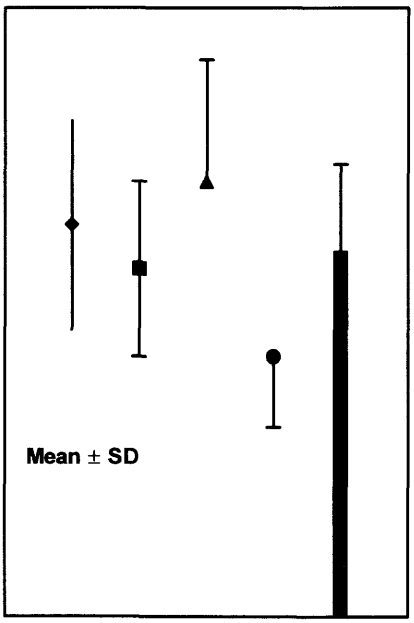


Figure 5.3. Error bars. Graphs show error bars in various ways. Note that the first bar is uncapped, and the others are capped. This is merely an artistic distinction that tells you nothing about the distribution of data. The first two error bars extend above and below the point, while the next two go only above or only below. This too is an artistic distinction that tells you nothing about the data. Note the legend telling us that these error bars represent the standard deviation. You can't interpret these error bars unless you know whether the size of the error bar shows you the standard deviation, the standard error of the mean, the extent of the 95% CI, or something else.

Remember that the scatter (however expressed) means different things in different contexts. Is the author showing the variability among replicates in a single experiment? Variability among experiments performed with genetically identical animals? Variability among cloned cells? Variability between patients? Your interpretation of error bars should depend heavily on such considerations.

SUMMARY

The sample mean you calculate from a list of values is unlikely to be exactly equal to the overall population mean (that you don't know). The likely discrepancy depends on the size of your sample and the variability of the values (expressed as the standard deviation). You can combine the sample mean, SD and sample size to calculate a

95% confidence interval of the mean. If your sample is representative of the population, you can be 95% sure that the true population mean lies somewhere within the 95% CI.

The *central limit theorem* states that the distribution of sample means will approximate a Gaussian distribution even if the population is not Gaussian. It explains why you can interpret a confidence interval from large samples even if the population is not Gaussian.

The *standard error of the mean* (SEM) is a measure of how close a sample mean is to the population mean. Although it is often presented in papers, it is easier to interpret confidence intervals, which are calculated from the SEM.

Papers often present the data as mean ± error, or show a graph with an *error bar.* These error values or error bars may reprsent the SD, the SEM or something else. You can not interpret them unless you know how they were defined.

OBJECTIVES

1. You must be familiar with the following terms:
   - Gaussian distribution
   - Central limit theorem
   - Standard error of the mean
   - Confidence interval of a mean
   - Confidence interval
   - Error bar
   - t distribution
2. Given a list of numbers, you should be able to calculate the 95% CI of the mean.
3. Given the mean, sample size, and SD or SEM, you should be able to calculate the 95% CI of the mean.
4. You should know the assumptions that must be true to interpret the CI.
5. You must be able to convert between SD and SEM without referring to a book.
6. You should be able to interpret error bars shown in graphs or tables of publications.

PROBLEMS

1. Figure 5.4 shows the probability distribution for the time that an ion channel stays open. Most channels stay open for a very short time, and some stay open longer. Imagine that you measure the open time of 10 channels in each experiment, and calculate the average time. You then repeat the experiment many times. What is the expected shape of the distribution of the mean open times?
2. An enzyme level was measured in cultured cells. The experiment was repeated on 3 days; on each day the measurement was made in triplicate. The experimental conditions were identical on each day; the only purpose of the repeated experiments was to determine the value more precisely. The results are shown as enzyme activity in units per minute per milligram of membrane protein.

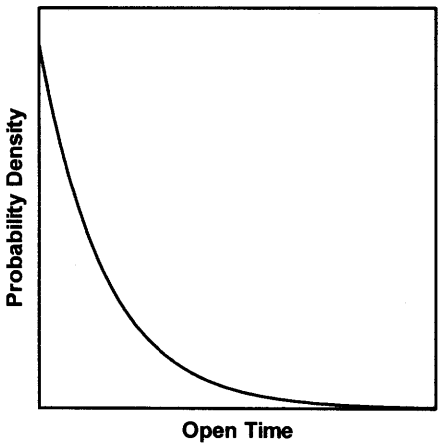


Figure 5.4.

	Replicate 1	Replicate 2	Replicate 3
Monday	234	220	229
Tuesday	269	967	275
Wednesday	254	249	246

Summarize these data as you would for publication. The readers have no interest in the individual results for each day; just one overall mean with an indication of the scatter. Give the results as mean, error value, and N. Justify your decisions.

3. Is the width of a 99% CI wider or narrower than the width of a 90% CI?
4. The serum levels of a hormone (the "Y factor") was measured to be 93 ± 1.5 (mean ± SEM) in 100 nonpregnant women and 110 ± 2.3 (mean ± SEM) in 100 women in the first trimester of pregnancy.
   A. Calculate the 95% CI for the mean value in each group.
   B. Assuming that the measurement of Y is easy, cheap, and accurate, is it a useful assay to diagnose pregnancy?
5. A paper reports that cell membranes have 1203 ± 64 (mean ± SEM) fmol of receptor per milligram of membrane protein. These data come from nine experiments.
   A. Calculate the 95% CI. Explain what it means in plain language.
   B. Calculate the 95% prediction interval. Explain what it means in plain language.
   C. Is it possible to sketch the distribution of the original data? If so, do it.
   D. Calculate the 90% CI.
   E. Calculate the coefficient of variation.
6. You measure BP in 10 randomly selected subjects and calculate that the mean is 125 and the SD is 7.5 mmHg. Calculate the SEM and 95% CI. Now you measure BP in 100 subjects randomly selected from the same population. What values do you expect to find for the SD and SEM?

7. Why doesn't it ever make sense to calculate a CI from a population SD (as opposed to a sample SD)?
8. Data were measured in 16 subjects, and the 95% CI was 97 to 132.
   A. Predict the distribution of the raw data.
   B. Calculate the 99% CI.
9. Which is larger, the SD or the SEM? Are there any exceptions?

III

INTRODUCTION TO P VALUES

I've put it off for nine chapters, but I can't delay any longer. It's time to confront P values. If you've had any exposure to statistics before, you've probably already heard about *P values* and *statistical significance.* It's time to learn what these phrases really mean. These chapters explain P values generally, without explaining any particular statistical tests in any detail. You'll learn more about specific tests in Part VII.

10

What Is a P Value?

INTRODUCTION TO P VALUES

When using statistics to compare two groups, there are two approaches you can use:

- You've already learned about one approach—calculating the confidence interval (CI) for the difference between means or the difference (or ratio) of two proportions. With this approach, you are focusing on the question: "How large is the difference in the overall population?" You start with the difference (or ratio) you know in the sample, and calculate a zone of uncertainty (the 95% CI) for the overall population.
- This chapter introduces you to the second approach, calculating P values. This approach focuses on a different question: "How sure are we that there is, in fact, a difference between the populations?" You observed a difference in your samples, but it may be that the difference is due to a coincidence of random sampling rather than due to a real difference between the populations. Statistical calculations cannot tell you whether that kind of coincidence has occurred but can tell you how rare such a coincidence would be.

Calculations of P values and CIs are based on the same statistical principles and the same assumptions. The two approaches are complementary. In this book, I separated the two approaches to aid learning. When analyzing data, the two approaches should be used together. The easiest way to understand P values is to follow an example.

A SIMPLE EXAMPLE: BLOOD PRESSURE IN MEDICAL STUDENTS*

You want to test whether systolic blood pressure differs between first- and second-year medical students (MS1 and MS2, respectively). Perhaps the stress of medical school increases blood pressure. Measuring the blood pressure in the entire class seems like a lot of work for a preliminary study, so instead you randomly selected five students from each class and measured his or her systolic blood pressure rounded to the nearest 5 mmHg:

MS1: 120, 80, 90, 110, 95

MS2: 105, 130, 145, 125, 115

*You've already encountered these fake data in Example 5.2, and you'll see them again late in the book.

First you should look at the data. It helps to see a graph. Figure 10.1 shows blood pressures for each individual. Clearly the blood pressure tends to be lower in the first-year students than in the second-year students, although the two overlap. The mean values are 99 for MS1 and 124 for MS2. The difference between the mean values is 25 mmHg.

Figure 10.2 shows only the mean and standard error of the mean (SEM). Note how easy it is to be misled by the SEM error bars. It is easy to forget that the SEM error bars do not directly show you the scatter of the data. Biomedical research papers often show data in the format of Figure 10.2, even though the format of 10.1 is more informative. To make any sense at all of Figure 10.2, you would need to read the figure legend or the methods section to find out the sample size and whether the error bars represented standard deviation (SD) or SEM (or something else). The right half of Figure 10.2 also shows mean and SEM, but the Y axis doesn't begin at 0. This appears to magnify the difference between the two classes. If you don't notice the scale of the axis, you could be deceived by the right half of Figure 10.2.

Next, you should think about the biological or clinical implications of the data. A change of 25 mmHg in blood pressure is substantial and would have important clinical ramifications if the difference were consistent. This interpretation does not come from any statistical calculation; it comes from knowing about blood pressure. In contrast, a change of 25 units in some other variable might be trivial. Statistical calculations can't determine whether differences are clinically or scientifically important.

Before continuing with data analysis, you need to think about the design of the study. Were the data collected in a way that any further analysis is useful? In this

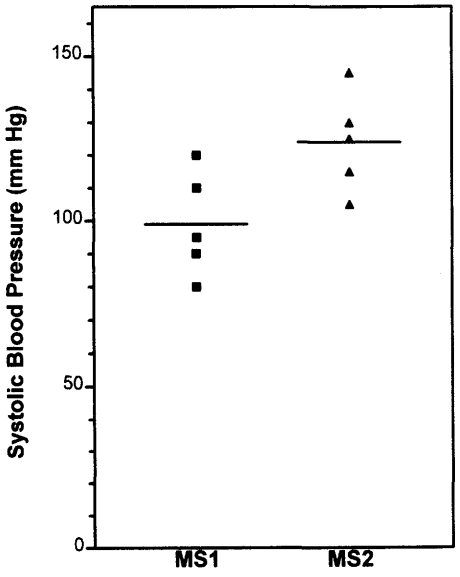


Figure 10.1. Sample data shown on a column scatter graph. Each square shows the systolic blood pressure of one first-year student (MS1). Each triangle shows the systolic blood pressure of one second-year student (MS2).

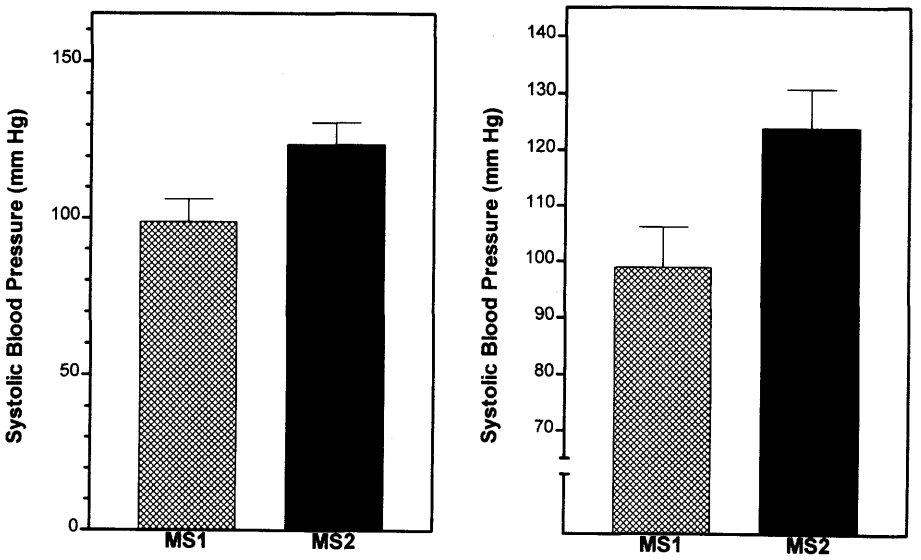


Figure 10.2. Sample data shown on a bar graph. This graph shows the mean and standard error of the blood pressures for the two samples. In the left panel, the Y axis begins at 0. In the right panel, the difference between the two means appears to be amplified because the axis doesn't begin at 0. You can be misled by this kind of graph unless you notice where the axis begins. You can't interpret these graphs unless you know that the error bar represents the standard error.

study, you would need to know how the students were selected and how their blood pressure was measured. We've already discussed this problem in Chapter 3. Let's assume that the experimental design was impeccable.

Two possible explanations remain for the difference between the two samples. The first possibility is that the overall populations of MS1s and MS2s have identical distributions of blood pressures, and that the difference observed in this experiment was just a coincidence—we just happened to select MS2s with higher blood pressure. The second possibility is that the mean systolic blood pressure of MS2s is really higher than that of MS1s. Before believing that these data represent a real difference, we would like to ask "what is the probability that the difference is due to chance?" Unfortunately, statistical calculations cannot answer that question.

Instead, statistical calculations can answer a related question: If one assumes that the distribution of blood pressure is identical in the two populations, what is the probability that the difference between the means of randomly selected subjects will be as large or larger than actually observed? The answer to this question is a probability, the P value.

There are several methods to calculate the P value, and these will be discussed later in the book. The general approach can be summarized as follows:

1. Assume that the individuals measured (the samples) are randomly selected from a larger group (the populations). If the subjects weren't randomly selected, at least assume that the subjects are representative of a larger population. Furthermore, assume that the experimental design is without biases or flaws.

