CHAPTER 8

Hypothesis Testing

Often in life just making a prediction is not good enough. People who really make a difference make decisions. People who make a positive difference either make their decisions based on sound data or are very lucky. Since you can't count on luck you'd better count on making decisions based on sound information. But, even sound information can't guarantee anything if the way you process that data is without good sense. This chapter presents you with the modern tools of the scientific method. Handle with care.

8.1 Intuition and Decision Making

There can be no more important thought processes in all of statistics than intuition and common sense. So often the variety of formulas can so intimidate the student that all normal thought processes are turned off. The student can begin to feel like their calculator. As they approach the classroom they "go into stat mode." While in this mode they are unfeeling machines pointing mechanically toward the goal of calculating some arcane formulas, often unaware of the real meaning of what they are trying to do. As we continue our quest for truth and an understanding of reality we must get out of this false "stat mode." The true stat mode requires all our humanity — our intuition, our judgment, and our reason. Machines really have no concept of truth or reality, no feeling of consciousness. They do only what they are told. Some students think that to be a good student is also to do only what you are told, but such is not the case. Good thinking that leads to the wrong answer because of some small mistake is much to be preferred to bad thinking that accidentally gives the right answer.

Warning! You must read the examples of this section to be prepared to do the exercises.

As the title indicates, this section will seek to awaken your intuition, your humanity — that which separates you from the blindingly fast but incredibly stupid machinery that serves you. You will be asked to respond to situations in this section based entirely on your common sense. The vehicles for this process will be examples and exercises. This is one section where skipping the reading of the section will definitely put you in real trouble when trying to complete the exercises.

Example 8.1: A software company claims that its customers will wait on average no longer than 5 minutes for technical support. A disgruntled customer does not believe their claim. To test the claim he makes 25 phone calls at random times, with the following results:

$$\bar{X} = 5.8 \text{ minutes and } s = 1.2 \text{ minutes.}$$

Note: The sample size of 25 is borderline for the Central Limit Theorem. Since it is a little bit below 30 the probabilities calculated in the example should be viewed as approximations unless the population of waiting times is known to be normal.

Do you think that the customer has sufficient cause to accuse the company of false claims? (You will learn most from this section if you play along and try to answer all questions in your mind as you read. There are no wrong answers. It is your intuition that must be used.) On the surface there is at least some cause to believe that the consumer has raised a legitimate issue. The sample mean of 5.8 *is* larger than the claimed average of 5. But is this *sufficient* evidence? After all, anytime one looks at a sample the results may not accurately reflect the nature of the true population.

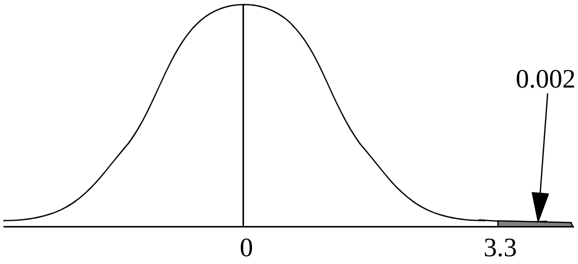
Note: The *t*-score is calculated just the same way as a *z*-score, but is used when s is used in place of σ.

One way to gauge whether a sample is extreme is to calculate a *t*-score.

$$t = \frac{\bar{X} - \mu}{s_{\bar{X}}} = \frac{5.8 - 5}{\frac{1.2}{\sqrt{25}}} = \frac{5.8 - 5}{0.24} = 3.3$$

This t (with $n - 1 = 24$ df as in chapter seven) is found in the long t-table. The following picture summarizes the findings:

Figure 8.1
Schematic For the *t*-score for Example 8.1



What does the tail area of 0.002 tell you? It tells you that if the population mean is really 5, then there is a very slim chance of getting a sample mean at least as far above 5 as the 5.8 the disgruntled customer obtained. How do you feel now? Do you think that the customer has a right to accuse the software company? With such a slim chance of obtaining a mean so far above 5 if 5 is really the mean, you should now be siding with the customer. It *does* appear as if he has a legitimate gripe. However, the practical importance of waiting 5.8 minutes instead of 5 minutes may be questioned. As one final note, notice that since the probability of 0.002 is so small, the fact we have applied the Central Limit Theorem to a sample of only 25 should not cause you to worry about your conclusion.

Example 8.2: Suppose that an airline has always gone on the assumption that the average weight of a piece of baggage at San Francisco International Airport is 27 lb. Baggage clerks have recently been complaining that they believe they are having to hoist more weight than that. The airline has decided to look into the situation. Fifty-two baggage items are randomly selected as they are off-loaded from incoming flights. Each item is then weighed. The results are:

$$\bar{X} = 30.2 \text{ lb.}, s = 11.1 \text{ lb.}$$

Once again the question is simple. Do these sample results disagree strongly enough with the assumption that the average baggage weight is 27 pounds to prove “beyond reasonable doubt” that the average weight is really more than 30 pounds? Once again, the best place to start is with a z or t-score. We will use a z-score in this example because the sample size is larger than our *t* table can handle.We know that in the case of very large samples the *t* and *z* are essentially identical anyway.

$$z = \frac{30.2 - 27}{\frac{11.1}{\sqrt{52}}} = \frac{30.2 - 27}{1.539293} = 2.08$$

Once again we are faced with a decision as to whether or not this *z*-score is extreme enough to force us to discard the old belief that baggage averages 27 pounds in weight. Let's look at the picture:

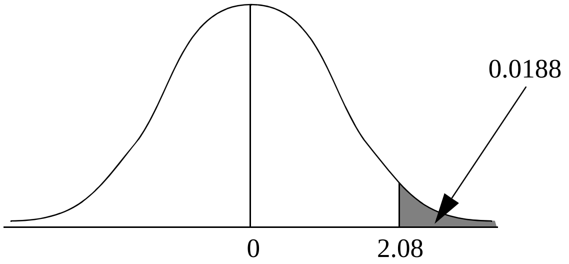


Figure 8.2
Schematic For the *z*-score for Example 8.2

There is about a 2% chance of getting a sample mean at least as high as 30 if the true population mean is 27. Most reasonable people would probably conclude that the average baggage weight is really higher than 27 based on these data.

Example 8.3: The management of a certain pizza parlor has believed that they average 56 customers per weekday. Recently they have been fearful that business has fallen off. A random sample of 16 days yielded the following numbers of customers:

51, 52, 60, 55, 59, 49, 56, 53, 58, 52, 55, 56, 48, 60, 56, 53

What do you think about management's concerns?

Note: The sample size of 16 days is too small to rely on the Central Limit Theorem. For the purposes of this example you are allowed to assume that the population in question is approximately normal.

