Lesson 8
Hypothesis Tests for Proportions

Outline of the Lesson

As we have discussed in the previous two lessons, *inferential statistics* has two important features:

- Information is obtained from a *sample*.
- The information from the sample is used to draw a conclusion (an *inference*) about the entire *population* from which the sample was drawn.

This course covers two major types of inferential statistics, *confidence intervals* and *hypothesis tests*. In the previous two lessons we studied the use of confidence intervals to draw conclusions about population

proportions. In this lesson we learn about hypothesis tests, again in the context of population proportions. Hypothesis tests, like confidence intervals, are used to make a statement about a population proportion:

- We use a confidence interval to indicate what we believe the population proportion *is*.
- We use a hypothesis test to indicate what we believe the population proportion *is not*.

Both methods seek to establish a conclusion about a population proportion. Both use evidence from a sample drawn from that population. The difference is in the emphasis ("is" vs. "is not").

To illustrate how the two are related, consider the example first presented in Lesson 6. A sample of Americans age 18 and above was asked the question, "Do you believe the amount of taxes you pay is fair?" In that lesson we reported the conclusion that 54% of all Americans age 18 and above do believe the amount of tax is fair, with margin of error ±4.2%. Written as a confidence interval, the result is (49.8%, 58.2%). For this same survey, a hypothesis test would be designed to answer questions such as these:

- Do you believe that 60% of Americans, age 18 or above, believe the amount of tax they pay is fair?
- Do you believe that 50% of Americans, age 18 or above, believe the amount of tax they pay is fair?

Based on the confidence interval (49.8%, 58.2%), we can see that the answers to the questions are, respectively:

- I believe the proportion ***is not*** 60%, because 60% is not within the confidence interval I calculated.
- I believe the proportion ***could be*** 50%, because 50% is within the confidence interval I calculated.

In the second answer, we do not conclude that the proportion ***is*** 50%, only that it ***could be*** 50%. The reason is clear if you think about the confidence interval. Although 50% is within the confidence interval, there are also many other plausible values within the interval.

To begin to move toward a *hypothesis testing* way of thinking, let's do a slight rewrite of the questions and conclusions, using a claim in place of a question. Here is the first question, rewritten:

> *Claim*. The population proportion is 60%.
> *Result of survey*. 54% was the proportion for the sample.
> *Conclusion.* I believe the claim **is not** correct.

and here is the second:

> *Claim*. The population proportion is 50%.
> *Result of survey*. 54% was the proportion for the sample.
> *Conclusion.* I believe the claim **could be** correct.

Notice that the claim about the population proportion is rejected when the distance between the claimed proportion and the proportion for the sample is too large – so large that the claimed proportion does not fall within the confidence interval. When that distance is smaller, we acknowledge that the claim might be correct.

You may wonder

Since the questions posed in the example can be answered using the confidence interval method we have already learned, why should we study the second form of inferential statistics, hypothesis tests?

The answer is simple: *the method of hypothesis tests applies to many situations for which confidence intervals are not applicable*. For example, as we will see in Lesson 10, we can use hypothesis tests to study whether there is an association between two variables such as political party affiliation and gender.

As we did for confidence intervals, we begin our explanation by thinking about probabilities for random events such as coin tosses and rolls of dice. Beginning in Section 8.5 we will apply what we learn to the polling process.

8.1 – An Example: Testing Dice for Fairness

Most people realize that when a coin is tossed, the probability that the result will be *heads* is ½, or 0.5, or 50%. Since there are two sides, and since each side is an equally likely outcome, we calculate the theoretical probability as 1 divided by 2 = 0.5. As we discussed in Lesson 6, what this implies is the following:

> *If the coin is tossed a large number of times, the proportion of heads will be approximately 0.5 = 50%. In addition, the more times the coin is tossed the closer to 50% you can expect the proportion to be.*

This same idea applies to more complicated probabilities, as we discuss in the following example.