2. Tentatively hypothesize that the distribution of values in the two populations are the same. This hypothesis is called the *null hypothesis,* sometimes abbreviated H_0. Most likely, the investigator did not believe (or did not want to believe) the null hypothesis. Rather, the investigator's experimental hypothesis (the whole reason for doing the experiment) was that the populations are different. This is sometimes called the *experimental* or *alternative hypothesis.*
3. Assuming the null hypothesis is true, calculate the likelihood of observing various possible results. You can use several methods (discussed later in this book) depending on the nature of the data and on which additional assumptions you are willing to make.
4. Determine the fraction of those possible results in which the difference between means is as large or larger than you observed. The answer is given in the form of a probability, called the *P value.*

Thinking about P values seems quite counterintuitive at first, as you must use backwards, awkward logic. Unless you are a lawyer or a Talmudic scholar used to this sort of argument by contradiction, you will probably find this sort of reasoning a bit uncomfortable. Four aspects are awkward:

- The hypothesis you are testing (the null hypothesis) is opposite to the hypothesis the experimenter expects or hopes to be true.
- Although mathematicians are comfortable with the idea of probability distributions, clinicians and scientists find it strange to calculate the theoretical probability distribution of the results of experiments that will never be performed.
- The derivation of the theoretical probability distributions depends on mathematics beyond the ready reach of most scientists.
- The logic goes in a direction that seems intuitively backwards. You observed a sample and want to make inferences about the population. Calculations of the P value start with an assumption about the population (the null hypothesis) and determine the probability of randomly selecting samples with as large a difference as you observed.

This book will present three methods for calculating a P value to compare a measured variable in two groups (t test, randomization test, and Mann-Whitney test). The t test is used most commonly. By plugging the data from our example into a computer program, we would learn that the P value is 0.034 (two tailed, see below).

Interpreting the P value is straightforward. If this null hypothesis were true, then 3.4% of all possible experiments of this size would result in a difference between mean blood pressures as large as (or larger) than we observed. In other words, if the null hypothesis were true, there is only a 3.4% chance of randomly selecting samples whose means are as far apart (or further) than we observed.

What conclusion should you reach? That's up to you. Statistical calculations provides the P value. You have to interpret it. Assuming that the study design was perfect, there are two possibilities: One possibility is that the two populations have different mean blood pressures. The other possibility is that the two populations are identical, and the observed difference is a coincidence of sampling. Statistical calculations determine how rare that coincidence would be. In this example we can say

that such a coincidence will occur 3.4% of the time *if* there really is no difference between populations.

Often, the P value is used to make a statement about statistical significance. This is explained in the next chapter.

OTHER NULL HYPOTHESES

When we compare the means of two groups, the null hypothesis is that the two populations have identical means. When you make other kinds of comparisons, the null hypotheses are logical:

- When you compare two proportions, the null hypothesis is that the two proportions are identical in the population.
- When you compare two survival curves, the null hypothesis is that the two survival curves are identical in the population.

COMMON MISINTERPRETATIONS OF P VALUES

P values are easy to misinterpret. The best way to avoid misinterpreting the P value is to keep firmly in mind what it does tell you: The P value is the probability of getting a difference as big (or bigger) than you got if the null hypothesis is really correct. Thus, a P value of 0.03 means that even if the two populations have identical means, 3% of experiments like the one you conducted would yield a difference at least as large as you found.

It is very tempting to jump from this and say, "Oh, well, if there is only a 3% probability that my difference would have been caused by random chance, then there must be a 97% probability that it was caused by a real difference." Wrong! What you can say is that if the null hypothesis were true, then 97% of experiments would lead to a difference smaller than the one you observed, and 3% of experiments would lead to a difference as large or larger than the one you observed.

Calculation of a P value is predicated on the assumption that the null hypothesis is correct. P values cannot tell you whether this assumption is correct. P value tells you how rarely you would observe a difference as larger or larger than the one you observed if the null hypothesis were true. The question that the scientist must answer is whether the result is so unlikely that the null hypothesis should be discarded.

ONE-TAILED VERSUS TWO-TAILED P VALUES

A two-tailed P value is the probability (assuming the null hypothesis) that random sampling would lead to a difference as large as or larger than the observed difference with either group having the larger mean. A one-tailed P value, in contrast, is the probability (assuming the null hypothesis) that random sampling would lead to a difference as large as or larger than the observed difference, and that the group specified in advance by the experimental hypothesis has the larger mean.

If the observed difference went in the direction predicted by the experimental hypothesis, the one-tailed P value is half the two-tailed P value.* The terms *one-sided* and *two-sided* P values are equivalent to one and two tails.

In the blood pressure example, I chose to use a two-tailed P value. Thus the P value answers this question: If the null hypothesis is true, what is the chance that the difference between two randomly selected samples of five subjects would have been 25 mmHg or greater with either group having the higher mean BP. A one-tailed P value is the probability (under the null hypothesis) of observing a difference of 25 mmHg or greater, with the second-year class having the larger mean value.

A one-tailed test is appropriate when previous data, physical limitations, or common sense tells you that the difference, if any, can only go in one direction. Here is an example in which you might appropriately choose a one-tailed P value: You are testing whether a new antibiotic impairs renal function, as measured by serum creatinine. Many antibiotics poison kidney cells, resulting in reduced glomerular filtration and increased serum creatinine. As far as I know, no antibiotic is known to decrease serum creatinine, and it is hard to imagine a mechanism by which an antibiotic would increase the glomerular filtration rate. Before collecting any data, you can state that there are two possibilities: Either the drug will not change the mean serum creatinine of the population, or it will increase the mean serum creatinine in the population. You consider it impossible that the drug will truly decrease mean serum creatinine of the population and plan to attribute any observed decrease to random sampling. Accordingly, it makes sense to calculate a one-tailed P value. In this example, a two-tailed P value tests the null hypothesis that the drug does not alter the creatinine level; a one-tailed P value tests the null hypothesis that the drug does not increase the creatinine level.

Statisticians disagree about when to use one- versus two-tailed P values. One extreme position is that it is almost never acceptable to report a one-tailed P value—that a one-tailed P value should be reported only when it is physically impossible for a difference to go in a certain direction. Thus a two-tailed P value ought to be used in the preceding example, because a drug-induced decrease in serum creatinine is not impossible (although it is unprecedented and unexplainable).

The other extreme position is that one-tailed P values are almost always appropriate. The argument is that formal studies (requiring formal statistical analysis) should only be performed after the investigator has proposed an experimental hypothesis based on theory and previous data. Such a hypothesis should specify which group should have the larger mean (or larger proportion, or longer survival, etc.). Indeed one could question whether taxpayer's money should be spent on research so vaguely conceived that the direction of the expected difference can't be specified in advance.

The issue here in deciding between one- and two-tailed tests is not whether or not you expect a difference to exist. If you already knew whether or not there was a difference, there is no reason to collect the data. Rather, the issue is whether the direction of a difference (if there is one) can only go one way. You should only use a one-tailed P value when you can state with certainty (and before collecting any data) that in the overall populations there either is no difference or there is a difference in a specified direction. If your data end up showing a difference in the "wrong" direction, you should be willing to attribute that difference to random sampling without even

*There are exceptions, such as Fisher's exact test.

considering the notion that the measured difference might reflect a true difference in the overall populations. If a difference in the "wrong" direction would intrigue you (even a little), you should calculate a two-tailed P value.

Two-tailed P values are used more frequently than one-tailed P values. I have chosen to use only two-tailed P values for this book for the following reasons:

- The relationship between P values and confidence intervals is more clear with two-tailed P values.
- Two-tailed P values are larger (more conservative). Since many experiments do not completely comply with all the assumptions on which the statistical calculations are based, many P values are smaller than they ought to be. Using the larger two-tailed P value partially corrects for this.
- Some tests compare three or more groups, which makes the concept of tails inappropriate (more precisely, the P value has more than two tails). A two-tailed P value is more consistent with P values reported by these tests.
- Choosing one-tailed P values can put you in awkward situations. If you decided to calculate a one-tailed P value, what would you do if you observed a large difference in the opposite direction to the experimental hypothesis? To be honest, you should state that the P value is large and you found "no significant difference." But most people would find this hard. Instead, they'd be tempted to switch to a two-tailed P value, or stick with a one-tailed P value, but change the direction of the hypothesis. You avoid this temptation by choosing two-tailed P values in the first place.

When interpreting published P values, note whether they are calculated for one or two tails. If the author didn't say, the result is somewhat ambiguous. The terms *one-sided* and *two-sided* P values mean exactly the same thing as *one-tailed* and *two-tailed* P values.

EXAMPLE 10.1. COMPARING TWO PROPORTIONS FROM AN EXPERIMENTAL STUDY

Later in the book (Part VII), you'll learn how to calculate many common statistical tests. But you can interpret P values without knowing too much about the individual tests. This example and the next four show you real examples of P values and how to interpret them.

This example uses the same data as example 8.1. This study compared disease progression in asymptomatic people infected with HIV. The relative risk was 0.57. This means that disease progressed in 57% as many treated patients as placebo-treated patients.

The null hypothesis is that AZT does not alter the probability of disease progression. We want to calculate the P value that answers this question: If the null hypothesis were true, what is the chance that random sampling of subjects would result in incidence rates different (or more so) from what we observed? The P value depends on sample size and on how far the relative risk is away from 1.0.

You can calculate the P value using two different tests: Fisher's exact test or the chi-square test. Both tests will be described in Chapter 27, but you don't need to know much about the tests to interpret the results. The chi-square test is used more often,

but only because it is easier to calculate by hand. Fisher's test calculates a more accurate P value and is preferred when computers do the calculating. The results from InStat are given in Table 10.1.

The P value is tiny, less than 0.0001. If you had chosen the chi-square test rather than Fisher's test, the P value would still have been less than 0.0001.

Interpreting the P value is straightforward: If the null hypothesis is true, there is less than a 0.01% chance of randomly picking subjects with such a large (or larger) difference in incidence rates. Note that InStat (like most statistical programs) also makes a statement about statistical significance. You'll learn what that means in the next chapter.

To interpret the P value you need to make the following assumptions:

- The subjects represent the population of all people who are infected with the HIV but have no symptoms who are or will be treated with AZT or placebo.
- Each subject was selected independently of the rest. Picking one subject should not influence the chance of picking anyone else.
- The only difference between the two groups is the treatment.

Are you ready to recommend AZT to these patients? Before deciding, remember that AZT is a toxic drug. You shouldn't base an overall conclusion from the study on one P value. Rather you should look at the benefit (which may be measured in several ways) and the risks or side effects of the treatment. You should also take into account other things you know about the AZT treatment from other studies.

EXAMPLE 10.2. COMPARING TWO PROPORTIONS FROM A CASE-CONTROL STUDY

This is a case-control study investigating the association between fleas and cat scratch fever. The results were shown in Table 9.1.

The odds ratio is 17.3. We want to calculate a P value. The null hypothesis is that there is no association between fleas and cat scratch disease, that the cats of cases are just as likely to have fleas as the cats of controls. The two-sided P value answers this question: If the null hypothesis were true, what is the chance of randomly picking subjects such that the odds ratio is 17.3 or greater or 0.058 (the reciprocal of 17.3) or lower?

To calculate a P value, you may use Fisher's exact test or the chi-square test. The same methods are used to analyze prospective and case-control studies. If a computer is doing the work, Fisher's test is your best choice. The results from InStat are shown

Table 10.1. Results from InStat Analysis of Example 10.1

Fisher's Exact Test
The two-sided P value is <0.0001, considered extremely significant.
There is a significant association between rows and columns.
Relative risk = 0.5718.
95% confidence interval: 0.4440 to 0.7363 (using the approximation of Katz).

Table 10.2. Results from InStat Analysis of Example 10.2

Fisher's Exact Test
The two-sided P value is <0.0001, considered extremely significant.
There is a significant association between rows and columns.
Odds ratio = 17.333
95% confidence interval: 5.506 to 54.563 (using the approximation of Woolf.)

in Table 10.2. Again the P value is less than 0.0001. If there were no association between fleas and cat scratch fever in the overall population, there is less than a 0.01% chance of randomly picking subjects with so much association.

To interpret the P value from a case-control study, you must accept these assumptions:

- The cases and controls are randomly selected from their respective populations, or at least are representative of those populations.
- Each subject was selected independently of the rest. Picking one subject should not influence the chance of picking anyone else.
- The controls do not differ systematically from the cases in any way, except for the absence of disease. It is very difficult to be sure you have satisfied this assumption.

EXAMPLE 10.3. COMPARING TWO MEANS WITH THE t TEST

In this study (same data as Example 7.1), the investigators compared stool output in babies treated conventionally with that in babies treated conventionally with the addition of bismuth salicylate. Control babies produced 260 ± 254 ml/kg of stool (N = 84, mean ± SD), while treated babies produced 182 ± 197 ml/kg of stool (N = 85). The difference between means is 78 ml/kg.

The null hypothesis is that the mean stool output in the overall population of babies with diarrhea treated with bismuth salicylate is equal to the mean stool output of babies treated conventionally. We need to calculate a P value that answers this question: If the null hypothesis is true, what is the probability of randomly choosing samples whose means are as different (or more different) as observed.

The unpaired t test calculates the P value. The test takes into account the size of the difference, the size of the samples, and the SD of the samples. The results from InStat are given in Table 10.3.

For this example, the two-tailed P value is 0.0269. If the drug is really ineffective, there is a 2.69% chance that two randomly selected samples of N = 84 and N = 85 would have means that are 78 ml/kg or further apart. This is a two-tailed P value, so it includes a 1.34% chance of randomly picking subjects such that the treatment reduces stool output by 78 ml/kg or more and a 1.34% of randomly picking subjects such that the treatment increases stool output by 78 ml/kg or more.

The t test is based on these familiar assumptions:

- The subjects are representative of all babies with diarrhea who have been or will be treated with oral rehydration or oral rehydration and bismuth salicylate.