The first step in answering this question is to calculate:

$$\bar{X} = 54.5625, s = 3.66912796, s_{\bar{X}} = \frac{3.66912796}{\sqrt{16}} = 0.91728199$$

Are these summary measures taken from the sample in substantial disagreement with the pizza parlor's standing belief that they average 56 customers per weekday? It is plain to see that the sample average is below 56. But, it's not much below. Is 54.5625 "far enough" below 56 to conclude that the pizza parlor is now getting fewer customers? Once again, let's calculate the probability of being at least as far below 56 as our sample mean of 54.5625.

$$t = \frac{54.5625 - 56}{0.91728199} = -1.5671299 \approx -1.6$$

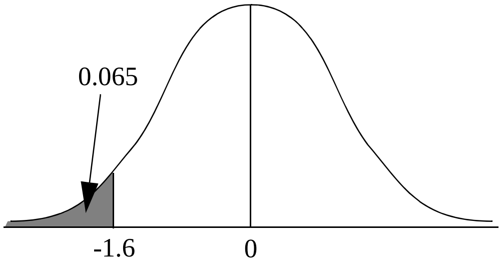


Figure 8.3
Schematic For the *t*-score for Example 8.3

If the belief of 56 customers per day is true, then there is a 6.5% chance of obtaining a sample mean at least as low as the one we obtained. This leaves us with a tough call. Many people would be willing to conclude from these results that the number of customers per day has fallen. Others may be reluctant to state that business has fallen off when there is still more than a 5% chance of our data being in harmony with the original belief of 56 customers per day. One thing is fairly clear. Even if this rather shaky data is enough evidence for you to conclude that business has fallen off, it probably hasn't fallen off a great deal.

Statistical decision making is really quite simple. You start with a statement of how things have always been believed to be (called the *null hypothesis*). You then gather data. The data is then compared with the null hypothesis. If the data is in reasonably close agreement with the null hypothesis then just go ahead and stay with that hypothesis. If, however, the data is in substantial disagreement with the null hypothesis, then you are forced to discard it. While the idea is simple, there are several words in the preceding sentences that should make you uneasy. The words I refer to are "reasonably close agreement" and "considerable disagreement." How are we to draw the line between these two possibilities? Statisticians have developed a very well defined methodology for doing this which is the subject of the next section.

Exercises

Concepts

1. A city dog catcher has over the years established that the average number of dogs caught per day is 37. She now believes that there is a new “puppy mill” in town breeding substandard puppies that are sold in pet stores, but which often end up unwanted and roaming the streets. Consequently, she believes that the number of dogs caught per day may have gone up. To test whether or not the old average of 37 dogs per day is still correct, she randomly selects 13 days and records the following numbers of dogs caught per day:

   39, 43, 41, 36, 46, 35, 45, 41, 50, 38, 37, 42, 31

   a. Just by *looking* at the numbers, what does your intuition tell you?

   b. Assuming that the population from which this sample is taken is approximately normal, follow the pattern outlined in the examples to calculate the approximate probability of getting a sample mean at least as large as the one that comes from this sample.

   c. Use the probability from part b. above to check your intuition from part a. above.

2. A house painter has from past experience found that a certain brand of house paint will cover 300 square feet per gallon when applied to wood siding. The paint company now claims their paint is “new and improved,” with greater coverage. He is not so sure. To test their claim, he randomly samples and applies 20 gallons of paint with the following results:

   312, 298, 288, 304, 306, 318, 290, 298, 301, 305, 286, 302,

   300, 291, 306, 302, 299, 305, 308, 298

   a. Just by *looking* at the numbers, what does your intuition tell you?

   b. Make a histogram with five classes.

   c. Does the histogram seem to support the assumption that the population from which the sample is drawn is approximately normal?

d. Applying your intuition to the histogram from part b., what can you say about the coverage of the “new and improved” paint?

e. Follow the pattern outlined in the examples to calculate the approximate probability of getting a sample mean at least as large as the one that comes from this sample.

f. Use the probability from part e. above to check your intuition as described in part d.

3. Suppose a random sample of 35 pieces of luggage is taken from the luggage of an airline. The airline claims that the average weight of a piece of luggage is 28.7 pounds. The labor union representing the baggage handlers claims that it is really higher than this. The random sample yielded the following information.

$$\bar{X} = 31.6 \text{ lb., } s = 14.3 \text{ lb.}$$

a. Just by *looking* at the sample results, what does your intuition tell you?

b. Follow the pattern outlined in the examples to calculate the approximate probability of getting a sample mean at least as large as the one that comes from this sample.

c. Use the probability from part b. above to check your intuition from part a. above.

d. Does your work require that you assume anything? Why, or why not?

A Look Back

4. Give a 95% confidence interval for the true mean number of dogs caught per day for exercise 1. Does this confidence interval help you decide if the number of dogs caught per day has really increased? If so, how?

5. Give a 95% confidence interval for the true mean number of square feet covered per gallon for the paint of exercise 2. Does this confidence interval help you decide if the coverage of the paint has really increased? If so, how?

8.2 The Scientific Method and Statistics

The scientific method has done a great deal for us. In the last couple centuries humankind has progressed far more in terms of technology than in all of recorded history before that time. A great deal of the responsibility for this explosion of knowledge is due to the development of the scientific method, and statistics is inseparably joined to the scientific method.

There are several facets of the scientific method that are little known or understood by the general public. The first is its *conservative nature*. The scientific method does not accept new ideas easily. This tendency to stay with past beliefs has no doubt saved society from a lot of crackpot ideas. On the other hand, a reluctance to accept new ideas can delay the acceptance of new ideas that really are good. Remember, there is *no way* to try to make decisions based on samples without the possibility of making a mistake.

Before going any farther, let's review the scientific method in some detail. What is the first step in the scientific method? Perhaps you said "formulate a hypothesis." If you did, you are on the right track, but there is something even before this. The first step in the scientific method is to be curious! That's right, be curious. Many great discoveries in science have been the result of serendipity. The right person in the right place at the right time *with a prepared and curious mind*. On the other hand how many great ideas just waiting to be discovered have been overlooked by people who just slog through life seeing little but their own problems? No doubt more opportunities have been missed than have been grabbed. Anyone who wants to be a scientist and carry out original research will find curiosity a most valuable ally. Without it a researcher will never know what to research!

Out of curiosity will grow the *research hypothesis*. This is the researcher's hunch of how things might be. But, until experimentation and data have their say, the research hypothesis will be doomed to be just a hunch. Along with the research hypothesis comes the *null hypothesis*. This is the hypothesis that basically says "nothing is going on."

Following the formulation of the null hypothesis is the design of the experiment. Ideally, a statistician should be involved in this step to make sure that the experiment really isolates the issue to be studied and controls for extraneous issues.

The next step is to carry out the experiment. This is where the expert in the field does her or his thing and the statistician just rests and trusts the scientist to do a proper job.

Then, the data obtained from the experiment are analyzed. This is the time for the statistician to get back into the picture to help with the analyses of the data. The honest researcher of integrity will also use the statistician as an unbiased person to be a check and balance on the ethical issues of reporting the facts free of bias and personal agendas.

When the cycle is complete, the results obtained from the data will invariably fuel further curiosity, which in turn will lead to another hypothesis and things will start all over again.

Unfortunately, this cycle is sometimes broken by researchers who take too much on themselves. Statisticians are all too often consulted only at the analyses stage, and then only when the analyses go beyond the routine. In some very unfortunate cases, many thousands of dollars worth of data must be trashed because precautions necessary to validate the research and its accompanying statistics were overlooked because a qualified statistician was not part of the planning stage.

Statisticians have developed a carefully defined vocabulary that describes the scientific method. The conservative old tried-and-true belief is called the null hypothesis and is denoted by H_0. This hypothesis is sometimes thought of as the “nothing is going on” or “things are as they’ve always been” hypothesis. The research or alternative hypothesis, denoted by H_A is usually the “pet” theory of the researcher. Of course, good researchers aren’t supposed to “want” any particular outcome, but humans being what they are, the truth is that most researchers can’t help but want their theory to be proven true.

As you well know by now, whenever you try to do anything based on a sample you run the risk of making incorrect judgments. This is very true when one is trying to decide between the research and null hypotheses. The following table shows the four possible outcomes to an experiment involving H_0 and H_A.