Fair and loaded dice

If we roll a pair of fair dice, the total on the two dice can be anything between 2 and 12, with different probabilities for the different results (obviously a 7 is more likely than a 2, since there is only one way to get a total of 2 and there are lots of ways to get a total of 7). It turns out that the probability of obtaining a total of 7 is 6 out of 36, or 1/6, or approximately 0.1667 = 16.67%. As for the coin toss situation, this means that if we roll a pair of fair dice a large number of times, the proportion of 7s will be approximately 0.1667.

It is possible to alter the dice to modify the probabilities for the various outcomes; such a pair of dice is commonly referred to as *loaded*. For use in a board game to be played in the home, this would not be of much concern. However, in a casino the use of loaded dice would be quite serious. State gaming commissions carry out an extensive battery of inspections and tests to ensure that casinos are using equipment, including dice, which perform as intended and as advertised. We will take a look at a very simple such test; the actual testing used in practice is far more sophisticated than what we do in this example.

For a particular pair of dice, we want to answer this question: *The dice manufacturer claims that the dice are fair. In particular, it claims that the probability (that is, long-term proportion) of rolling a 7 is 1/6 ≈ 16.67%. Should we believe the claim or not?* An obvious way to investigate the claim is to start

rolling the dice, observing the percentage of 7s obtained. Of course, even for fair dice we cannot expect the proportion in such a sampling to be *exactly* 1/6. However, we would expect it to be *close to* 1/6. Moreover, the more times we roll the dice the closer to 1/6 the proportion should be. The experiment can be summarized as follows:

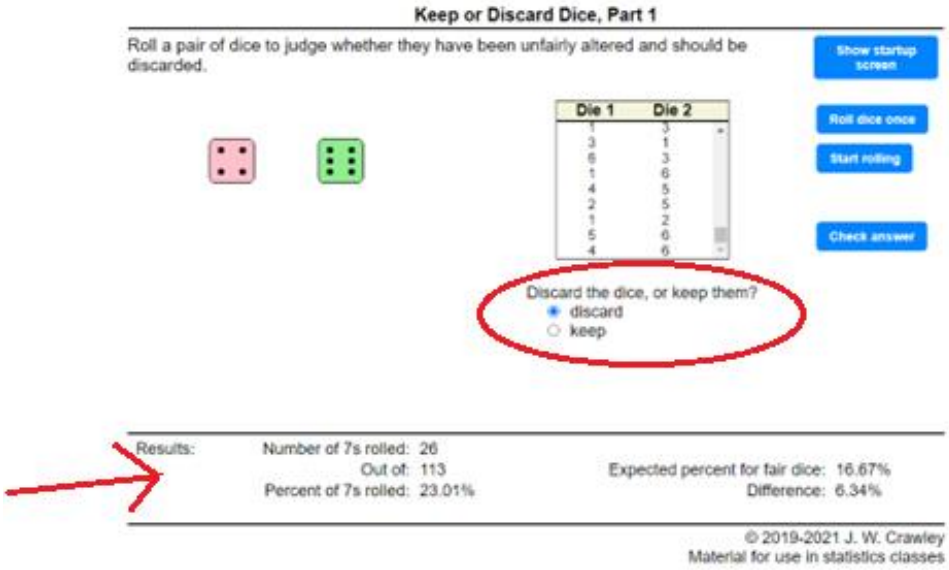
- Roll the dice many times.
  - If the proportion of 7s is not close to 1/6, we have evidence that the probability is not 1/6. We will **reject** the claim, and conclude that the dice are loaded. We will describe our decision as "discard the dice."
  - If the proportion of 7s is close to 1/6, we will acknowledge that the claim **could be** true. We will describe our decision as "keep the dice."