Table 10.3. Results from InStat Analysis of Example 10.3

Unpaired t Test
Are the means of control and treated equal?
Mean difference = −78.000 (mean of column B minus mean of column A)
The 95% confidence interval of the difference: −146.99 to −9.013
t = 2.232 with 167 degrees of freedom.
The two-tailed P value is 0.0269, considered significant.

- The two groups are not matched or paired. (You should try to pair data whenever possible, but you need to use appropriate analyses. See Chapter 25.)
- Each subject was selected independently of the rest. Choosing one subject from the population should not influence the chance of choosing anyone else.
- The distribution of values in the overall population follows a roughly Gaussian distribution, and the two populations have identical SDs.

The last assumption presents a problem with these data. As discussed in Chapter 7, the data are very unlikely to come from a Gaussian distribution. The SD is about equal to the mean, yet stool output cannot be negative. There is no way this can happen with a Gaussian distribution. Since the sample size is so large, we can rely on the central limit theorem and ignore the assumption.

EXAMPLE 10.4. COMPARING MEANS OF PAIRED SAMPLES

This study (same as Example 7.2) measured the volume change of renal cysts when incubated in culture medium. The investigators studied nine cysts and measured the weight change (as a proxy for volume) of each. The average weight change was 0.50 grams with a standard error of 0.23.

The null hypothesis is that renal cysts are as likely to shrink as to grow in culture. More precisely, the null hypothesis is that the mean volume of renal cysts does not change in culture. We want to calculate a P value that answers this question: If the null hypothesis is true, what is the chance of randomly picking cysts with an average weight change of 0.50 grams or more? Use the paired t test to calculate the answer. The results from InStat are given in Table 10.4. The P value is 0.0614. If renal cysts on average do not change weight in culture, there is a 6.14% chance that nine randomly selected cysts would change weight as much or more than these did.

Table 10.4. Results from InStat Analysis of Example 10.4

Paired t Test for Renal Cysts Example
Does the mean change from BEFORE to AFTER equal 0.000?
Mean difference = 0.5000 (mean of paired differences)
The 95% confidence interval of the difference: −0.03038 to 1.030
t = 2.174 with 8 degrees of freedom.
The two-tailed P value is 0.0614, considered not quite significant.

EXAMPLE 10.5. COMPARING TWO SURVIVAL CURVES WITH THE LOG-RANK TEST

After a bone marrow transplant, the white cells from the donor can mount an immune response against the recipient. This condition, known as graft-versus-host disease (GVHD) can be quite serious. Immunosuppressive drugs are given in an attempt to prevent GVHD. Cyclosporine plus prednisone are commonly used. Chao et al.* compared those two drugs alone (two drugs) with those two drugs plus methotrexate (three drugs). Figure 10.3 shows their data.

They recorded time until GVHD and plotted the data as Kaplan-Meier survival curves. Note that the term *survival curve* is a bit misleading as the end point is onset of GVHD rather than death. The authors chose to plot the data going "uphill," showing the percent of patients who had GVHD by a certain time. Most investigators would have plotted these Kaplan-Meier curves going "downhill," showing the percentage of patients who had NOT had GVHD by a certain time. This is an artistic decision that does not affect the analyses.

The authors compared the two groups with the log-rank test (other investigators might use the Mantel-Haenszel test, which is nearly identical). You will learn a little bit about these tests in Chapter 33. Even though you have not yet learned about the test, you can understand the results.

As you would expect, the null hypothesis is that the two populations have identical survival curves and that the observed difference in our samples is due to chance. In other words, the null hypothesis is that treatment with three drugs is no better (and no worse) than treatment with two drugs in preventing GVHD. The P value from the log-rank test answers this question: If the null hypothesis is true, what is the probability of obtaining such different survival curves with randomly selected subjects?

The calculations are quite involved and should be left to computer programs. The authors report that the P value is 0.02. If addition of the third drug did not alter onset of GVHD, you'd see such a large difference in survival curves in only 2% of experiments of this size.

You've already learned about the assumptions that must be true to interpret a survival curve in Chapter 6. The log-rank test depends on those same assumptions (reviewed below):

- The subjects are representative of all bone-marrow transplant recipients.
- The subjects were chosen independently.
- Consistent criteria. The entry criteria and the definition of *survival* must be consistent during the course of the study.
- The survival of the censored subjects would be the same, on average, as the survival of the remaining subjects.

The data in Figure 10.3 show only one end point—the time to onset of GVHD. The investigators also compared other variables, such as renal function, liver function,

*NJ Chao, GM Schmidt, JC Niland et al. Cyclosporine, methotrexate and prednisone compared with cyclosporine and prednisone for prophylaxis of acute graft-versus-host disease. *N Engl J Med* 329:1225–1230, 1993.

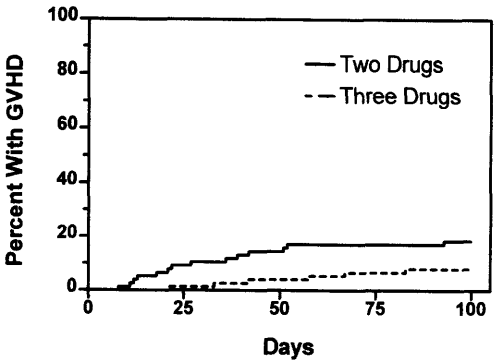


Figure 10.3. Data for Example 10.5. The investigators compared treating bone marrow transplants with two drugs or three drugs. The X axis shows time after the transplant, while the Y axis shows the percent of patients who have graft-versus-host disease (GVHD).

survival time, and recurrence of leukemia. Analysis of all these data suggest that addition of the third drug helps prevent GVHD without causing other problems. The authors recommendation to treat with three-drug therapy is based on analysis of all these variables, not just the survival curves.

SUMMARY

Most statistical tests calculate a P value. Even if you don't learn the details of all the tests it is essential that anyone reading the biomedical literature understand what a P value is (and what it isn't). A P value is simply a probability that answers the following question: If the null hypothesis were true (i.e., there is no difference between populations), what is the probability that random sampling (given the sample size actually used) would result in a difference as big or bigger than the one observed?

OBJECTIVES

1. You must be familiar with the following terms:
   - P value
   - Null hypothesis
   - One tailed test
   - Two tailed test
2. Whenever you see a P value reported in the biomedical literature, you should be able to identify the null hypothesis (even if you are not familiar with the statistical test being used) and pose the question that the P value answers.
3. You should know the difference between one- and two-tailed P values and know the arguments for and against using each.
4. Whenever you see a P value, you should be able to state the question it answers.

PROBLEMS

1. You wish to test the hypothesis that a coin toss is fair. You toss the coin six times and it lands on heads each time. What is the null hypothesis? What is the P value? Is a one- or two-tailed test more appropriate? What do you conclude?
2. You conduct an experiment and calculate that the two-tailed P value is 0.08. What is the one-tailed P value? What assumptions are you making?
3. (Same data as Problem 4 in Chapter 5 and Problem 1 in Chapter 7.) The serum levels of a hormone (the Y factor) was measured to be 93 ± 1.5 (mean ± SEM) in 100 nonpregnant women and 110 ± 2.3 (mean ± SEM) in 100 women in the first trimester of pregnancy. When the two groups are compared with a t test, the P value is less than 0.0001.
   A. Explain what the P value means in plain language.
   B. Review your answers to Problem 4 in Chapter 5 and Problem 1 in Chapter 7. Explain how the results complement one another.
4. (Same data as Problem 4 in Chapter 8.) Cohen et al. investigated the use of active cardiopulmonary resuscitation (CPR).* In standard CPR the resuscitator compresses the victim's chest to force the heart to pump blood to the brain (and elsewhere) and then lets go to let the chest expand. Active CPR is done with a suction device. This enables the resuscitator to pull up on the chest to expand it as well as pressing down to compress it. These investigators randomly assigned cardiac arrest patients to receive either standard or active CPR. Eighteen of 29 patients treated with active CPR were resuscitated. In contrast, 10 of 33 patients treated with standard CPR were resuscitated. Using Fisher's test, the two-sided P value is 0.0207. Explain what this means in plain language.

*TJ Cohen, BG Goldner, PC Maccaro, AP Ardito, S Trazzera, MB Cohen, SR Dibs. A comparison of active compression-decompression cardiopulmonary resuscitation with standard cardiopulmonary resuscitation for cardiac arrests occurring in the hospital. *N Engl J Med* 329:1918–1921, 1993.

11

Statistical Significance and Hypothesis Testing

In the previous chapter we treated the P value as a number and left the interpretation to the readers. Looked at this way, calculating a P value is analogous to calculating an average, ratio, or percentage. It is a useful way to summarize data to aid understanding and communication. This chapter shows how to make statements about statistical significance.

STATISTICAL HYPOTHESIS TESTING

When interpreting many kinds of data, you need to reach a decision. In a pilot experiment of a new drug, you need to decide whether the results are promising enough to merit a second experiment. In a phase III drug study, you need to decide whether the drug is effective and should be recommended for all patients. In a study comparing two surgical procedures, you need to decide which procedure to recommend.

Hypothesis testing is an approach that helps you make decisions after analyzing data. Follow these steps:

1. Assume that your samples are randomly selected from the population.
2. State the null hypothesis that the distribution of values in the two populations is the same.
3. Define a threshold value for declaring a P value significant. This threshold called the *significance level* of the test is denoted by α and is commonly set to 0.05.
4. Select an appropriate statistical test and calculate the P value.
5. If the P value is less than α, then conclude that the difference is *statistically significant* and decide to *reject the null hypothesis.* Otherwise conclude that the difference is not statistically significant and decide to not reject the null hypothesis.

Note that statisticians use the term *hypothesis testing* quite differently than scientists. Testing scientific hypotheses requires hard work involving many kinds of experiments. To test a new hypothesis, it is often necessary to design new experimental methodology

and to design clever control experiments. Statistical hypothesis testing, in contrast, is easy. Just check whether one P value is above or below a threshold.

Converting a P value to the conclusion "significant" or "not significant" reminds me of the movie reviewers Siskel and Ebert. In their written reviews, they rate each movie on a scale (i.e., three and a half stars or B−). This is analogous to a P value. It is a concise way to summarize their opinions about the movie. When reviewing movies on television, they make a decision for you: see it (thumbs up) or don't see it (thumbs down).

The terminology of hypothesis testing is easiest to understand in the context of quality control. For example, let's assume that you run a brewery and that you have a warehouse full of the latest batch of beer. Before selling this batch, you need to test whether the batch meets various quality standards. Rather than test the entire warehouse full of beer (the population), you randomly choose some bottles (the sample) to compare to a "gold standard." The results can be expressed as a P value that answers the following question: If the new batch of beer is identical to the standard batch, what is the probability that a randomly selected sample of bottles would be as different from the standard as actually observed? If the P value is less than α (usually 0.05), then you reject the null hypothesis and reject the batch of beer (or at least do further tests to find out what is wrong). If the P value is greater than α, you do not reject the null hypothesis and do not reject the batch.

THE ADVANTAGES AND DISADVANTAGES OF USING THE PHRASE *STATISTICALLY SIGNIFICANT*

There are three advantages to using the phrase *statistically significant:*

- In some situations, it is necessary to reach a crisp decision from one experiment. Make one decision if the results are significant and the other decision if the results are not significant.
- With some statistical tests, it is difficult or impossible to obtain the exact P value, but it is possible to determine whether or not the P value exceeds α.
- People don't like ambiguity. The conclusion that "the results are statistically significant" is much more satisfying than the conclusion that "random sampling would create a difference this big or bigger in 3% of experiments if the null hypothesis were true."

The disadvantage to using the phrase *statistically significant* is that many people misinterpret it. Once some people have heard the word *significant,* they stop thinking about the data.

In biology and clinical medicine, it is not always necessary to reach a crisp decision from each P value. Instead, you can often base important decisions on several kinds of data, perhaps from several studies. Often, the best conclusion is wait and see. If it is not necessary to make a crisp decision from one experiment, there is no need to summarize the findings as being significant or not significant. Instead, just report the P value.

AN ANALOGY: INNOCENT UNTIL PROVEN GUILTY

Table 11.1 shows the analogy between the steps that a jury must follow to determine guilt (at least in the United States) and the steps that a scientist follows to determine statistical significance.

A jury reaches the verdict of guilty when the evidence is inconsistent with the assumption of innocence. Otherwise the jury reaches a verdict of not guilty. Note that a jury can never reach the verdict of innocent. The only choices are guilty or not guilty ("not proven" in Britain). A jury reaches a verdict of not guilty when the evidence is not inconsistent with the presumption of innocence. A jury does not have to be convinced that the defendant is innocent to reach a verdict of not guilty.

When performing a statistical test, you never conclude that you accept the null hypothesis. You reach the conclusion not significant whenever it is plausible that the data are consistent with the null hypothesis. A not significant result does not mean that the null hypothesis is true.

Journalists also evaluate evidence at a trial, and have different goals than jurors. As a journalist, you don't have to reach a verdict of guilty or not guilty. Instead your job is to summarize the proceedings. You want to present enough evidence so the readers can reach their own conclusions. You may include evidence unavailable to the jurors (evidence that wasn't presented at trial, or evidence that was ruled inadmissible). As a journalist, you don't have to reach any conclusion about the guilt of the defendant. If you do reach a conclusion, it is appropriate for it to be somewhat hazy if you think the evidence was inconclusive.