		Decision	
		Reject H_0	Don’t reject H_0
Reality	H_0 is false	Correct decision!	Type II error
	H_0 is true	Type I error	Correct decision!

A Look Back: Here is the reason that statisticians can’t always tell the truth. When testing hypotheses there are two ways to reach incorrect decisions, and when we are dealing with a sample, we can never be sure that the data has guided us in the right direction. It is always possible that a sample — even a random sample — may not accurately represent the population.

The statistical community should apologize for the names “type I error” and “type II error.” This is like naming your kids “child I” and “child II.” Statisticians have such a history of meaningful names (binomial, Central Limit Theorem, etc.) that it is a real shame to use such unimaginative names here. It also makes them tough to remember, but you must make the effort necessary to remember them because you will need a quick recall of what they mean as we get further into hypothesis testing.

In keeping with the conservative nature of statistics a type I error will in most situations be thought of as more serious than a type II error. A well-known and accessible example is our court system. The American judicial system runs on the scientific method. We hold the famous axiom “innocent until proven guilty.” The null hypothesis in this situation is “this person is innocent.” The “innocent until proven guilty” axiom contains the conservative nature of our legal system. It is considered a greater error to condemn an innocent man (type I error) than it is to set a guilty man free (type II error). Some in society might argue this point, but that is how our legal system works. Statistics also follows the “innocent until proven guilty” axiom in the sense that the scientific method demands that we not reject H_0 (the old fashioned tried-and-true belief) unless we have strong evidence telling us that we must. This is the reason for the terminology "don’t reject H_0" in the table above. Since we are conservative about rejecting H_0 we know that not rejecting H_0 doesn’t mean that H_0 is necessarily true. It just means that we didn’t have overwhelming evidence to disprove H_0.

Example 8.4: A bungee jumper has the null hypothesis “this bungee cord will not hold me.” The alternative hypothesis is “this bungee cord will hold me.” Comment on the nature and consequences of both types of error.

A type I error is rejecting H_0 when it is true. In this case the bungee jumper decides that the cord will hold when it really won't. The consequences are obvious and terrible.

A type II error is failing to reject H_0 when it is false. In this case the jumper incorrectly believes the cord will not hold. If the jumper has not yet jumped he will just get another cord and start over again. If he has jumped, he will have the scare of his life on the way down (believing the cord won't hold), but unless he has heart failure he will suffer no long-term ill effects.

Example 8.5: A paramedic arriving at the scene of an accident must test the null hypothesis "this victim can be saved." Comment of the nature and seriousness of a type I and a type II error.

Rejecting H_0 when it is true (a type I error) means that the paramedic will leave someone who might be revived. This is a very serious mistake indeed — the stuff of which lawsuits are made.

Failing to reject H_0 when it is false (a type II error) means that the paramedic will expend time and energy trying revive someone who cannot be revived. This wasted time might result in the death of other victims who are still alive.

Example 8.6: A pharmaceutical company has developed a new drug to combat high blood pressure. Their null hypothesis is H_0: "This drug is dangerous." Comment on the nature and consequences of both possible types of error.

First, a comment on the nature of the null hypothesis. It may seem a bit strange to have the null hypothesis state that there *is* a problem with the drug. After all, isn't the null hypothesis supposed to be the "nothing is going on" hypothesis? Yes, but in this case, saying that the drug is dangerous really is the "nothing is going on" hypothesis. From the perspective of the drug company, saying that the drug is dangerous is the same as saying it is useless. If it is dangerous it will have no market value.

If the drug is really dangerous, but is deemed safe (a type I error) the result will be the approval of a dangerous drug for use by the public. This could cost lives and result in lawsuits.

If the drug is really safe, but is deemed dangerous (a type II error) the drug company would keep a potentially useful drug from coming to market. This could also cause deaths. Such issues really do face pharmaceutical companies, and the federal agencies that police them, on a regular basis.

In this example, the company would deem the type I error as the more serious error. It is much more likely that they will be held responsible for marketing a dangerous drug than it is that they will be held responsible for withholding a safe drug from the market. In fact, it is fairly likely that the public will never even know of the existence of a drug that the company fears is unsafe.

A Look Ahead: The probability of a type I error (α) is very important, and will be seen on virtually every page throughout the rest of this book.

As in the previous example, it is most often the case that a type I error is the more serious of the two possible errors. In this spirit, the statistical methods that we will be studying make a decided effort to control the rate at which type I errors are made. A lot of attention is paid to the probability that a procedure will lead to a type I error. This probability of a type I error is very important and is denoted by α (pronounced “alpha”). The probability of a type II error is denoted by β (pronounced “beta”), and is also important, but is also much more difficult to deal with. There are statistical methods designed to deal with both α and β, but the methods that deal with β are much more sophisticated and are not in as wide use as those that control α. Formal methods to control β will not be covered in this book. One of the reasons that β is so difficult to deal with is that it depends on the unknown state of reality. If you are testing a null hypothesis which says that $\mu = 7$, while in reality $\mu = 200$ with $\sigma = 3$ it is very likely that even a test with small sample size will correctly reject H_0. That is to say, the probability of a type II error (β) is low, and the power $(1 - \beta)$ is high. On the other hand, if the null hypothesis says that $\mu = 7$, while in reality $\mu = 6.999$ with $\sigma = 3$ it is very likely that even a test with large sample size will fail to detect this very small deviation from the value specified by H_0. That is to say, the probability of a type II error (β) is high, and the power $(1 - \beta)$ is low. In most practical cases the situation is not as extreme as either of those just presented, but the problem remains. The only way to know the real power of a test is to know the values of μ and σ. But if we knew the values of μ and σ we certainly wouldn’t need to be testing hypotheses! About all that we can do about β is to try to keep sample sizes as large as possible.

One thing is certain. In any given situation, *the bigger the sample, the higher the power of the test*. In practical research the sample size is all too often determined by how much money is available to carry out the research. In such cases the value of β is beyond the control of the researcher. In those cases where funding will only allow a relatively small sample, the size of β is often much too high. It is sometimes the case that a deviation from the null hypothesis that is large enough to be of practical significance still has a small probability of being detected. This means that the *power of the test* $(1-\beta)$ is low and the probability of a type II error (β) is high.

The number α is chosen by the researcher. It should be chosen *before* the experiment is conducted so that no claim of bias can be made against the researcher. If α is picked after the data has been collected, in such a way as to prove some pet theory, then the researcher is in big trouble. As you will see in the next section, the choice of α can in some cases make all the difference in the decision of whether the null hypothesis is rejected. Remember, α is the probability of a type I error (rejecting H_0 when it is true). The smaller you make α the harder it is to reject H_0. α is chosen to be small when the consequences of a type I error are very serious. α can be larger when the consequences of a type I error are not so serious. Traditional values of α are 0.10, 0.05, and 0.01 with 0.05 being the most frequently used. There is absolutely nothing special about these values except their place in history. During the great depression mathematicians were not immune to the rampant unemployment of the time. Many of them were put to work in federally funded public works programs. One of the things they did was calculate the values in statistical tables such as the *z* and *t* tables. Since there were no electronic computers or calculators the work was very tedious. Tables were calculated for selected values and someone made some arbitrary choices as to what values of α should be included in the tables. Although computers can now fill out tables for any value of α in a fraction of a second, the traditional values that were originally put into tables still reign supreme because of historical inertia.