An applet for experimenting

Just as we did for confidence intervals, we have developed a series of applets to allow you to experiment with this concept. The first applet allows you to roll the dice as many times as you like, either by clicking repeatedly on the "Roll dice once" button, or by using the "Start rolling" and "Stop rolling" buttons. The applet keeps track of how many 7s are rolled, and the proportion of 7s rolled. Here is the link to the applet:

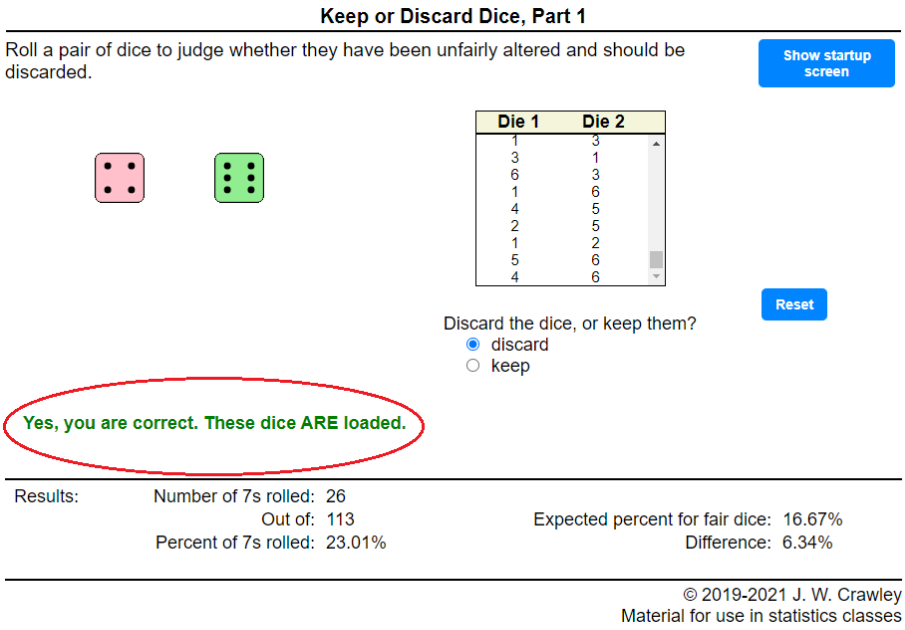
Keep or Discard Dice Part 1

For example, here is what happened when the author used the applet to roll the dice 113 times.