Table 11.1. Significant vs. Not Significant Decision Compared with Guilty vs. Not Guilty Decisions

Step	Trial by Jury	Statistical Significance
1.	Start with the presumption that the defendant is innocent.	Start with the presumption that the null hypothesis is true.
2.	Listen to factual evidence presented in the trial. Don't consider other data, such as newspaper stories you have read.	Base your conclusion only on data from this one experiment. Don't consider whether the hypothesis makes scientific or clinical sense. Don't consider any other data.
3.	Evaluate whether you believe the witnesses. Ignore testimony from witnesses who you think have lied.	Evaluate whether the experiment was performed correctly. Ignore flawed data.
4.	Think about whether the evidence is consistent with the assumption of innocence.	Calculate a P value.
5.	If the evidence is inconsistent with the assumption, then reject the assumption of innocence and declare the defendant to be guilty. Otherwise, reach a verdict of not guilty. As a juror, you can't conclude "maybe," can't ask for additional evidence, and can't say "wait and see."	If the P value is less than a preset threshold (typically 0.05), conclude that the data are inconsistent with the null hypothesis. Reject the null hypothesis, and declare the difference to be statistically significant. Otherwise, conclude that the difference is not significant. You can't conclude "maybe" or "wait and see."

I think that many scientists find themselves in the role of a journalist more often than they find themselves in the role of a juror. A scientist does not always need to evaluate data using the approach of statistical hypothesis testing. When presenting data, the most important thing is to present the methods and results clearly so the readers can reach their own conclusions. In many scientific situations, it is not necessary to reach a crisp conclusion of significant or not significant. Instead, it is sometimes appropriate to reach a more general conclusion that might be somewhat uncertain. And when reaching that conclusion, it is appropriate to consolidate data from several experiments and to take into account whether the hypothesis being tested fits theory and previous data.

TYPE I AND TYPE II ERRORS

When you conclude that results are significant or not significant, there are two ways you can be wrong:

- You can find that a result is statistically significant and reject the null hypothesis when in fact the null hypothesis is true. This is a Type I error. Convicting an innocent person is a Type I error. The probability of making a Type I error is α.
- You find that a result is not statistically significant and fail to reject the null hypothesis when the null hypothesis is in fact false. This is a Type II error. Failing to convict a guilty person is a Type II error.

We'll return to Type I and Type II errors in the next chapter.

CHOOSING AN APPROPRIATE VALUE FOR α

By tradition, α is usually set to equal 0.05. This cutoff is purely arbitrary, but is entrenched in the literature. It is amazing that scientists (who disagree about so much) all agree on such an arbitrary value! Ideally, the value of α should not be set by tradition but rather by the context of the experiment.

If you set α to a very low value, you will make few Type I errors. That means that if the null hypothesis is true, there is only a small chance that you will mistakenly call a result significant. However, there is also a larger chance that you will not find a significant difference even if the null hypothesis is false. In other words, reducing the value of α will decrease your chance of making a Type I error but increase the chance of a Type II error.

If you set α to a very large value, you will make many Type I errors. If the null hypothesis is true, there is a large chance that you will mistakenly find a significant difference. But there is a small chance of missing a real difference. In other words, increasing the value of α will increase your chance of making a Type I error but decrease the chance of a Type II error. The only way to reduce the chances of both a Type I error and a Type II error is to collect bigger samples. See Chapter 22.

You must balance the costs or consequences of Type I and Type II errors, and alter the value of α accordingly. Let's look at three examples:

- You are screening a new drug in an *in vitro* assay. If the experiment yields significant results, you will investigate the drug further. Otherwise you will stop investigating the drug. In this circumstance, the cost of a Type I error is a few additional experiments, and the cost of a Type II error is abandoning an effective drug. It makes sense to set α high to minimize Type II errors, even at the expense of additional Type I errors. I'd set α to 0.10 or 0.20 in this situation.
- You are conducting a phase III clinical trial of that drug to treat a disease (like hypertension) for which there are good existing therapies. If the results are significant, you will market the drug. If the results are not significant, work on the drug will cease. In this case, a Type I error results in treating future patients with a useless drug and depriving them of a good standard drug. A Type II error results in aborting development of a good drug for a condition that can be treated adequately with existing drugs. Thinking scientifically (not commercially), you want to set a low to minimize the risk of a Type I error, even at the expense of a higher chance of a Type II error. I'd set α to 0.01 in this situation.
- You are conducting a phase III clinical trial of that drug to treat a disease for which there is no good existing therapy. If the results are significant, you will market the drug. If the results are not significant, work on the drug will cease. In this case, a Type I error results in treating future patients with a useless drug instead of nothing. A Type II error results in cancelling development of a good drug for a condition that is currently not treatable. Here you want to set α high value because a Type I error isn't so bad but a Type II error would be awful. I'd set α to 0.10.

In this chapter, I've been deliberately vague when talking about the chances of a Type II error, and have not told you how to calculate that probability. You'll learn how to do that in Chapters 23 and 27.

Let's continue the analogy between statistical significance and the legal system. The balance between Type I and Type II errors depends on the type of trial. In the United States (and many other countries) a defendant in a criminal trial is considered innocent until proven guilty "beyond a reasonable doubt." This system is based on the belief that it is better to let many guilty people go free than to falsely convict one innocent person. The system is designed to avoid Type I errors in criminal trials, even at the expense of many Type II errors. You could say that α is set to a very low value. In civil trials, the court or jury finds for the plaintiff if the evidence shows that the plaintiff is "more likely than not" to be right. The thinking is that it is no worse to falsely find for the plaintiff than to falsely find for the defendant. The system attempts to equalize the chances of Type I and Type II errors in civil trials.

THE RELATIONSHIP BETWEEN α AND P VALUES

The P value and α are closely related. You calculate a P value from particular data. You set α in advance, based on the consequences of Type I and Type II errors. α is the threshold P value below which a difference is termed *statistically significant.*

THE RELATIONSHIP BETWEEN α AND CONFIDENCE INTERVALS

Although I have presented confidence intervals (CIs) and P values in separate sections of this book, the two are closely related. They both are based on the same assumptions and the same statistical principles. Ask yourself whether the 95% CI includes the value stated in the null hypothesis. If you are comparing two means, ask whether the 95% CI for the difference between the means includes 0. If you are analyzing a prospective study, ask whether the 95% CI for the relative risk includes 1.0. If the 95% CI includes the value stated in the null hypothesis, then the P value must be greater than 0.05 and the difference must be not significant. If the 95% CI does not include the null hypothesis, then the P value must be less than 0.05 and the result must be significant.

The rule in the previous paragraph works because 95% (CI) and 5% (threshold P value) add to 100%. Other pairs can be used as well. If the 99% CI includes the null hypothesis, then you can be sure that the P value must be greater than 0.01. If the 90% CI includes the null hypothesis, then you can be sure that the P value must be greater than 0.10.

STATISTICAL SIGNIFICANCE VERSUS SCIENTIFIC IMPORTANCE

A result is said to be statistically significant when the P value is less than a preset value of α. This means that the results are surprising and would not commonly occur if the null hypothesis were true. There are three possibilities:

- The null hypothesis is really true, and you observed the difference by coincidence. The P value tells you how rare the coincidence would be. If the null hypothesis really is true, α is the probability that you will happen to pick samples that result in a statistically significant result.
- The null hypothesis is really false (the populations are really different), and the difference is scientifically or clinically important.
- The null hypothesis is really false (the populations are really different), but the difference is so small as to be scientifically or clinically trivial.

The decision about scientific or clinical importance requires looking at the size of the difference (or the size of the relative risk or odds ratio). With large samples, even very small differences will be statistically significant. Even if these differences reflect true differences between the populations, they may not be interesting. You must interpret scientific or clinical importance by thinking about biology or medicine. For example, few would find a mean difference of 1 mmHg in blood pressure to be clinically interesting, no matter how low the P value. It is never enough to think about P values and significance. You must also think scientifically about the size of the difference.

SUMMARY

After calculating a P value, you "test a hypothesis" by comparing the P value with an arbitrary value α, usually set to 0.05. If the P value is less than α, the result is said

to be "statistically significant" and the null hypothesis is said to be "rejected." If the P value is greater than α, the result is said to be "not statistically significant" and the null hypothesis is "not rejected." As used in the context of statistical hypothesis testing, the word *significant* has almost no relationship to the conventional meaning of the word to denote importance. Statistically significant results can be trivial. Differences that might turn out to be enormously important may result in "nonsignificant" P values. The scheme of hypothesis testing is useful in situations in which you must make a crisp decision from one experiment. In many biological or clinical studies, it is not always necessary to reach a crisp decision and it makes more sense to report the exact P value and avoid the term *significant.*

OBJECTIVES

1. You must be familiar with the following terms:
   - Hypothesis testing
   - Null hypothesis
   - α
   - Statistically significant
   - Type I error
   - Type II error
2. You must know the distinction between the statistical meaning of the terms *significant* and *hypothesis testing,* and the scientific meaning of the same terms.
3. You must know what *not significant* really means.

PROBLEMS

1. For Problems 3 and 4 of the last chapter, explain what a Type I and Type II error would mean.
2. Which of the following are inconsistent?
   A. Mean difference = 10. The 95% CI: -20 to 40. P = 0.45
   B. Mean difference = 10. The 95% CI: -5 to 15. P = 0.02
   C. Relative risk = 1.56. The 95% CI: 1.23 to 1.89. P = 0.013
   D. Relative risk = 1.56. The 95% CI: 0.81 to 2.12. P = 0.04
   E. Relative risk = 2.03. The 95% CI: 1.01 to 3.13. P < 0.001.

12

Interpreting Significant and Not Significant P Values

THE TERM *SIGNIFICANT*

You've already learned that the term *statistically significant* has a simple meaning: The P value is less than a preset threshold value, α. That's it! In plain language, a result is statistically "significant" when the result would be surprising if there really were no differences between the overall populations.

It's easy to read far too much into the word *significant* because the statistical use of the word has a meaning entirely distinct from its usual meaning. Just because a difference is *statistically significant* does not mean that it is important or interesting. A statistically significant result may not be scientifically significant or clinically significant. And a difference that is not significant (in the first experiment) may turn out to be very important. This is important, so I'll repeat it: *Statistically significant results are not necessarily important or even interesting.*

EXTREMELY SIGNIFICANT RESULTS

Intuitively, you'd think that $P = 0.004$ is more significant than $P = 0.04$. Using the strict definitions of the terms, this is not correct. Once you have set a value for α, a result is either *significant* or *not significant.* It doesn't matter whether the P value is very close to α or far away. Many statisticians feel strongly about this, and think that the word *significant* should never be prefaced by an adjective. Most scientists are less rigid, and refer to *very significant* or *extremely significant* results when the P value is tiny.

When showing P values on graphs, investigators commonly use a "Michelin Guide" scale. $*P < 0.05$ (significant), $**P < 0.01$ (highly significant); $***P < 0.001$ (extremely significant). When you read this kind of graph, make sure that you look at the key that defines the symbols, as different investigators use different threshold values.

BORDERLINE P VALUES

If you follow the strict paradigm of statistical hypothesis testing and set α to its conventional value of 0.05, then a P value of 0.049 denotes a significant difference and a P value of 0.051 denotes a not significant difference. This arbitrary distinction is unavoidable since the whole point of using the term *statistically significant* is to reach a crisp conclusion from every experiment without exception.

Rather than just looking at whether the result is significant or not, it is better to look at the actual P value. That way you'll know whether the P value is near α or far from it. When a P value is just slightly greater than α, some scientists refer to the result as *marginally significant* or *almost significant.*

When the two-tailed P value is between 0.05 and 0.10, it is tempting to switch to a one-tailed P value. The one-tailed P value is half the two-tailed P value and is less than 0.05, and the results become "significant" as if by magic. Obviously, this is not an appropriate reason to choose a one-tailed P value! The choice should be made before the data are collected.

One way to deal with borderline P values would be to choose between three decisions rather than two. Rather than decide whether a difference is significant or not significant, add a middle category of *inconclusive.* This approach is not commonly used.

THE TERM *NOT SIGNIFICANT*

If the P value is greater than a preset value of α, the difference is said to be *not significant.* This means that the data are not strong enough to persuade you to reject the null hypothesis. People often mistakenly interpret a high P value as proof that the null hypothesis is true. That is incorrect. A high P value does not prove the null hypothesis. This is an important point worth repeating: *A high P value does not prove the null hypothesis.* As you've already learned in Chapter 11, concluding that a difference is not statistically significant when the null hypothesis is, in fact, false is called a Type II error.

When you read that a result is not significant, don't stop thinking. There are two approaches you can use to evaluate the study. First, look at the confidence interval (CI). Second, ask about the power of the study to find a significant difference if it were there.

INTERPRETING *NOT SIGNIFICANT* RESULTS WITH CONFIDENCE INTERVALS

Example 12.1

Ewigman et al. investigated whether routine use of prenatal ultrasound would improve perinatal outcome.* They randomly divided a large group of pregnant women into two

*BG Ewigman, JP Crane, FD Frigoletto et al. Effect of prenatal ultrasound screening on perinatal outcome. *N Engl J Med* 329:821–827, 1993.

groups. One group received routine ultrasound sonogram exams twice during the pregnancy. The other group received sonograms only if there was a clinical reason to do so. The physicians caring for the women knew the results of the sonograms and cared for the patients accordingly. The investigators looked at several outcomes. Table 12.1 shows the total number of adverse events, defined as fetal or neonatal deaths of moderate to severe morbidity.

The null hypothesis is that the rate of adverse outcomes is identical in the two groups. In other words, the null hypothesis is that routine use of ultrasound neither prevents nor causes perinatal mortality or morbidity. The two-tailed P value is 0.86. The data provide no reason to reject the null hypothesis.

Before you can interpret the results, you need to know more. You need to know the 95% CI for the relative risk. For this study, the relative risk is 1.02, with the 95% CI ranging from 0.88 to 1.17. The null hypothesis can be restated as follows: In the entire population, the relative risk is 1.00. The data are consistent with the null hypothesis. This does not mean that the null hypothesis is true. Our CI tells us that the data are also consistent (within 95% confidence) with relative risks ranging from 0.88 to 1.17.

Different people might interpret these data in two ways:

> The confidence interval is very narrow, and is centered close to 1.0. These data convince me that routine use of ultrasound is neither helpful nor harmful. To reduce costs, I'll use ultrasound only when there is an identified problem.