Since smaller α ‘s make it harder to reject the null hypothesis, they make it *easier* to make type II errors (failing to reject a false null hypothesis). In short, smaller α ‘s mean larger β ‘s.

The following box summarizes what you should know about hypothesis testing:

There are two hypotheses: the null hypothesis H_0 and the alternative hypothesis H_A.

The null hypothesis is the “nothing is going on” or “things are always as they have been” hypothesis. This hypothesis usually represents the status quo.

The alternative hypothesis (also called the research hypothesis) is the “something is going on” hypothesis.

There are two types of errors possible:

- Type I — Rejecting a true null hypothesis.
- Type II — Failing to reject a false null hypothesis.
- The probability of a Type I error is denoted by α. α is also called the *significance level* of the test. The lower the α, the more conservative the test.
- The probability of a Type II error is denoted by β.
- The larger the sample size, the smaller the β.
- The smaller the α, the larger the β.

The researcher has the luxury of predetermining the value of α, but controlling β is more difficult.

The double negative found in the terminology “fail to reject” H_0 bothers many students, but is necessary to accurately convey the conservative spirit of hypothesis testing.

Exercises

Concepts

6. A doctor has pioneered a new surgical procedure for aortic aneurysms. He is testing the null hypothesis: “The survival rate for this procedure is no better than that for the old procedure.” After taking a sample of cases treated using the new procedure he will compare the results to those already known for the old procedure. There is one additional fact that may influence your answers — the new procedure is considerably more expensive to perform than the old one.
   a. Comment on the nature and seriousness of the two types of error that can be made.
   b. How would you choose α for this problem?
7. A grocery store chain wants to know how the sales of its house brand bread compare to the sales of a nationally advertised brand. Their null hypothesis is “in our stores our bread sells no more than the national brand.” The alternative hypothesis states: “in our stores our brand *does* sell more than the national brand.”
   a. Comment on the nature and seriousness of the two types of errors that can be made.
   b. How would you choose α for this problem?
8. You are testing a new pig slop to see if it causes quicker weight gain than the old slop. The null hypothesis is “the new slop is no better than the old.” If, in reality the new slop *is* better than the old, but you fail to reject H_0, what type of error have you committed?
9. You are the principal of an elementary school. You believe that this year’s group of sixth graders is the smartest ever. The null hypothesis is “this is just an average class,” while the alternative hypothesis is “this is an outstandingly bright group of kids.” You test the hypotheses by giving them achievement tests (which of course aren’t perfect measuring devices). If, in reality they are just a normal group of kids, but you reject H_0 anyway, what type of error have you made?
10. In the real world of scientific research, which of the two types of error do you think are most often made? Why?

11. Give conditions under which making a Type I error would not be considered a serious problem.
12. Give conditions under which making a Type II error would not be considered a serious problem.
13. In a court of law, which type of error do you think is made more often? Why?
14. Suppose a medical researcher writes an exciting "breakthrough" journal article declaring that "people who hang upside down five minutes per day live longer than people who don't." Further, suppose that his research was well done and accurately analyzed, but the difference he found was that the "hangers" live an average of nine hours longer than the "non-hangers." How should he report his results? What does this tell you about the difference between statistical significance and practical significance?
15. Give conditions under which making a Type II error would be considered a serious problem?

8.3 One-Tailed Tests of the Mean

One of the most important parameters used to describe a population is the mean μ. Very often if we can say what the mean is with some degree of certainty we will be satisfied. In chapter seven you learned how to estimate the mean with confidence intervals. In this chapter our emphasis will be on making decisions. We will be testing hypotheses such as

$$H_0 : \mu = 112 .$$

If you were careful in your reading of the first two sections of this chapter the stage should be set for you. Recall that statistics operating hand-in-glove with the scientific method has two hypotheses — the null and alternative hypotheses. Recall further that our methods are conservative. They stick to the null hypothesis until they are forced by the weight of evidence to reject it. In many cases it doesn't take any sophisticated methods to tell whether or not H_0 should be rejected. Sometimes the data is in nearly perfect agreement with H_0, and it is clear that H_0 should not be rejected. At other times, the data are equally clearly out of line with H_0, and it is clear that H_0 should be rejected. It is the borderline cases in between that require all the formal machinery of statistics. After all, the only reason for the formal machinery of statistics is to have an objective and systematic way to decide the close calls, independent of any possible bias on the part of the researcher. The best way to learn the use of this formal machinery is through examples.

Example 8.7: An automobile manufacturer has just brought a new turbocharged V-6 engine into production. They believe that it will produce a higher peak horsepower than the old model which produced 212 horsepower. To test this they put 36 of the newly produced engines on the dynamometer and obtained the following results:

$$\bar{X} = 216 \text{ hp}, s = 12 \text{ hp}.$$

Formulate the null and alternative hypotheses and decide whether or not to reject H_0 using an α of 0.05.

The hypotheses are:

$$H_0 : \mu = 212 \text{ and } H_A : \mu > 212 .$$

Before making any calculations, note that with a sample size of 36 the Central Limit Theorem guarantees that we do not need to worry about whether the population of all the engine horsepowers is normal. To make the test, first calculate:

$$s_{\bar{X}} = \frac{12}{\sqrt{36}} = 2 .$$

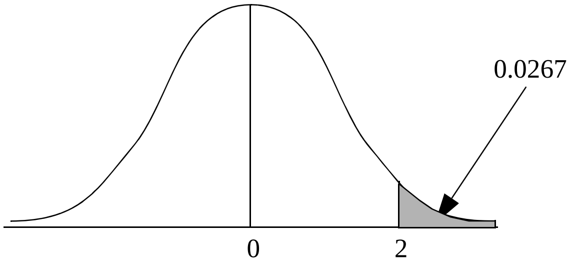
Use this result to calculate:

$$t = \frac{216 - 212}{2} = 2 .$$

Now, the question is, “is this t-score of 2 large enough to be quite sure that the data do not agree with H_0?” If we are to reject H_0 then we must be sure that there is no more than a 5% chance of making a mistake by so doing.

To help us make the decision, we ask a related question, “If H_0 is really true, what is the probability of getting a sample at least as far out of line with H_0 as ours?” This probability is easy to calculate. A picture will clarify the situation:

Figure 8.4
Schematic For the *t*-score For Example 8.7



The probability of obtaining a sample mean at least two standard deviations from the hypothesized mean of 212 is 0.0267. Since this is smaller than 0.05 we will reject H_0. The reasoning behind our rejection of H_0 is simple. The chance of obtaining a sample mean at least as large as ours from a population described by H_0 is slim (0.0267). So, we are led to believe that our sample results *did not* come from a population with mean 212.

The number 0.0267 from the example above, has a very special meaning, and its own name. It is called the *P-value*. The following definition precisely tells what the *P-value* is all about.

Assuming H_0 is true, the *P-value* is the probability of obtaining sample results at least as extreme as what was actually obtained.

If the *P*-value is as small or smaller than α, then the data is in substantial disagreement with H_0 and H_0 is rejected. If the *P*-value is larger than α then the data is not in substantial disagreement with H_0 and H_0 cannot be rejected.

Following the rules in the preceding paragraph H_0 will be rejected in error exactly $\alpha \times 100\,\%$ of the time. This is because we have set it up so that if H_0 is true, we will reject H_0 based on sample results that will occur only $\alpha \times 100\,\%$ of the time.