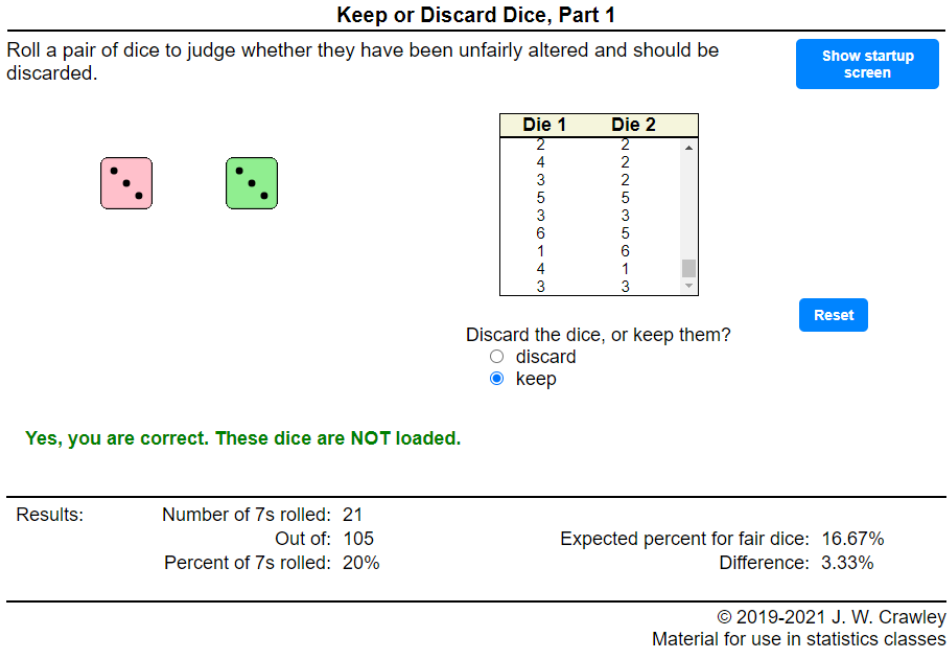


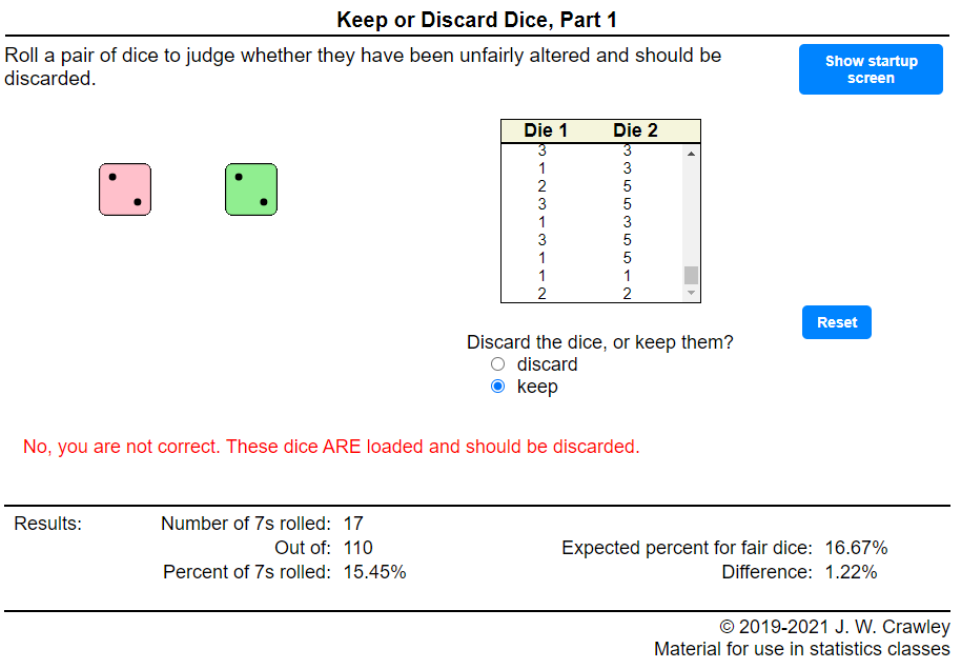
Refer to the results (highlighted with an arrow) and the author's decision (circled). In this experiment, 26 out of the 113 rolls (23.01%) were sevens. This is a good deal larger than the 16.67% we would expect for fair dice, so the author is concluding that the dice are loaded and should be discarded.

The applet was written so that the applet itself knows the correct answer to the question. When the author clicked on the "Check answer" button he was told the correct answer, as shown here:



The author did a second sample, for a second pair of dice, then a third sample, with yet another pair of dice, with these results:





In the second sample, the difference between the expected percent and the actual percent was 3.33%. The author made the judgement that this was "close enough" and decided to keep the dice. A similar judgement for the third sample led to the same conclusion. Notice that in one case the author was correct, but in the other he was incorrect.

Exercise 1: Use the applet several times, each time rolling the dice at least 100 times. Record your results here: the difference between the percentage of 7s and the expected value of 16.67%, your decision ("yes" or "no" to the question if you believe the dice are loaded), and whether you were correct or not. The first three rows are filled in based on the results shown above. **Note:** *The author had decided to keep the dice if the difference was under 4%, but you may use your own strategy for your decisions.*

Difference	Do you discard the dice?	Correct or incorrect?
6.69%	yes	correct
3.14%	no	correct
2.11%	no	incorrect

Types of errors; sample size; how close is close enough?

When you decide whether you believe the dice should be discarded or kept, sometimes you are correct and sometimes you are incorrect. When you roll the dice (whether using the applet or using real dice), you make two choices. First, you decide how many times to roll the dice before making your decision. Second, you decide how far away from the expected 16.67% your sample must be to choose the *discard* option (or, equivalently, how close to 16.67% your sample must be to choose the *keep* option). We will informally refer to this as a "measure of closeness." If, based on this chosen measure, the proportion of 7s is close to the expected 16.67%, you will keep the dice. Otherwise, you will discard the dice.