> The confidence interval is narrow, but not all that narrow. There have been plenty of studies showing that ultrasound doesn't hurt the fetus, so I'll ignore the part of the confidence interval above 1.00. But ultrasound gives the obstetrician extra information to manage the pregnancy, and it makes sense that using this extra information will decrease the chance of a major problem. The confidence interval goes down to 0.88, a reduction of risk of 12%. If I were pregnant, I'd certainly want to use a risk-free technique that reduces the risk of a sick or dead baby by as much as 12%! The data certainly don't prove that routine ultrasound is beneficial, but the study leaves open the possibility that routine use of ultrasound might reduce the rate of truly awful events by as much as 12%. I think the study is inconclusive. Since ultrasound is not prohibitively expensive and appears to have no risk, I will keep using it routinely.

To interpret a not significant P value, you must look at the CI. If the entire span of the CI contains differences (or relative risks) that you consider to be trivial, then you can make a strong negative conclusion. If the CI is wide enough to include values you consider to be clinically or scientifically important, then the study is inconclusive. Different people will appropriately have different opinions about how large a difference

Table 12.1. Results of Example 12.1

	Adverse Outcome	Healthy Baby	Total
Routine sonograms	383	7,302	7,685
Sonograms only when indicated	373	7,223	7,596
Total	756	14,525	15,281

(or relative risk) is scientifically or clinically important and may interpret the same not significant study differently.

In interpreting the results of this example, you also need to think about benefits and risks that don't show up as a reduction of adverse outcomes. The ultrasound picture helps reassure parents that their baby is developing normally and gives them a picture to bond with and to show relatives. This can be valuable regardless of whether it reduces the chance of adverse outcomes. Although statistical analyses focus on one outcome at a time, you must consider all the outcomes when evaluating the results.

INTERPRETING *NOT SIGNIFICANT* P VALUES USING POWER ANALYSES*

A not significant P value does not mean that the null hypothesis is true. It simply means that your data are not strong enough to convince you that the null hypothesis is not true. As you have already learned, obtaining a not significant result when the null hypothesis is really false is called a Type II error. When you obtain (or read about) a not significant P value, you should ask yourself this question: "What is the chance that this is a Type II error?" That question can only be answered if you specify an alternative hypothesis—a difference Δ (or relative risk R) that you hypothesize exists in the overall population. Then you can ask, "If the true difference is Δ (or the true relative risk is R), what is the chance of obtaining a significant result in an experiment of this size? The answer is termed the *power* of the study.

Table 12.2 shows the power of example 12.1 for various hypothetical relative risks. This table uses the sample size of the example and sets $\alpha = 0.05$ and the risk in the control subjects to 5.0%. Calculating the power exactly is quite difficult, but there are several ways to calculate power approximately. You'll learn about the approximation used to create this table in Chapter 27. In Chapter 23 you'll learn how to calculate a similar table for studies that compare two means (rather than two proportions).

If the experimental treatment (routine ultrasound) reduced the risk by 25% (so the relative risk = 0.75), the study had a 97% power to detect a significant difference. If the treatment was really that good, then 97% of studies this size would wind up with a significant result, while 4% would come up with a not significant result. In contrast, the power of this study to detect a risk reduction of 5% (relative risk = 0.95) is only 11%. If the treatment truly reduced risk by 5%, only 11% of studies this size would come up with a significant result.

The numbers are particular to this study. The principle is universal. All studies have very little power to detect tiny differences and enormous power to detect large differences. If you increase the number of subjects, you will increase the power.

As you can see, calculations of power can aid interpretation of nonsignificant results. When reading biomedical research, however, you'll rarely encounter power calculations in papers that present not significant results. This is partly a matter of tradition, and it is partly because it is difficult to define the smallest difference or relative risk that you think it is important.

*This section is more advanced than the rest. You may skip it without loss of continuity.

Table 12.2. A Power Analysis of Example 12.1

Relative Risk	Power
0.95	11%
0.90	30%
0.85	60%
0.80	84%
0.75	97%

SUMMARY

A result is statistically significant when the P value is less than the preset value of α. This means that the results would be surprising if the null hypothesis were really true. The statistical use of the term *significant* is quite different than the usual use of the word. Statistically significant results may or may not be scientifically or clinically interesting or important.

When results are statistically not significant it means that the results are not inconsistent with the null hypothesis. This does *not* mean that the null hypothesis is true. When interpreting results that are not significant, it helps to look at the extent of the CI and to calculate the power that the study would have found a significant result if the populations really were different (with a difference of a defined size).

IV

BAYESIAN LOGIC

When you interpret the results of an experiment, you need to consider more than just the P value. You also need to consider the experimental context, previous data, and theory. Bayesian logic allows you to integrate the current experimental data with what you knew before the experiment.

Since Bayesian logic can be difficult to understand in the context of interpreting P values, I first present the use of Bayesian logic in interpreting the results of clinical laboratory tests in Chapter 14. Then in Chapter 16 I explain how Bayesian logic is used in interpreting genetic data.

14

Interpreting Lab Tests: Introduction to Bayesian Thinking

Note to basic scientists: Don't skip this chapter because it appears to be too clinical. This chapter sets the stage for the discussion in the next two chapters.

What do laboratory tests have to do with P values? Understanding how to interpret "positive" and "negative" lab tests will help you understand how to interpret "significant" and "not significant" statistical tests.

THE ACCURACY OF A QUALITATIVE LAB TEST

We will consider first a test that yields a simple answer: positive or negative. Results can be tabulated on a two by two contingency table (Table 14.1). The rows represent the outcome of the test (positive or negative), and the columns indicate whether the disease is present or absent (based upon some other method that is completely accurate, perhaps the test of time). If the test is "positive," it may be true positive (TP), or it may be a false positive (FP) test in a person without the condition being tested for. If the test is "negative," it may be a true negative (TN) or it may be a false negative (FN) test in a person who does have the condition.

How accurate is the test? It is impossible to express the accuracy in one number. It takes at least two: sensitivity and specificity. An ideal test has very high sensitivity and very high specificity:

- The *sensitivity* is the fraction of all those with the disease who get a positive test result.
- The *specificity* is the fraction of those without the disease who get a negative test result.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \qquad \text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{14.1}$$

It is easy to mix up sensitivity and specificity. Sensitivity measures how well the test identifies those with the disease, that is, how sensitive it is. If a test has a high sensitivity, it will pick up nearly everyone with the disease. Specificity measures how well the test excludes those who don't have the disease, that is, how specific it is. If a test has a very high specificity, it won't mistakenly give a positive result to many people without the disease.

Table 14.1. Accuracy of a Qualitative Lab Test

	Disease Present	Disease Absent
Test positive	TP	FP
Test negative	FN	TN

Sackett and colleagues have published a clever way to remember the difference between sensitivity and specificity.* Remember the meaning of sensitivity with this acronym SnNOut: If a test has high *sen*sitivity, a *n*egative test rules *out* the disorder (relatively few negative tests are false negative). Remember the meaning of specificity with this acronym: SpPln. If a test has high *sp*ecificity, a *p*ositive test rules *in* the disorder (relatively few positive tests are false positive).

THE ACCURACY OF A QUANTITATIVE LAB TEST

Many lab tests report results on a continuous scale. For the purposes of this chapter, we will simplify things so that it reports either "normal" or "abnormal." Figures 14.1 and 14.2 show the distribution of test results in patients with the condition being tested (dashed curve) and in those without the condition (solid curve). In Figure 14.1, the two distributions don't overlap, and it is easy to pick the cutoff value shown by the vertical dotted line. Every individual whose value is below that value (the left of the dotted line) does not have the condition, and every individual whose test value is above that threshold has it.

Figure 14.2 shows a more complicated situation where the two distributions overlap. Again, the solid curve represents patients without the condition and the dashed curve represents patients with the condition. Wherever you set the cutoff, some patients will be misclassified. The dark shaded area shows false positives, those patients classified as positive even though they don't have the disease. The lighter shaded area shows false negatives, those patients classified as negatives even though they really do have the disease.

Choosing a threshold requires you to make a tradeoff between sensitivity and specificity. If you increase the threshold (move the dotted line to the right), you will increase the specificity but decrease the sensitivity. You have fewer false positives but more false negatives. If you decrease the threshold (move the dotted line to the left), you will increase the sensitivity but decrease the specificity. You will have more false positives but fewer false negatives.

Choosing an appropriate threshold requires knowing the consequences (harm, cost) of false-positive and false-negative test results. Consider a screening test for a disease that is fatal if untreated but completely treatable. If the screening test is positive, it is followed by a more expensive test that is completely accurate and without risk.

*DL Sackett, RB Haynes, GH Guyatt, P Tugwell. *Clinical Epidemiology. A Basic Science for Clinical Medicine*, 2nd ed. Boston, Little Brown, 1991.

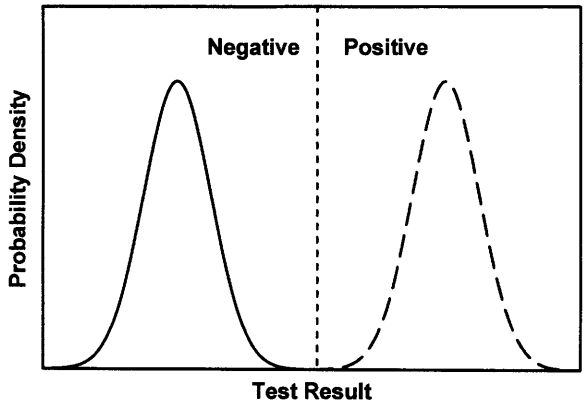


Figure 14.1. A perfect test. The solid line shows the distribution of values in people without the disease being tested for. The dashed curve shows the distribution of values in people with the disease. The vertical line shows the demarcation between normal and abnormal results. The two distributions do not overlap.

For this screening test you want to set the sensistivity very high, even at the expense of a low specificity. This ensures that you will have few false negatives but many false positives. That's OK. False positives aren't so bad, they just result in a need for a more expensive and more accurate (but safe) test. False-negative tests would be awful, as it means missing a case of a treatable fatal disease. Now let's consider another example, a screening test for an incurable noncommunicable disease. Here you want to avoid false positives (falsely telling a healthy person that she will die soon), while

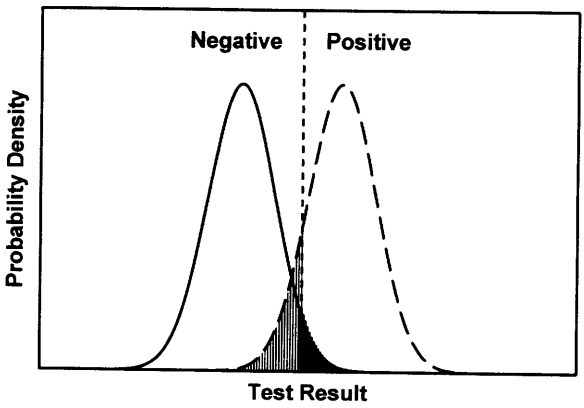


Figure 14.2. A typical test. As in Figure 14.1, the solid line shows the distribution of values in people without the disease, and the dashed curve shows the distribution of values in people with the disease. The two distributions overlap, so it is not obvious where to draw the line between negative and positive results. Our decision is shown as the dotted line. False-positive results are shown in the solid region: These are individuals without the disease who have a positive test result. False negatives are shown as the shaded area: These are individuals with the disease who have a negative test result.

false negatives aren't so bad (since you can't treat the disease anyway). In such a case, you would want the specificity to be high, even at the expense of a low sensitivity. It is impossible to make generalizations about the relative consequences of false-positive and false-negative tests, and so it is impossible to make generalizations about where to set the trade-off between sensitivity and specificity. Since the consequences of false-positive and false-negative tests are usually not directly comparable, value judgments are needed and different people will appropriately reach different conclusions regarding the best threshold value.

THE PREDICTIVE VALUE OF A TEST RESULT

Neither the specificity nor sensitivity answer the most important questions: If the test is positive, what is the chance that the patient really has the disease? If the test is negative, what is the chance that the patient really doesn't have the disease? The answers to those questions are quantified by the positive predictive value and negative predictive value:

$$\text{Positive predictive value} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

$$\text{Negative predictive value} = \frac{\text{TN}}{\text{TN} + \text{FN}}. \tag{14.2}$$

The sensitivity and specificity are properties of the test. In contrast, the positive predictive value and negative predictive value are determined by the characteristics of the test and the prevalence of the disease in the population being studied. The lower the prevalence of the disease, the lower the ratio of true positives to false positives.

Look back at Figure 14.2. The two curves have equal areas, implying that there are equal numbers of tested patients with the condition and without the condition. In other words, Figure 14.2 assumes that the prevalence is 50%. In Figure 14.3, the

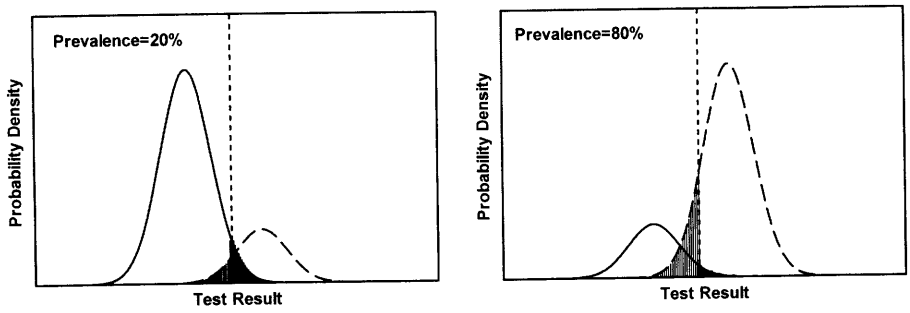


Figure 14.3. The effect of prevalence on the predictive value of the test. In Figures 14.1 and 14.2 the two curves had equal area, implying that half the tested population had the disease. The left half of this figure shows the results if the prevalence is 20%, while the right half shows the results if the prevalence is 80%. The fraction of all positive tests that are false positives is higher in the left panel. The fraction of all negative tests that are false negatives is higher in the right panel.

prevalence is changed to 20% (left panel) or 80% (right panel). As before, the solid curve represents people without the condition, and the dashed curve represents people with the condition. Any test result to the right of the dotted line is considered to be positive. This includes a portion of the area under the dashed curve (true positives) and a portion of the area under the solid curve (false positives).