The approach to hypothesis testing presented here is called the *P-value approach* (for obvious reasons). There is another approach (called the *classical approach*). In keeping with the meaning of the word classical, this approach is the one that came first and was employed for a long time before the *P*-value approach became popular. However, the newer *P*-value approach is better than the older classical approach. The reason for this is that the *P*-value approach not only tells you whether to reject H_0, it also tells you how likely a population described by H_0 would be to give a sample like the one obtained. This probability tells us whether the decision is clear-cut or a close call. It gives us not only a decision, but some shading of degree as well.

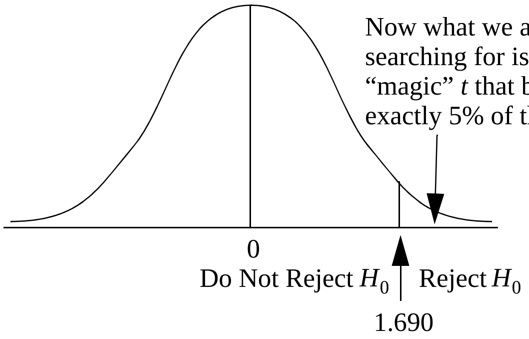
A Look Ahead: In the remainder of the book we will be emphasizing the *P-value* approach wherever possible. However, there are some cases where it will be impractical to reproduce tables large enough to find *P*-values for every situation. If you have a computer available, the *P*-value approach will still be the way to go. If not, you will then be forced to use the classical approach.

Although the previous paragraph declared the *P*-value approach "better" than the classical approach, there are still some reasons to learn the classical approach. First, some people find the classical approach simpler. Second, many studies reported in professional journals employ the classical approach. Third, studying the classical approach may enrich your understanding of the whole subject of hypothesis testing. Fourth, the inadequacy of some of our probability tables mentioned in the "Look Ahead" side-bar.

Example 8.8: We return to example 8.7 to illustrate the classical approach to hypothesis testing. There is a special table made just to facilitate this approach — Table 3 in Appendix B. This simple table gives *critical points* for hypothesis tests that use *t* or *z*-scores.

Returning to example 8.7, you are now asked to consider a new question: "Just how big would the *t*-score have to be to cause you to reject H_0?" Recalling that the α for this problem is 0.05, you should respond that we should reject H_0 for any value of *t* that would be expected to occur no more than 5% of the time. A picture will illustrate:

Figure 8.5
Schematic For the Classical Approach To the Horsepower Example



Now what we are searching for is the “magic” *t* that bounds exactly 5% of the area.

The magic *t* (found in Table 3 of Appendix B) is 1.690.

The classical method is indeed simple. It gives us this one number (1.690) upon which hangs the entire decision. If our calculated *t*-score is 1.690 or larger, then we will reject H_0. If our calculated *t*-score is smaller than 1.690 then we fail to reject H_0.

You may be snowed at this point. Let’s go back over what we’ve done and verify that it really all makes sense. What kind of *P*-values lead to rejection of H_0? The answer, “small ones.” This is because the *P*-value is the chance of a population described by H_0 yielding a sample like the one obtained. On the other hand, what kind of calculated *t*-scores lead to rejection of H_0? In this case the answer is “large ones.” Some students wonder why the *P*-value needs to be small, while the *t*-value needs to be large to reject H_0. The reason is that the *P*-value measures the *probability* of getting a sample at least as far from H_0 as the one obtained. If this value is small then the chances of H_0 “generating” our sample are small, which leads us to believe that H_0 is false. The *t*-score, on the other hand is a measure of distance. It measures how far the sample mean is from the mean specified in H_0. If this distance is *large*, it means the sample mean is far from the value hypothesized in H_0, again leading us to reject H_0. Also, note that a large *t*-value is precisely what will give rise to a small *P*-value. The *P*-value is computed from the tail area *beyond* the calculated *t*-score. The *larger* the *t*-score, the *less* tail area is left beyond it.

Summary:

A P-value $\leq \alpha$ leads to rejection of H_0.

A calculated t-score $\geq t_{\alpha, n-1}$ leads to rejection of H_0.

$t_{\alpha, n-1}$ refers to the t-value from Table 3 in Appendix B with tail area α and $n-1$ degrees of freedom.

Example 8.9: A tailor believes the neck size of shirts she has been ordering from a certain supplier may be larger than advertised. To check this, she takes a sample of 31 shirts that supposedly have a 15 inch neck size. She very carefully measures the actual neck size and obtains the following data (measured to the nearest sixteenth of an inch, although the data are presented here as decimals). She will use $\alpha = 0.05$ to make her test.

15.0, 15.25, 14.9375, 15.125, 15.5, 15.1875, 14.875, 15.25,
15.375, 14.9375, 15.125, 15.0625, 15, 15.25, 15.125, 15, 15.5,
15.125, 15.1875, 15.375, 15, 14.875, 15.0625, 15.0625, 14.875
14.9375, 14.8125, 15.1875, 15.5, 14.875, 15.35.

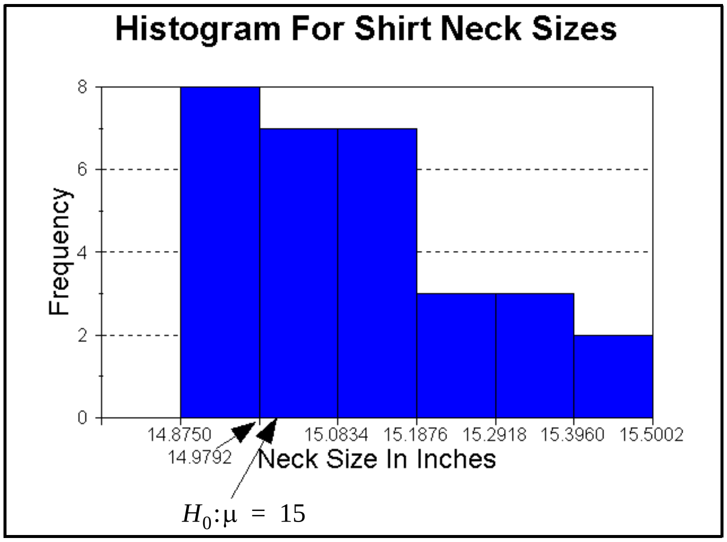
Test the tailor's research hypothesis that states "the shirt maker is making the necks too large" against the null hypothesis that states "the necks are the correct size."

The first step to solving this example is to carefully state the null and alternative hypotheses in mathematical form:

$$H_0: \mu = 15$$
$$H_A: \mu > 15$$

Before proceeding with the usual mechanics of hypothesis testing, let's take a look at a histogram of the data and see whether the data appears to coincide with H_0 or not.

Figure 8.6
Histogram For Shirt Neck Sizes



Two things should be clear from this chart. First, the data clearly do not appear to have come from a normal distribution. But, no matter. Remember the Central Limit Theorem? It's time to remind ourselves of its tremendous importance. Even though we are very likely sampling from a population that is not normal, the Central Limit Theorem tells us that the distribution of $\bar{X}$ will be well approximated by a normal distribution. The sample size of 31 is large enough to assure this. Second, there is a lot of the "weight" of the sample that is larger than 15. Visually, there appears to be considerable evidence that H_0 should be rejected.

Now it's time to pursue the formal mechanics of hypothesis testing to see if intuition is right. First, some preliminary calculations:

Note: Intermediate calculations should not be rounded, but the final *t*-score is rounded to 3.4 since our *t*-table is only given to one decimal place.