There are two different ways to be correct, and two different ways to be incorrect, as shown in this table:

		Your decision	
		Keep	**Discard**
Actual status of dice	**Fair**	Correct	Incorrect
	Loaded	Incorrect	Correct

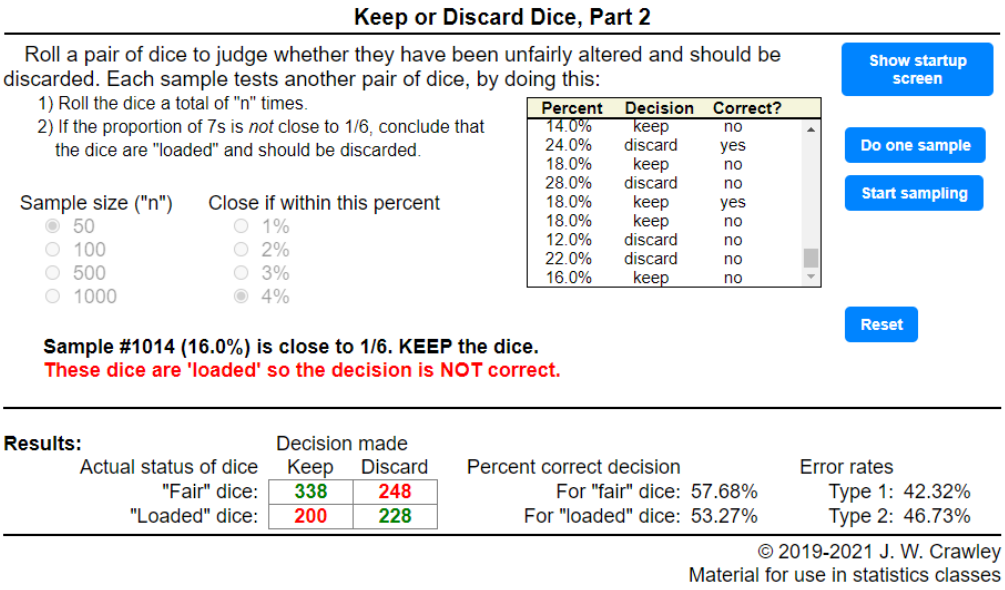
In the first type of error, we have incorrectly discarded fair dice. The manufacturer's claim, "The proportion of 7s is 16.67%" was true, but we have concluded that it was false. This is called a Type 1 error, and it can be controlled in a predictable manner by our choice of sample size and measure of closeness. We will study this in more detail as we proceed.

In the second type of error, called Type 2 error, we have allowed a loaded pair of dice to sneak past us. Although sample size and measure of closeness play a role in controlling this type of error, so also does the amount by which the dice have been altered. (Dice that roll a 7 nearly 100% of the time would be much less likely to sneak past us than dice whose proportion of 7s has been modified only slightly.) This makes the study of Type 2 error much more complex, and it is well beyond the scope of this lesson, and indeed of this course. However, one should always keep in mind that the more one reduces Type 1 error, the more likely one is to encounter Type 2 error.

We have created a second applet to allow you to experiment with these concepts, at this link:

Keep or Discard Dice Part 2

This "Keep or Discard Dice, Part 2" applet lets you choose a sample size and a measure of closeness. For example, when you set the measure of closeness at 4%, this means that: if the sample's proportion is more than 4% distant from the expected 16.67%, you will conclude that the dice were loaded and should be discarded; otherwise your decision will be to keep the dice. You can then run multiple samples of this size using this measure of closeness, and the applet will keep track of how many times the resulting decision is correct and incorrect, as illustrated in this sample run:



Keep or Discard Dice, Part 2

Roll a pair of dice to judge whether they have been unfairly altered and should be discarded. Each sample tests another pair of dice, by doing this:

1) Roll the dice a total of "n" times.
2) If the proportion of 7s is *not* close to 1/6, conclude that the dice are "loaded" and should be discarded.

Sample size ("n"): 50, 100, 500, 1000

Close if within this percent: 1%, 2%, 3%, 4%

Percent	Decision	Correct?
14.0%	keep	no
24.0%	discard	yes
18.0%	keep	no
28.0%	discard	no
18.0%	keep	yes
18.0%	keep	no
12.0%	discard	no
22.0%	discard	no
16.0%	keep	no

Show startup screen | Do one sample | Start sampling | Reset

Sample #1014 (16.0%) is close to 1/6. KEEP the dice.
These dice are 'loaded' so the decision is NOT correct.