The predictive values depend on the prevalence of the condition. A much larger fraction of positive tests are false positives in the left panel than in the right panel, so the predictive value of a positive test is lower in the left panel. Conversely, a much lower fraction of negative tests are false negative, so the predictive value of a negative test is much higher in the left panel.

CALCULATING THE PREDICTIVE VALUE OF A POSITIVE OR NEGATIVE TEST

Acute intermittent porphyria is an autosomal dominant disease that is difficult to diagnose clinically. It can be diagnosed by reduced levels of porphobilinogen deaminase. But the levels of the enzyme vary in both the normal population and in patients with porphyria, so the test does not lead to an exact diagnosis.

The sensitivity and specificity have been tabulated for various enzyme activities. Published data show that 82% of patients with porphyria have levels < 99 units (sensitivity = 82%) and that 3.7% of normal people have levels < 99 units (specificity = 100 – 3.7% = 96.3%). What is the likelihood that a patient with 98 units of enzyme activity has porphyria? The answer depends on who the patient is. We'll work through three examples.

Patient A

In this example, the test was done as a population screening test. Patient A has no particular risk for the disease. Porphyria is a rare disease with a prevalence of about 1 in 10,000. Since patient A does not have a family history of the disease, his risk of porphyria is 0.01%. Equivalently, we can say that the prior probability (prior to knowing the test result) is 0.01%. After knowing the test result, what is the probability that patient A has the disease? To calculate this probability, we need to fill in the blanks (A through I) in Table 14.2 for a large population similar to patient A.

Follow these steps to fill in the table. (Don't skip these steps. I show similar tables many times in this book, and you need to understand them.)

Table 14.2. Results of a Lab Test: Definitions of A through I

	Disease Present	Disease Absent	Total
Test positive	A	D	G
Test negative	B	E	H
Total	C	F	I

1. Assume a population of 1,000,000 (arbitrary). Enter the number 1,000,000 in position I. All we care about is ratios of values, so the total population size is arbitrary. I = 1,000,000.
2. Since the prevalence of the disease is 1/10,000, the total number in the disease present column is 0.0001 × 1,000,000 or 100. C = 100.
3. Subtract the 100 diseased people from the total of 1,000,000, leaving 999,900 disease absent people. F = 999,900.
4. Next calculate the number of people with disease present who also test positive. This equals the total number of people with the disease times the sensitivity. Recall that sensitivity is defined as the fraction of those with the disease whose test result is positive. So the number of people with disease and a positive test is 0.82 * 100 = 82. A = 82.
5. Next calculate the number of people without disease who test negative. This equals the total number of people without the disease (999,900) times the specificity (.963). Recall that the specificity is the fraction of those without the disease whose test result is negative. E = 962,904.
6. Calculate the number of people with the disease who test negative by subtraction: B = 100 − 82 = 18. So B = 18.
7. Calculate the number of people without the disease who test positive by subtraction: D = F − E = 36,996.
8. Calculate the two row totals. G = A + D = 37078. H = B + E = 962,922.

Table 14.3 is the completed table.

If you screen 1 million people, you expect to find a test result less than 99 units in 37,078 people. Only 82 of these cases will have the disease. The predictive value of a positive test is only 82/37,078, which equals 0.22%. Only about 1 in about 500 of the positive tests indicate disease! The other 499 out of 500 positive tests are false positives. The test is not very helpful.

Patient B

This patient's brother has the disease. The disease is autosomal dominant, so there is a 50% chance that each sibling has the gene. The prior probability is 50%. Table 14.4 gives results you expect to see when you screen 1000 siblings of patients.

You expect to find 429 positive tests. Of these individuals, 410 will actually have the disease and 19 will be false positives. The predictive value of a positive test is 410/429, which is about 96%. Only about 5% of the positive test are false positives.

Table 14.3. Porphyria Example A: Screening Test

Patient A	Disease Present	Disease Absent	Total
<99 units	82	36,996	37,078
>99 units	18	962,904	962,922
Total	100	999,900	1,000,000

Table 14.4. Porphyria Example B: Siblings of People with the Disease

	Disease Present	Disease Absent	Total
<99 units	410	19	429
>99 units	90	481	571
Total	500	500	1000

Patient C

This patient does not have a family history of porphyria, but he has all the right symptoms. You suspect that he has it. What pretest probability should you use? Base it on (informed) clinical judgment. In this case, you feel 30% sure that the patient has the disease. So the prior probability is 30%. Filling in the table is easy, as illustrated in Table 14.5.

If the test is positive, you can conclude that there is about a 246/272 = 90% chance that the patient has porphyria. Unlike the other examples, the "prior probability" in this example is a fairly fuzzy number. In the other two examples, the prior probability came from epidemiological data (prevalence) or from genetic theory. Here it just comes from clinical judgment. Since the prior probability is not an exact probability, the answer (90%) is also not exact. It is an estimate, but you can be fairly certain that patient C has porphyria. If the prior probability was really 20% or 40% (instead of 30%), you can calculate that the predictive value would be 85% or 94%.

The three patients had exactly the same test result, but the interpretation varies widely. To interpret the test result, it is not enough to know the sensitivity and specificity. You also need to know the prior probability (prior to seeing the test result) that the patient had the disease. For patient A, you obtained the prior probability from population prevalence. In patient B, you obtained the prior probability from genetic theory. In patient C, you estimated the prior probability using clinical intuition.

The predictive value of a positive test depends on who you are testing. Your interpretation of a positive test should depend partly on the characteristics of the test (as measured by sensitivity and specificity) but also on the prevalence of the disease in the group being tested. A test that is very useful for testing patients strongly suspected of having the disease (high prevalence) may turn out to be useless as a screening test in a more general population (low prevalence). Screening tests for rare diseases in the general population are only useful when both sensitivity and specificity are extremely high.

Table 14.5. Porphyria Example C: Clinical Suspicion

Patient C	Disease Present	Disease Absent	Total
<99 units	246	26	272
>99 units	54	674	728
Total	300	700	1000

Table 14.6. Porphyria Example D

	Disease Present	Disease Absent	Total
<79 units	219	2	221
>79 units	281	498	779
Total	500	500	1000

Patient D

Like patient B, this patient is the brother of an affected patient. But patient D's enzyme level is lower, 79 units. Since a low level is abnormal, the lower activity is associated with lower sensitivity and higher specificity. Fewer patients and many fewer normals have such a low enzyme level. For this level, the sensitivity is 43.8% and the specificity is 99.5%. If you test 1000 siblings of porphyria patients, the results you expect to find are given in Table 14.6.

You'd only expect to find 221 people whose enzyme level is so low, and 219 of those have the disease. The predictive value of a positive test is 219/222 or 98.6%. As you'd expect, a lower (more abnormal) test result has a higher predictive value.

BAYES' THEOREM

In these examples, calculating the predictive values using the tables took several steps. These can be combined into a single equation named for Thomas Bayes, an English clergyman who worked out the mathematics of conditional probability in the late 1700s. The equation can be written in terms of either probabilities or odds, but the equation is much simpler when expressed in terms of odds. Therefore, you need to review the difference between probability and odds before reading about Bayes theorem.

A REVIEW OF PROBABILITY AND ODDS

Likelihood can be expressed either as a probability or as odds.

- The *probability* that an event will occur is the fraction of times you expect to see that event in many trials.
- The *odds* are defined as the probability that the event will occur divided by the probability that the event will not occur.

A probability is a fraction and always ranges from 0 to 1. Odds range from 0 to infinity. Any probability can be expressed as odds. Any odds can be expressed as a probability. Convert between odds and probability with Equations 14.3 and 14.4:

$$\text{Odds} = \frac{\text{probability}}{1 - \text{probability}}. \tag{14.3}$$

$$\text{Probability} = \frac{\text{odds}}{1 + \text{odds}}. \tag{14.4}$$

If the probability is 0.50 or 50%, then the odds are 50:50 or 1. If you repeat the experiment often, you expect to observe the event (on average) in one out of two trials (probability = 1/2). That means you'll observe the event once for every time it fails to happen (odds = 1:1).

If the probability is 1/3, the odds equal 1/3/(1 − 1/3) = 1:2 = 0.5. On average, you'll observe the event once in every three trials (probability = 1/3). That means you'll observe the event once for every two times it fails to happen (odds = 1:2).

BAYES' EQUATION

The Bayes' equation for clinical diagnosis is Equation 14.5:

$$\text{Post-test odds} = \text{pretest odds} \cdot \frac{\text{sensitivity}}{1 - \text{specificity}}. \tag{14.5}$$

The post-test odds are the odds that a patient has the disease, taking into account both the test results and your prior knowledge about the patient. The pretest odds are the odds that the patient has the disease determined from information you know before running the test. The ratio sensitivity/(1 − specifity) is called the *likelihood ratio.* It is the probability of obtaining the positive test result in a patient with the disease (sensitivity) divided by the probability of obtaining a positive test result in a patient without the disease (1 − specificity). So Bayes' equation can be written in a simpler (and more general) form:

$$\text{Post-test odds} = \text{pretest odds} \cdot \text{likelihood ratio}. \tag{14.6}$$

Using this equation we can rework the examples with intermittent porphyria. The test used in the example has a sensitivity of 82% and a specificity of 96.3%. Thus the likelihood ratio (sensitivity/1 − specificity) is .82/(1.0 − .963) = 22.2.* Analysis of patients A through C is shown in Table 14.7.

The first column shows the pretest probability, which came from epidemiological data (A), genetic theory (B), or clinical experience (C). The second column expresses the pretest probability as odds, calculated from the pretest probability using Equation 14.3. The third column was calculated from Equation 14.6. This is the odds that the patient has the disease, considering both the results of the test and the pretest odds. The last column converts this result back to a probability using Equation 14.4. The

Table 14.7. Intermittent Porphyria Examples Using Bayes' Equation

Patient	Pretest Probability	Pretest Odds	Post-Test Odds	Post-Test Probability
A	0.0001	0.0001	0.0022	0.0022
B	0.50	1.0000	22.2	0.957
C	0.30	0.4286	9.514	0.905

*If you express sensitivity and specificity as percents, rather than fractions, the likelihood ratio is defined as (sensitivity/100 − specificity).

results, of course, match those calculated earlier from the individual tables. Using Bayes' equation is more efficient.

SOME ADDITIONAL COMPLEXITIES

- Bayesian logic integrates the result of one lab test into the entire clinical picture. Many clinicians do a pretty good job of combining probabilities intuitively without performing any calculations and without knowing anything about Bayesian thinking. The formal Bayesian approach is more explicit and exact. Moreover, it clearly shows the necessity of knowing the prevalence of diseases and the sensitivity and specificity of tests.
- When thinking about the predictive value of tests, distinguish screening tests from confirmatory tests. It is OK if the quick and cheap screening tests turns up a lot of false positives, as long as the confirmatory test (often slower and more expensive) gives the correct result. When a positive screening test is followed by a confirmatory test, you really only care about the predictive value of the pair of tests.
- For some genetic diseases you need to distinguish the sensitivity to detect the genetic defect from the sensitivity to detect clinical disease. Some genetic diseases have poor penetrance, meaning that some people with the abnormal gene do not get the disease. A test that detects the gene with few false positives would produce a lot of false positives when assessed for its ability to detect the clinical disease.
- For many tests, sensitivity and specificity are tabulated for various values of the test. It is not necessary to pick a single threshold between positive and negative. Patient D described earlier demonstrates this point.
- The closer you can test for the real cause of the disease, the higher the sensitivity and specificity. If you tested for the abnormal gene sequence (rather than enzyme activity), the sensitivity and specificity would be extremely high (unless the penetrance is low). In this case, the test would probably be definitive, and there would be little need for Bayesian analyses.

SUMMARY

There are many ways of summarizing the accuracy of a diagnostic test. *Sensitivity* quantifies how well the test correctly detects the presence of the condition; *specificity* quantifies how well it correctly detects the absence of the condition.

The rates of false-negative and false-positive test results depend not only on the sensitivity and specificity of the test, but also on the prior probability that the subject has the disease. In some situations, you know the prior probability from population epidemiology. The prior probability is the prevalence of the disease. In other cases, you know the prior probability from genetic theory. In still other situations, you can estimate the prior probability from clinical experience.

Bayesian logic can be used to combine the result of the test with the prior probability to determine the probability that the patient has the condition. The Bayesian approach lets you combine the objective results of a test with your prior clinical suspicion to calculate the probability that a patient has a disease. Although formal

Bayesian analysis is seldom used in clinical settings, good clinicians intuitively combine these probabilities routinely.

OBJECTIVES

1. You must be familiar with the following terms:
   - Sensitivity
   - Specificity
   - False positives
   - False negatives
   - Predictive value
   - Bayes' equation
   - Bayesian logic
   - Likelihood ratio
2. You must understand why the rates of false positives and negatives depends on the prevalence of the condition being tested for.
3. Given the specificity, sensitivity, and prevalence, you should be able to calculate the rate of false positives and false negatives.
4. Using a book for reference, you should be able to calculate the probability that a patient has a disease if you are given the specificity and sensitivity of a test and the prior odds (or prior probability).