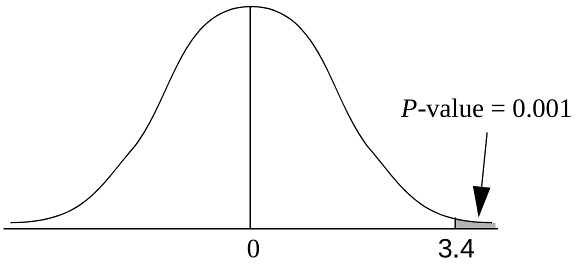
$$\bar{X} = 15.120165,\ s = 0.198554,\ s_{\bar{X}} = \frac{0.198554}{\sqrt{31}} = 0.035661$$

Next, a *t*-score is calculated[1]:

$$t = \frac{15.120165 - 15}{0.035661} = 3.3696475 \approx 3.4$$

Now, a trip to the *t*-table and a picture will help us to make the decision. When looking in the table, remember that the *t*-score is associated with $n - 1 = 31 - 1 = 30$ degrees of freedom.

Figure 8.7
Schematic For the Shirt Neck Size Example



The *P*-value (0.001) is *much* smaller than $\alpha = 0.05$, so H_0 is rejected with gusto! Intuition based on the histogram is exonerated.

As a final point, the critical *t*-value for this test is 1.697. Since the calculated *t*-score (3.4) is so much higher than the critical *t*-value, the same conclusion of rejection of H_0 is reached. Of course, it should be emphasized that the *P*-value and classical approaches *always* lead to the same conclusion when correctly applied. For this reason, it is not a bad idea for you to perform each test both ways just as a way to check your work.

Example 8.10: An agronomist is studying the moisture content of range grass (suitable for feeding cattle). She knows that in the spring the moisture content is 74%. She believes that in late summer the moisture percentage will be less. The following data represent the results of the laboratory analysis of 25 grass samples.

$$\bar{X} = 72.7\,\%,\ s = 5.0\,\%$$

State and test the null and alternative hypotheses with $\alpha = 0.10$.

The hypotheses are:

[1] At this point, many books would have you calculate a *z*-score since the Central Limit Theorem tells us sample means are normally distributed. Our reasoning for using a *t*-score anyway is based on two facts. First, for large samples the *t* and the *z* are essentially identical. Second, what little difference there is, is in the direction of the critical *t* values always being slightly larger than the corresponding *z* values. This makes the *t* test slightly more conservative than the *z* test, which seems a good idea when we realize that we are working with approximations anyway.

$$H_0: \mu = 74\%$$

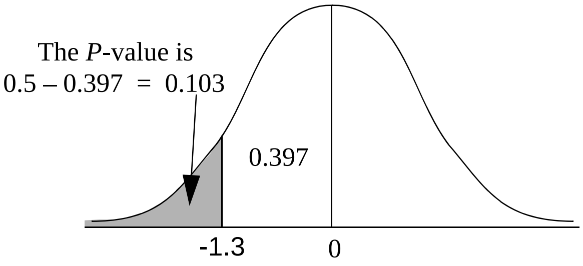
$$H_A: \mu < 74\%$$

Before any tests are made, recall that with a sample size of 25 the Central Limit Theorem may not come to our rescue. A sample size of at least 30 is needed to feel very confident that the sample mean can be accurately approximated by a normal distribution. However, a sample of 25 is large enough to still proceed as long as we realize that the *P*-value that we get may not be exactly correct. Another way out is to simply assume that the population of grass moisture contents is approximately normal. This assumption is probably at least approximately correct. Many, if not most, biological processes have distributions that are roughly normal. So, having decided that it is safe to proceed, the basic calculations are:

$$s_{\bar{X}} = \frac{5.0}{\sqrt{25}} = 1\%, \; t = \frac{72.7 - 74}{1} = -1.3$$

The picture is:

Figure 8.8
Schematic For the Moisture Content of Range Grass



Since the *P*-value is greater than the α of 0.10 we must fail to reject H_0.

Important Point: The paragraph to the right is the explanation of why we use the terminology “fail to reject H_0.”

The previous example powerfully illustrates why we should almost never say “accept H_0.” To accept H_0 when we know that the scientific method is conservatively leaning in the direction of H_0 in the first place, and when the decision is so borderline, would not be wise. In the previous example, it might well be possible to prove a difference if a larger sample size were taken in a subsequent study.

For these reasons the fallacy of the phrase "accept H_0" is once again emphasized.

Example 8.11: A forest ranger knows from past measurements that the average diameter of a mature Ponderosa pine on the east side of the mountain is about 3.1 feet. He hypothesizes that due to higher rainfall on the west side of the mountain, the average diameter of a mature Ponderosa on there is more than 3.1 feet. A random sample of 13 trees on the west side of the mountain yields the following summary statistics:

$$\bar{X} = 2.9 \text{ ft.}, s = 0.9 \text{ ft.}$$

State the hypotheses and make the test using $\alpha = 0.05$.

The hypotheses are $H_0 : \mu = 3.1$ ft. and $H_A : \mu > 3.1$ ft. This is admittedly a trick question. If you were about to calculate $s_{\bar{X}}$, the *t*-score and *P*-value, stop! It's time to get out your common sense again. The alternative hypothesis says that the mean should be *larger* than 3.1. The sample mean is *smaller* than 3.1, so there is *no* evidence against H_0. And, remember, the scientific method only rejects H_0 if there is *strong* evidence against H_0. Thus, you must fail to reject H_0, with gusto! In fact, this is the one and only type of situation where you could get by with saying "accept H_0."

Exercises

Basic Practice

16. The hypothesis $H_0 : \mu = 15$ is to be tested against

    $H_A : \mu > 15$ with $\alpha = 0.05$. A random sample results in:

    $$n = 49, \bar{X} = 16.0, s = 5.3.$$

    a. Find the *P*-value for the test.
    b. Decide whether to reject H_0.

17. Repeat the hypothesis test of exercise 16. using the classical approach.

18. The hypothesis $H_0 : \mu = 12$ is to be tested against

$H_A: \mu < 12$ with $\alpha = 0.05$. A random sample results in:

$$n = 64, \bar{X} = 11.4, s = 5.3 .$$

a. Find the *P*-value for the test.

b. Decide whether to reject H_0.

19. Repeat the hypothesis test of exercise 18. using the classical approach.

20. The hypothesis $H_0: \mu = 20$ is to be tested against

$H_A: \mu > 20$ with $\alpha = 0.05$. A random sample results in:

$$n = 36, \bar{X} = 20.9, s = 2.9 .$$

a. Find the *P*-value for the test.

b. Decide whether to reject H_0.

c. Does any assumption need to be made for these calculations to be valid? If so, what?

21. The hypothesis $H_0: \mu = 15$ is to be tested against

$H_A: \mu < 15$ with $\alpha = 0.01$. A random sample results in:

$$n = 16, \bar{X} = 13.8, s = 3.9 .$$

a. Find the *P*-value for the test.

b. Decide whether to reject H_0.

c. Does any assumption need to be made for these calculations to be valid? If so, what?

22. Repeat the hypothesis test of exercise 20. using the classical approach.

Applied

23. A consumer action group believes that a can of soup that supposedly has 670 mg. of sodium per serving really has a higher sodium content. To test this they take a random sample of 64 cans of soup and find:

$$\bar{X} = 712 \text{ mg}, s = 31 \text{ mg} .$$

State and test the appropriate hypotheses using $\alpha = 0.05$.

24. A manufacturer of ball bearings has reason to suspect that the machine that manufactures 3 cm. ball bearings is out of adjustment. A test is run to see if the machine is now producing bearings with an average diameter that is greater than 3 cm. Nineteen ball bearings are randomly sampled from a production run with the following results:

$$\bar{X} = 3.21, s = 0.08 .$$

State and test the relevant hypotheses with $\alpha = 0.05$. What assumption needs to be made to guarantee the validity of the test? Given the nature of the results of the test you performed above, do you think that concerns regarding this assumption are of great importance in this problem?

25. A teacher has kept careful records over a long career that indicate the average score on one of his final exams is $\mu = 76$%. Late in his career he has had a stroke of what he believes to be genius regarding his teaching style. He thinks that in the future his students will improve on the old final average of 76%. Viewing his next class of 41 students as a random sample, do the following results substantiate his claim at the $\alpha = 0.05$ level of significance?

$$\bar{X} = 79.3, s = 12.1 .$$

26. An automobile manufacturer thinks that they have improved on the horsepower of their 3.8 liter V6 engine. The old engine produced an average of $\mu = 145$ horsepower. Do the following results of a sample of 21 of the new engines substantiate the belief of improved horsepower at the 0.01 level of significance?

$$\bar{X} = 157, s = 17 .$$

27. The length of time it takes to dry a new dental bonding agent is under consideration. The old standard takes an average of $\mu = 53$ seconds. The hypothesis is that the new bonding agent takes less time to dry. Do the following results substantiate the belief of decreased drying time at the 0.05 level of significance?

$$n = 25, \bar{X} = 48.6, s = 15.2 .$$

A Look Back

28. The manufacturer of an automatic soft drink dispenser is concerned about whether a certain machine is filling cups with the correct amount of 10 oz. He decides to do a hypothesis test of $H_0: \mu = 10$ vs. $H_A: \mu \neq 10$ with the understanding that if the machine is malfunctioning, he will have to hire someone to perform expensive repairs.
    a. Describe a Type I error and the consequences of making one in this situation.
    b. Describe a Type II error and the consequences of making one in this situation.

8.4 Two-Tailed Tests of the Mean

This section really introduces only one new concept. Instead of considering alternative hypotheses such as $H_A: \mu > 15$, we will be considering hypotheses of the type $H_A: \mu \neq 15$. Because this type of hypothesis allows for the possibility of μ being either greater than or less than 15, it is called a two-tailed test.

The two differences between one and two tailed tests regard the way in which the *P*-value and critical *t*-scores are found. The philosophy has not changed, however. The *P*-value is still what it has always been — the probability of obtaining a sample result at least as far from H_0 as the one obtained. The difference is that there are now *two* ways to be far from H_0 — smaller or larger. This means that when calculating *P*-values you will need to double the value you have been getting. In the case of the classical approach the α must be *split* equally between the two sides. This means that if you are performing a two-tailed test with $\alpha = 0.05$ that when you look up the critical *t*-score you must look up 0.025 (exactly *one-half* of 0.05).

Example 8.12: The manager of a large department store keeps close tabs on the parking situation at her store. Past evidence has shown that about 612 cars have been parking in the store's parking lots on an average Thursday (not during any holiday season). She wants to know if this number has changed. To test this claim she has her security department carefully count all cars that use the lot during 10 randomly selected Thursdays (not near any holiday). The following results are obtained:

$$\bar{X} = 652.0 \text{ cars}, s = 112.8 \text{ cars}.$$

State and test the relevant hypotheses using $\alpha = 0.10$.

The word “changed” above is a tip-off. Change is a two-sided thing. The number of cars parking in the lot could go up or down. This single word “changed” is the entire key to understanding that this is a two-tailed hypothesis test. The hypotheses are:

$$H_0 : \mu = 612 \text{ and } H_A : \mu \neq 612 .$$

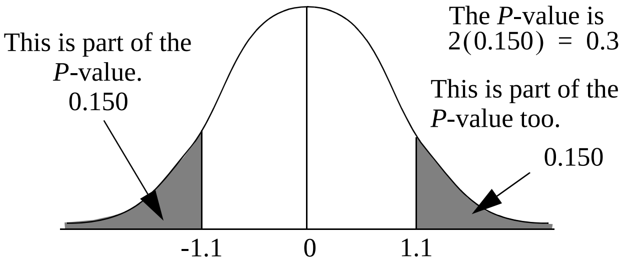
Note: With an *n* of only 10, all the subsequent calculations assume that the population of all numbers of cars parked per day is normally distributed.

The preliminary calculations are:

$$s_{\bar{X}} = \frac{112.8}{\sqrt{10}} = 35.670492, t = \frac{652.0 - 612}{35.670492} = 1.1213751 \approx 1.1 .$$

To find the *P*-value we resort once again to a picture.

Figure 8.9
Schematic For the Parking Lot Problem



Since the *P*-value (0.3) is greater than α (0.10) the null hypothesis cannot be rejected. Here again is an example of why one should not say “accept H_0.” The sample mean of 652 is 40 above the hypothesized mean, but the large standard deviation and the small sample size make it so that the difference would have to be huge to be detected. This is an example of a test with *low power.* There may indeed be an undetected but substantial change in the number of cars now using the parking lot. By failing to reject H_0 we are taking a large risk of making a type II error. One would not wish to make this possible error even worse by saying “accept H_0.”

The classical approach would have critical values of -1.833 and 1.833. The picture is:

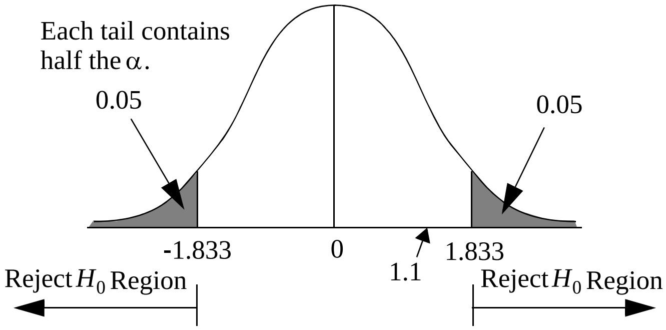


Figure 8.10
Classical Approach For the Parking Lot Problem

Since the calculated *t*-score of 1.1 is less extreme than the critical values, we are again led to the conclusion "fail to reject H_0."

Example 8.13: A fast food chain has changed suppliers for catsup. The old catsup had 25 calories per serving. Since they want to appear health-conscious to consumers the fast food chain constantly monitors all nutritional aspects of the food they serve. In the case of the new catsup they have no idea whether it has more or less calories than the old, so they wish to test:

$$H_0 : \mu = 25 \text{ vs. } H_A : \mu \neq 25 .$$

They will perform the test with $\alpha = 0.05$. To do so, they take a random sample of 25 individually wrapped servings and analyze the caloric content with the following results:

Note: As so often before, we should notice again that with a sample of size 25 we must either assume the population to be normal or consider our probability calculations to be approximations.

$$\bar{X} = 28.2 \text{ calories}, s = 1.1 \text{ calories}$$

This is a two-tailed test because of the "not equal" sign in the alternative hypothesis. The preliminary calculations are:

Note: It is always the *alternative* hypothesis that tells you whether the test is one or two tailed.

$$s_{\bar{X}} = \frac{1.1}{\sqrt{25}} = 0.22, t = \frac{28.2 - 25}{0.22} = 14.545455 \approx 14.55$$

At this point, you should be saying "WOW!" This is by far the biggest *t*-score that you have ever seen (at least in this book). One tail or two, it doesn't matter! *T*-scores of this size mean *REJECT* H_0 with great gusto! We won't even bother to look at a picture because 14.55 is so far out in the tail that there is virtually no area beyond it. The fact that the test is a two-tailed test means that the area beyond 14.55 must be doubled, but two times essentially zero is still essentially zero. (This *t*-score is so large that it is *way* beyond the bounds of the *t*-table.) Thus, the *P*-value is essentially zero.