PROBLEMS

1. A test has a specificity of 92% and a sensitivity of 90%. Calculate the predictive values of positive and negative tests in a population in which 5% of the individuals have the disease.
2. A test has a specificity of 92% and a sensitivity of 99%. Calculate the predictive values of positive and negative tests in a population in which 0.1% of the individuals have the disease.
3. A woman wants to know if her only son is color blind. Her father is color blind, so she must be a carrier (because color blindness is a sex-linked trait). This means that, on average, half her sons will be color blind (she has no other sons). Her son is a smart toddler. But if you ask him the color of an object, his response seems random. He simply does not grasp the concept of color. Is he color blind? Or has he not yet figured out what people mean when they ask him about color? From your experience with other kids that age, you estimate that 75% of kids that age can answer correctly when asked about colors. Combine the genetic history and your estimate to determine the chance that this kid is color blind.
4. For patient C in the porphyria example, calculate the predictive value of the test if your clinical intuition told you that the prior probability was 75%.

15

Bayes and Statistical Significance

Setting the value of α, the threshold P value for determining significance, is similar to selecting the threshold value for a lab test that distinguishes "normal" from "abnormal." Recall from the previous chapter that selecting a threshold value for a lab test involves a trade-off between false positives and false negatives. Similarly, you learned in the previous chapter that selecting a value for α involves a tradeoff between Type I errors and Type II errors. The analogy is shown in Tables 15.1 and 15.2. You should memorize the differences between Type I and Type II errors. Those terms are sometimes mentioned in papers without being defined, and the distinction is not always clear from the context.

If a lab test measures the concentration of a chemical in the blood, you don't have to worry about false-positive and false-negative lab tests if you think about the actual concentration, rather than the conclusion positive or negative. Similarly, you don't have to worry about Type I and II errors if you think about the actual value of P (as a way to describe or summarize data) rather than the conclusion significant or not significant.

TYPE I ERRORS AND FALSE POSITIVES

You have made a Type I error when you reject the null hypothesis ($P < \alpha$) when the null hypothesis is really true. Note the subtle distinction between P values and α. Before collecting data, you choose the value of α, the threshold value below which you will deem a P value significant. Then you calculate the P value from your data.

The following statements summarize the analogy between specificity and P values:

Lab. If the patient really does not have the disease, what is the chance that the test will yield a positive result? The answer is 1 minus the specificity.

Statistics. If we assume that the two populations have identical means (or proportions), what is the chance that your study will find a statistically significant difference? The answer is α.

TYPE II ERRORS AND FALSE NEGATIVES

To define a Type II error, you must define your experimental hypothesis. To distinguish it from the null hypothesis, the experimental hypothesis is sometimes called the

Table 15.1. False Negatives and Positives in Diagnostic Tests

Diagnostic Test	Disease Is Really Present	Disease Is Really Absent
Test positive	No error (true positive)	False positive
Test negative	False negative	No error (true negative)

alternative hypothesis. It is not enough to say that the experimental hypothesis is that you expect to find a difference; you must define how large you expect the difference to be.

You have made a Type II error when you conclude that there is no significant difference between two means, when in fact the alternative hypothesis is true. The probability of making a Type II error is denoted by β and is sometimes called a *beta error.* The value of β depends on how large a difference you specify in the alternative hypothesis. If you are looking for a huge difference, the probability of making a Type II error is low. If you are looking for a tiny difference, then the probability of making a Type II error is high. Thus, one cannot think about β without defining the alternative hypothesis. This is done by deciding on a value for the minimum difference (or relative risk) that you think is clinically or scientifically important and worth detecting. This minimum difference is termed Δ (delta). Your choice of Δ depends on the scientific or clinical context. Statisticians or mathematicians can't help, the alternative hypothesis must be based on your scientific or clinical understanding. β is the probability of randomly selecting samples that result in a nonsignificant P value when the difference between population means equals Δ.

The *power* of a test is defined as $1 - \beta$. The power is the probability of obtaining a significant difference when the difference between population means equals Δ. Like β, the power can only be defined once you have chosen a value for Δ. The larger the sample size, the greater the power. The lower you set α, the lower the power. Increasing Δ will increase the power, as it is easier to find a big difference than a small difference.

The following statements summarize the analogy between sensitivity and power:

Lab. If the patient really has a certain disease, what is the chance that the test will correctly give a positive result? The answer is the sensitivity. If the test can detect several diseases, the sensitivity of the test depends on which disease you are looking for.

Statistics. If there really is a difference (Δ) between population means (or proportions), what is the chance that analysis of randomly select subjects will result in a significant difference? The answer is the power, equal to one minus β. The answer depends on the size of the hypothesized difference, Δ.

Table 15.2. Type I and Type II Errors in Statistical Tests

Statistical Test	Populations Have Different Means (or Proportions)	Populations Have Identical Means (or Proportions)
Significant difference	No error	Type I error
No significant difference	Type II error	No error

PROBABILITY OF OBTAINING A FALSE-POSITIVE LAB RESULT: PROBABILITY THAT A SIGNIFICANT RESULT WILL OCCUR BY CHANCE

What is the probability of obtaining a false-positive lab result? This question is a bit ambiguous. It can be interpreted as two different questions:

- What fraction of all disease free individuals will have a positive test? This answer equals FP/(FP + TN), which is the same as one minus the specificity.
- What fraction of all positive test results are false positives? The answer is FP/(FP + TP). This is the conventional definition of the rate of false positives. As you learned in the previous chapter, this question can be answered only if you know the prevalence of the disease (or prior probability) in the population you are studying.

What is the probability of obtaining a statistically significant P value by chance? Again, this question is ambiguous. It can be interpreted as two different questions:

- If the null hypothesis is true, what fraction of experiments will yield a significant P value? Equivalently, if the null hypothesis is true, what is the probability of obtaining a statistically significant result ($P < \alpha$)? The answer is α, conventionally set to 5%.
- In what fraction of all experiments that result in significant P values is the null hypothesis true? Equivalently, if a result is statistically significant, what is the probability that the null hypothesis is true? The answer is not necessarily 5%. Conventional statistics cannot answer this question at all. Bayesian logic can answer the question, but only if you can define the prior probability that the null hypothesis is true. The next section discusses how to apply Bayesian logic to P values.

In each case (lab tests and statistical tests) the logic of the first question goes from population to sample, and the logic of the second goes from sample to population. When analyzing data, we are more interested in the second question.

THE PREDICTIVE VALUE OF SIGNIFICANT RESULTS: BAYES AND P VALUES

You perform a statistical test and obtain a significant result. Repeated from the last section, here is the question you wish to answer:

> In what fraction of all experiments that result in significant P values is the null hypothesis true? Equivalently, if the result is statistically significant, what is the probability that the null hypothesis is really true?

Here is an imaginary example. You are working at a drug company and are screening drugs as possible treatments for hypertension. You test the drugs in a group of animals. You have decided that you are interested in a mean decrease of blood pressure of 10 mmHg and are using large enough samples so that you have 80% power to find a significant difference ($\alpha = 0.05$) if the true difference between population means is 10 mmHg. (You will learn how to calculate the sample size in Chapter 22.)

You test a new drug and find a significant drop in mean blood pressure. You know that there are two possibilities. Either the drug really works to lower blood pressure, or the drug doesn't alter blood pressure at all and you just happened to get lower pressure readings on the treated animals. How likely are the two possibilities?

Since you set α to 0.05, you know that 5% of studies done with inactive drugs will demonstrate a significant drop in blood pressure. But that isn't the question you are asking. You want to know the answer to a different question: In what fraction of experiments in which you observe a significant drop in pressure is the drug really effective? The answer is not necessarily 5%. To calculate the answer you need to use Bayesian logic and need to think about the prior probability. The answer depends on what you knew about the drug before you started the experiment, expressed as the prior probability that the drug works. This point is illustrated in the following three examples.

Drug A

This drug is known to weakly block angiotensin receptors, but the affinity is low and the drug is unstable. From your experience with such drugs, you estimate that there is about a 10% chance that it will depress blood pressure. In other words, the prior probability that the drug works is 10%. What will happen if you test 1000 such drugs? The answer is shown in Table 15.3.

These are the steps you need to follow to create the table:

1. We are predicting the results of 1000 experiments with 1000 different drugs, so the grand total is 1000. This number is arbitrary, since all we care about are ratios.
2. Of those 1000 drugs we screen, we expect that 10% will really work. In other words, the prior probability equals 10%. So we place 10% of 1000 or 100 as the total of the first column, leaving 900 for the sum of the second column.
3. Of the 100 drugs that really work, we will obtain a significant result in 80% (because our experimental design has 80% power). So we place 80% of 100, or 80, into the top left cell of the table. This leaves 20 experiments with a drug that really works, but $P > 0.05$ so we conclude that the drug is not effective.
4. Of the 900 drugs that are really ineffective, we will by chance obtain a significant reduction in blood pressure in 5% (because we set α equal to 0.05). Thus the top cell in the second column is 5% $\times$ 900 or 45. That leaves 855 experiments in which the drug is ineffective, and we correctly observe no significant difference.
5. Determine the row totals by addition.

Out of 1000 tests of different drugs, we expect to obtain a significant difference ($P < 0.05$) in 125 of them. Of those, 80 drugs are really effective and 45 are not. When

Table 15.3. Statistical Significance When Testing Drug A

Drug A Prior Probability = 10%	Drug Really Works	Drug Is Really Ineffective	Total
Significant difference	80	45	125
No significant difference	20	855	875
Total	100	900	1000

you see a significant result for any particular drug, you can conclude that there is a 64% chance (80/125) that the drug is really effective and a 36% chance (45/125) that it is really ineffective.

Drug B

Here the pharmacology is much better characterized. Drug B blocks the right kinds of receptors with reasonable affinity and the drug is chemically stable. From your experience with such drugs, you estimate that the prior probability that the drug is effective equals 80%. What would happen if you tested 1000 such drugs? The answer is shown in Table 15.4.

If you test 1000 drugs like this one, you expect to see 650 significant results. Of those, 98.5% (640/650) will be truly effective. When you see a significant result for any particular drug, you can conclude that there is a 98.5% chance that it will really lower blood pressure and a 1.5% chance that it is really ineffective.

Drug C

This drug was randomly selected from the drug company's inventory of compounds. Nothing you know about this drug suggests that it affects blood pressure. Your best guess is that about 1% of such drugs will lower blood pressure. What would happen if you screen 1000 such drugs? The answer is shown in Table 15.5.

If you test 1000 drugs like this one, you expect to see 58 significant results. Of those, you expect that 14% (8/58) will be truly effective and that 86% (50/58) will be ineffective. When you see a significant result for any particular drug, you can conclude that there is a 14% chance that it will really lower blood pressure and an 85% chance that it is really ineffective.

These examples demonstrate that your interpretation of a significant result appropriately depends on what you knew about the drug before you started. You need to integrate the P value obtained from the experiment with the prior probability.

When you try to do the calculations with real data, you immediately encounter two problems:

- You don't know the prior probability. The best you can do is convert a subjective feeling of certainty into a "probability." If you are quite certain the experimental hypothesis is true, you might say that the prior probability is 0.99. If you are quite certain the experimental hypothesis is false, you might say that the prior probability is 0.01. If you think it could go either way, you can set the prior probability to 0.5.

Table 15.4. Statistical Significance When Testing Drug B

Drug B Prior Probability = 80%	Drug Really Works	Drug Is Really Ineffective	Total
Significant difference	640	10	650
No significant difference	160	190	350
Total	800	200	1000

Table 15.5. Statistical Significance When Testing Drug C

Drug C Prior Probability = 1%	Drug Really Works	Drug Is Really Ineffective	Total
Significant difference	8	50	58
No significant difference	2	940	942
Total	10	990	1000

- You don't know what value to give Δ, the smallest difference that you think is scientifically or clinically worth detecting. While it is usually difficult to choose an exact value, it is usually not too hard to estimate the value.

Despite these problems, it is often possible to make reasonable estimates for both the prior probability and Δ. It's OK that these values are estimated, so long as you treat the calculated probability as an estimate as well.

THE CONTROVERSY REGARDING BAYESIAN STATISTICS

It is possible to combine all the steps we took to create the tables into one simple equation called the *Bayes' equation,* as you saw in the last chapter. The entire approach discussed in the previous section is called *Bayesian thinking.* The Bayesian approach to interpreting P values is rarely used. If you knew the prior probability, applying Bayesian logic would be straightforward and not controversial. However, usually the prior probability is not a real probability but is rather just a subjective feeling. Some statisticians (Bayesians) think it is OK to convert these feelings to numbers ("99% sure" or "70% sure"), which they define as the prior probability. Other statisticians (frequentists) think that you should never equate subjective feelings with probabilities.

There are some situations where the prior probabilities are well defined. For example, see the discussion of genetic linkage in the next chapter. The prior probability that two genetic loci are linked is known, so Bayesian statistics are routinely used in analysis of genetic linkage. There is nothing controversial about using Bayesian logic when the prior probabilities are known precisely.

The Bayesian approach explains why you must interpret P values in the context of what you already know or believe, why you must think about biological plausibility when interpreting data. When theory changes, it is appropriate to change your perception of the prior probability and to change your interpretation of data. Accordingly, different people can appropriately and honestly reach different conclusions from the same data. All significant P values are not created equal.

APPLYING BAYESIAN THINKING INFORMALLY

When reading biomedical research, you'll rarely (if ever) see Bayesian calculations used to interpret P values. And few scientists use Bayesian calculations to help interpret P values. However, many scientists use Bayesian thinking in a more informal way without stating the prior probability explicitly and without performing any additional calculations. When reviewing three different studies, the thinking might go like this:

This study tested a hypothesis that is biologically sound and that is supported by previous data. The P value is 0.04, which is marginal. I have a choice of believing that the results are due to a coincidence that will happen 1 time in 25 under the null hypothesis, or of believing that the experimental hypothesis is true. Since the hypothesis makes so much sense, I'll believe it. The null hypothesis is probably false.