Just for practice, what would the critical *t*-score be? With 24 degrees of freedom and a tail area of 0.025, the critical *t*-score is 2.064. Of course, there are really two critical *t*-scores — 2.064 and -2.064. Our calculated *t*-score of 14.55 is certainly outside the critical values, so the conclusion is once again reject H_0 !

Example 8.14: Suppose a pharmaceutical company has developed a new drug for insomnia. They know that the average time to fall asleep after taking the leading drug for insomnia is 21 minutes. They would like to test to see if they can demonstrate that the new drug results in a different amount of time necessary to fall asleep. To perform the test 100 randomly selected people with insomnia are given the new drug, and the time until sleep is carefully noted. Test the relevant hypotheses with $\alpha = 0.05$ and the following sample results:

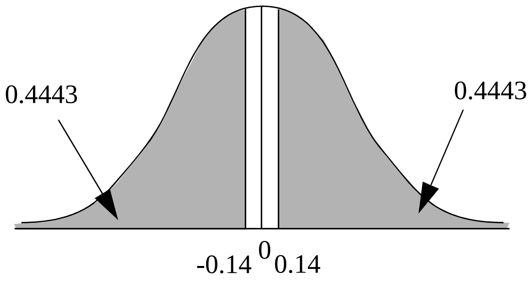
$$\bar{X} = 21.1 \text{ minutes}, s = 7.2 \text{ minutes}.$$

With a sample of 100 there is absolutely no need to worry about whether the population of times to fall asleep is normally distributed. A quick look at the sample results shows that the sample result of $\bar{X} = 21.1$ minutes is in strong agreement with the null hypothesis. Even if our statistics can demonstrate a significant difference between the old and new drugs, it is probably not large enough to be of much practical importance. To formally tackle the problem we calculate:

$$s_{\bar{X}} = \frac{7.2}{\sqrt{100}} = 0.72, t = \frac{21.1 - 21}{0.72} = 0.13888889 \approx 0.14 .$$

With a sample of 100 the bounds of the *t*-table are exceeded, and we may use the standard normal table. It is really unnecessary, however. The calculated *t*-score of 0.14 is very small as *t*-scores go. It is so small in fact that there should be no need to do any further math. The decision is, “fail to reject H_0.” For the practice, and in hopes of your deeper understanding, the following details are provided in spite of the fact that the decision is rather clear-cut.

Figure 8.11
Schematic For the Insomnia Problem



The P-value is $2(0.4443) = 0.8886$. This is certainly larger than α, leading to the decision “fail to reject H_0.” This is also clear from the fact that the sample mean fell very close to the hypothesized population mean. This is what caused the very small t-score of 0.14.

This leaves the drug company free to make its decision of whether to market the drug based on other factors such as cost and side effects.

Exercises

Basic Practice

29. The hypothesis $H_0: \mu = 15$ is to be tested against

    $H_A: \mu \neq 15$ with $\alpha = 0.05$. A random sample results in:

$$n = 32, \bar{X} = 16.0, s = 5.3.$$

    a. Find the P-value for the test.
    b. Decide whether to reject H_0.
    c. Does your work require the truth of an assumption to be valid?

30. Test the hypothesis of exercise 29. using the classical approach.

31. The hypothesis $H_0 : \mu = 20$ is to be tested against $H_A : \mu \neq 20$ with $\alpha = 0.01$. A random sample results in:

$$n = 100, \bar{X} = 19.1, s = 5.7 .$$

a. Find the *P*-value for the test.

b. Decide whether to reject H_0 .

c. Does your work require the truth of an assumption to be valid?

32. The hypothesis $H_0 : \mu = 15$ is to be tested against $H_A : \mu \neq 15$ with $\alpha = 0.05$. A random sample results in:

$$n = 8, \bar{X} = 13.8, s = 3.9 .$$

a. Find the *P*-value for the test.

b. Decide whether to reject H_0 .

c. Does your work require the truth of an assumption to be valid?

33. Test the hypothesis of exercise 32. using the classical approach.

Applied

34. Suppose an engineer has just developed a new shape for steel beams used in heavy construction. The new beam will be cheaper to manufacture, but it is not known how the strength of the new beams will compare with the strength of beams made with the old design which can hold an average of 12.3 tons. State and test the relevant hypotheses with $\alpha = 0.05$ and the following random sample results:

$$n = 10 , \bar{X} = 14.3 \text{ tons}, s = 1.1 \text{ tons}.$$

Does your work require the truth of an assumption to be valid?

35. An English professor is trying to determine authorship of an old literary document. She has counted the frequency of occurrence of a word that is known to appear with an average frequency of 3.2 times per paragraph in the works of Shakespeare. If she can prove that the frequency of the word is

significantly different from 3.2 then she would have evidence that Shakespeare was not the author. Considering the document's 31 paragraphs to be a random sample from the population of all possible paragraphs from whoever wrote the work, test to see if the frequency of use of the word is different from that of Shakespeare using $\alpha = 0.01$. The following results summarize the sample:

$$\overline{X} = 3.6, s = 2.1 .$$

Can Shakespeare be ruled out as a potential author based on this evidence? Does your work require the truth of an assumption to be valid?

36. A contractor does not know whether a new technique will allow his crew to frame a certain model of house in more, less or the same amount of time than it used to take them ($\mu = 71$ hours). Do the following data support the belief that the new technique finishes framing a house in a different amount of time than the old technique? Use the $\alpha = 0.10$ level of significance?

    73, 68, 59, 71, 70, 66, 68, 72, 68, 67, 66.

37. A new flea medication for dogs has been developed. Is there sufficient evidence to conclude at the $\alpha = 0.01$ level of significance that the duration of the new product differs from that of the old one (average effective time of 15 days)? The sample results (in days) follow.

    14, 13, 16, 15, 17, 18, 14, 15, 15, 15, 15, 16

A Look Back

The following exercises are a mixture of one and two-tailed situations. Textbook exercises are often misleadingly simple because they are sectioned in such a way that you know what kind of a test is involved just by where you are in the book. It is the purpose of these exercises to help you develop the critical thinking skills necessary to distinguish between one and two-tailed tests.

38. A pet food company has just created a new formula for cat food. From extensive testing they know that cats eat an average of 4.1 ounces of the old formula per meal. They want to

test to see if the new food has improved palatability (and will therefore be consumed in larger quantities). Should a one or two-tailed test be used?

39. The mayor of a city wants to know if library usage at the city's 12 public libraries has changed. Past evidence has indicated that 2,112 persons visit the libraries per day. A random sample of days is to be taken during the next year to see if the number of persons using the library has changed. Should a one or two-tailed test be used?
40. A building contractor knows that on average it takes a crew 13 days to build a certain design of tract house. He and an architect have just redesigned the home in an effort to make it easier to build. They want to test whether it really can be built quicker. Should they use a one or two-tailed test?
41. An animal psychologist is testing the learning rate of mice. Past research has shown that it takes mice who have practiced a certain maze many times an average of 31 seconds to make it through it. The psychologist believes that a new mouse training technique may help mice to complete the maze in a shorter period of time. Is a one or two-tailed test applicable?