This study tested a hypothesis that makes no biological sense and has not been supported by any previous data. The P value is 0.04, which is lower than the usual threshold of 0.05, but not by very much. I have a choice of believing that the results are due to a coincidence that will happen 1 time in 25 under the null hypothesis, or of believing that the experimental hypothesis is true. Since the experimental hypothesis is so crazy, I find it easier to believe that the results are due to coincidence. The null hypothesis is probably true.

This study tested a hypothesis that makes no biological sense and has not been supported by any previous data. I'd be amazed if it turned out to be true. The P value is incredibly low (0.000001). I've looked through the details of the study and cannot identify any biases or flaws. These are reputable scientists, and I believe that they've reported their data honestly. I have a choice of believing that the results are due to a coincidence that will happen one time in a million under the null hypothesis or of believing that the experimental hypothesis is true. Even though the hypothesis seems crazy to me, the data force me to believe it. The null hypothesis is probably false.

You should interpret experimental data in the context of theory and previous data. That's why different people can legitimately reach different conclusions from the same data.

MULTIPLE COMPARISONS

Experienced clinicians do not get excited by occasional lab values that are marginally abnormal. If you perform many tests on a patient, it is not surprising that some are labeled "abnormal," and these may tell you little about the health of the patient. You need to consider the pattern of all the tests and not focus too much on any one particular test. If the test is quantitative, you also need to consider whether the test is just barely over the arbitrary line that divides normal from abnormal, or whether the result is really abnormal and far from the dividing line.

Similarly, experienced scientists do not get excited by occasional "significant" P values. If you calculate many P values, you expect some to be small and significant just by chance. When you interpret significant P values, you must take into account the total number of P values that were calculated. If you make multiple comparisons and calculate many P values, you expect to encounter some small P values just by chance. Chapter 13 discussed this problem in great detail.

SUMMARY

The analogy between diagnostic tests and statistical hypothesis tests is summarized in Table 15.6.

Table 15.6. Comparison Between Diagnostic Tests and Statistical Hypothesis Tests

	Lab Test	Statistical Hypothesis Test
Result	The result is a measurement, but it can be compared to a threshold and reported as "normal" or "abnormal."	The result is a P value, but it can be compared to a threshold and reported as "statistically significant" or "not statistically significant."
Scope	A lab test is performed for one individual and yields the diagnosis of positive or negative.	A P value is calculated from one experiment and yields the conclusion of significant or not significant.
Errors	A lab test can result in two kinds of errors: false positives and false negatives.	A statistical hypothesis test can result in two kinds of errors: Type I and Type II.
Threshold	You should choose the threshold between "normal" and "abnormal" based on the relative consequences of false-positive and false-negative diagnoses.	You should choose a value for α (the threshold between "not significant" and "significant" P values) based on the relative consequences of making a Type I or Type II error.
Accuracy	The accuracy of the lab test is expressed as two numbers: sensitivity and specificity.	The accuracy of the statistical test is expressed as two numbers: α and β (or power).
Interpretation	When interpreting the result of a lab test for a particular patient, you must integrate what is known about the accuracy of the laboratory test (sensitivity and specificity) with what is known about the patient (prevalence, or prior probability that the patient has disease). Bayesian logic combines these values precisely.	When interpreting the result of a statistical test of a particular hypothesis, you must integrate what is known about the accuracy of the statistical test (α and β) with what is known about the hypothesis (prior probability that the hypothesis is true). Bayesian logic combines these values precisely.
Multiple comparisons	If you perform many tests on one patient, you shouldn't be surprised to see occasional "abnormal" results. If you perform many tests, you need to look at overall patterns and not just individual results.	If you perform many statistical tests, you shouldn't be surprised to see occasional "significant" results. If you perform many tests, you need to look at overall patterns and not just at individual P values.

OBJECTIVES

1. You must be familiar with the following terms:
   - Type I error
   - Type II error
   - α error
   - β error
   - Power

2. You should be able to explain the analogy between false-positive and false-negative lab tests and Type II and Type I statistical errors.
3. You should understand why it is hard to answer this question: In what fraction of all experiments that result in a significant P value is the null hypothesis really true?
4. You should be able to explain why the answer to that question depends on the nature of the hypothesis being tested.
5. Given a prior probability and power, you should be able to calculate the predictive value of a statistically significant P value.

PROBLEMS

1. A student wants to determine whether treatment of cells with a particular hormone increases the number of a particular kind of receptors. She and her advisor agree that an increase of less than 100 receptors per cell is too small to care about. Based on the standard deviation of results you have observed in similar studies, she calculates the necessary sample size to have 90% power to detect an increase of 100 receptors per cell. She performs the experiment that number of times, pools the data, and obtains a P value of 0.04.

   The student thinks that the experiment makes a lot of sense and thought that the prior probability that her hypothesis was true was 60%. Her advisor is more skeptical and thought that the prior probability was only 5%.

   A. Combining the prior probability and the P value, what is the chance that these results are due to chance? Answer from both the student's perspective and that of the advisor.
   B. Explain why two people can interpret the same data differently.
   C. How would the advisor's perspective be different if the P value were 0.001 (and the power were still 90%)?

2. You go to Las Vegas on your 25th birthday, so bet on the number 25 in roulette. You win. You bet a second time, again on 25, and win again! A roulette wheel has 38 slots (1 to 36, 0, and 00), so there is a 1 in 38 chance that a particular spin will land on 25.

   A. Assuming that the roulette wheel is not biased, what is that chance that two consecutive spins will land on 25?
   B. If you were to spend a great deal of time watching roulette wheels, you would note two consecutive spins landing on 25 many times. What fraction of those times would be caused by chance? What fraction would be caused by an unfair roulette wheel?

16

Bayes' Theorem in Genetics

BAYES' THEOREM IN GENETIC COUNSELING

In genetic counseling you want to determine the probability that someone has a particular genetic trait.

Example 16.1

A woman wants to know her chances of being a carrier for Duchenne's muscular dystrophy, an X-linked recessive trait. Since her brother and maternal uncle both have the disease, it is clear that the gene runs in her family and is not a new mutation. From her family history, her mother must be a carrier and the woman had a 50% chance of inheriting the gene at birth.

Knowing that the woman has two sons without the disease decreases the chance that the woman is a carrier. Bayesian logic allows you to combine this evidence (two healthy sons) with the family history (50% chance of being a carrier). We'll first perform the calculations step by step with a table and then use Bayes' equation. Table 16.1 shows what you would expect to see if you were to examine many women with the same family history and two sons. The calculations are explained later.

To generate the table, follow these steps:

1. Set the grand total to 1000. This is arbitrary as we only care about ratios.
2. We know that half the women are carriers, so place $1/2 \times 1000$ or 500 into each column total.
3. If a woman is a carrier, there is a 1/4 chance ($1/2 \times 1/2$) that both her sons would not have the disease. So place $1/4 \times 500 = 125$ in box A. That leaves 375 cases in box C.
4. If a woman is not a carrier, then none of her sons will have this disease (barring new mutations, which are very rare). So $D = 0$ and $B = 500$.
5. Compute the row totals.

Of the 1000 hypothetical women with two sons and this family history, 375 would have at least one son with the disease. We know that the woman in our example is not in this category. She is in the group of 625 women who have two sons without the disease. Of these 125 are carriers. So $125/625 = 20\%$ of the women with two healthy sons are carriers. Thus we can say that the woman in our example has a 20% chance, or 1 in 5, of being a carrier.

Table 16.1. Calculations of Chance of Being a Carrier of Duchenne's Muscular Dystrophy in the Example

	Woman Is a Carrier	Woman Is Not a Carrier	Total
Both sons without disease	A = 125	B = 500	625
At least one son has the disease	C = 375	D = 0	375
Total	500	500	1000

From the laws of Mendelian genetics, we knew that her risk of being a carrier at birth was 1/2. Taking into account her two unaffected sons, using Bayesian logic lowers the risk to 1/5. Now let's use the Bayes' equation to streamline the calculations. Bayes' equation is as follows:

$$\text{Post-test odds} = \text{pretest odds} \cdot \text{likelihood ratio.} \qquad (16.1)$$

The likelihood ratio is the probability a carrier will have two unaffected sons divided by the probability that a noncarrier will have two unaffected sons. The probability that a carrier will have an unaffected son is 1/2. Therefore, the probability that both sons will be unaffected is $1/2 \times 1/2 = 1/4$ or 25%. The probability that a noncarrier will have two sons without this disease is 100% (barring new mutations, which are extremely rare). So the likelihood ratio is 25%/100% or 0.25.

From her family history, we know that this woman had a 50% chance of being a carrier at birth. This is the pretest probability. Therefore the pretest odds are 50:50 or 1.0. Multiply the pretest odds by the likelihood ratio to calculate the post-test odds, which equal 0.25 or 1:4. If you saw many people with the same family history as this woman, you'd see one carrier for every four noncarriers. Converting from odds to probability, the post-test probability is 20%. She has a 20% chance of being a carrier.

BAYES AND GENETIC LINKAGE

When two loci (genes or DNA sequences) are located near each other on the same chromosome, they are said to be linked. If the two loci are very close, crossing over or recombination between the two loci occurs rarely. Thus, alleles of linked loci tend to be inherited together. If the loci are further apart, recombination (a normal process) occurs more frequently. If the loci are very far apart, the two loci segregate independently just as if they were on different chromosomes.

Linkage is useful in genetic diagnosis and mapping. Since it is not possible to detect all abnormal genes directly, geneticists try to identify a marker gene (such as those for variable antigens or isozymes) or a variable DNA sequence that is linked to the disease gene. Once you know that the disease gene is linked to a marker, the presence of the marker (which you can identify) can then be used to predict the presence of the disease gene (which you cannot identify directly). This allows detection of genetic diseases prenatally or before they cause clinical problems. It also allows diagnosis of unaffected heterozygotes (carriers) who can pass the abnormal gene on

to their children. This method works best for diseases caused by an abnormality of a single gene.

Before linkage can be useful in diagnosis, you need to identify a marker linked to the gene. This is usually done by screening lots of potential markers. How can you tell if a marker is linked to a disease gene? Geneticists study large families and observe how often the disease and marker are inherited together and how often there is recombination. If there are few recombination events between the marker and the disease, there are two possible explanations. One possibility is that the two are linked. The other possibility is that the two are not linked, but—just by coincidence—there were few recombination events.

Bayesian logic combines the experimental data with the prior probability of linkage to determine the probability that the gene is truly linked to the disease. To calculate Bayes' equation, we need to define the likelihood ratio in the context of linkage. When calculating the predictive values of lab tests in Chapter 14, we defined the likelihood ratio as sensitivity divided by one minus specificity—the probability that someone with the disease will have an abnormal test result divided by the probability that someone without the disease will have an abnormal test result. For studies of linkage, therefore, the likelihood ratio is the probability of obtaining the data if the genes really are linked* divided by the probability of observing those data if the genes are really not linked. The details of the calculations are beyond the scope of this book. When you read papers with linkage studies, you'll rarely see reference to the likelihood ratio. Instead you'll see the *lod score* (*l*og of *od*ds), which is simply the logarithm (base 10) of the likelihood ratio.

The higher the lod score, the stronger the evidence for linkage. A lod score of 3 means that the likelihood ratio equals 1000 (antilog of 3). This means that the data are 1000 times more likely to be observed if the marker is linked to the disease than if the marker is not linked.

To calculate the probability that the marker is linked to the gene requires accounting for the prior probability of linkage using Bayesian logic. Bayes' equation for linkage can be written as follows:

$$\begin{aligned}\text{Post-test odds of linkage} &= \text{pretest odds of linkage} \cdot \text{likelihood ratio.}\\ \text{Post-test odds of linkage} &= \text{pretest odds of linkage} \cdot 10^{\text{lod}}.\end{aligned} \tag{16.2}$$

To calculate Bayes' equation, you must know the prior (or pretest) odds of linkage. Since there are 23 pairs of chromosomes, the chance that any particular randomly selected marker will be located on the same chromosome as the disease gene is 1/23 or 4.3%. But it is not enough to be on the same chromosome. To be linked to a disease, the marker must be close to the disease gene. So the prior probability that a random marker is linked to a particular gene must be less than 4.3%. In fact, genetic data tell us that 2% of randomly selected markers are linked to any particular disease gene.† Converting to odds, the pretest odds of linkage are about 0.02. The values presented here assume that the marker was randomly selected, as is often the case. If you pick a marker known to be on the same chromosome as the disease, then the pretest odds of linkage are higher.

*This can only be calculated once you specify a hypothetical genetic distance θ.

†For these calculations, we define linkage to mean that the probability of recombination is 30% or less.

Let's assume that a lod score equals 3. What is the probability that the marker and disease are truly linked? The post-test odds equal the pretest odds (0.02) times the likelihood ratio ($10^3 = 1000$), which is 20. Converting to a probability, the post-test probability equals 20/21 (Equation 14.4), which is about 95%. If you observe a lod score of 3.0, you will conclude that the marker and gene are linked. When you make that conclusion, there is a 95% chance that you will be correct, leaving a 5% chance that you will be wrong.

If a lod score equals or exceeds -2, geneticists usually conclude that the marker and disease are linked. If a lod score is less than or equal to -2, geneticists conclude that the marker and disease are not linked. See Problem 2 to calculate the probability that this conclusion is wrong. If the lod score is between -2 and 3, geneticists conclude that the evidence is not conclusive.

PROBLEMS

1. In Example 16.1, assume that the woman had three unaffected sons. What is the probability that she is a carrier?
2. If the lod score is -3, what is the probability that the marker is linked to the disease?
3. It would be possible to calculate a P value from linkage data. Explain in plain language what it would mean.
4. You perform a t test and obtain a P value of 0.032. You used enough subjects to ensure that the experiment had a 80% power to detect a specified difference between population means with $P < 0.05$. Does it make sense to calculate a likelihood ratio? If so, calculate the ratio and explain what it means.