Results:

Actual status of dice	Decision made: Keep	Decision made: Discard
"Fair" dice:	338	248
"Loaded" dice:	200	228

Percent correct decision: For "fair" dice: 57.68%; For "loaded" dice: 53.27%

Error rates: Type 1: 42.32%; Type 2: 46.73%

How good a strategy was it to use a sample size of 50 rolls, with 4% as the measure of closeness? Let's analyze the results for both fair dice and loaded dice.

Fair dice: In this sample run, 586 (338 + 248) pairs of fair dice were tested, and we falsely accused 248 of those of being loaded dice. We were correct only 57.68% of the time, with Type 1 errors occurring the other 42.32% of the time.

Loaded dice: On the other hand, 428 (200 + 228) pairs of loaded dice were tested, and 200 of these slipped through our testing. We were correct only 53.27% of the time, with Type 2 errors occurring the other 46.73% of the time.

Obviously, using a sample of size 50 with 4% as our measure of closeness is not a very good testing strategy!!

Exercise 2. a. Use the applet to run between 900 and 1000 samples, with sample size 50 and closeness measure 4%. Record the results; are your results fairly consistent with ours?

b. Do the same, but using sample size 1000 and closeness measure 4%. Record the results, and comment.

c. Do the same, but using sample size 1000 and closeness measure 2%. Record the results, and comment.

d. Experiment with other combinations of sample size and closeness measure. Suppose you want to keep the Type 1 error rate near or below 5%. Are there any combinations that seem to meet this goal?

Note: Refer to the answer keys at the end of the section to see what happened when the author did this exercise. Your results will not match those exactly, but will likely be similar.

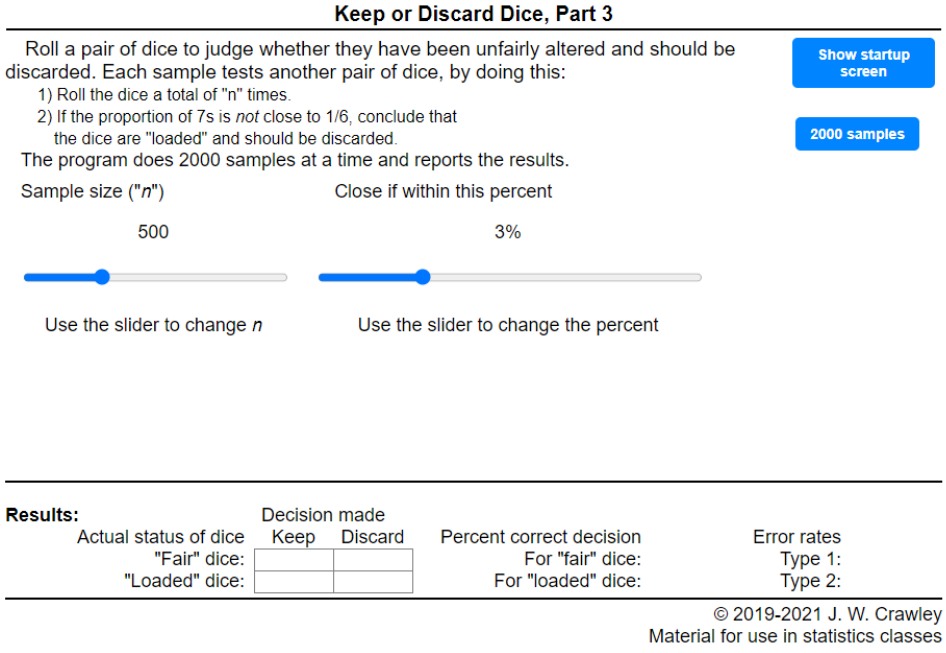
Controlling Type 1 error

Several conclusions can be drawn from the previous exercise.

- For a given measure of closeness, we can reduce the percentage of Type 1 error by using a larger sample size.
- For a given sample size, we can reduce the percentage of Type 1 error by using a larger measure of closeness. (However, this comes at the expense of increasing the percentage of Type 2 error.)
- Perhaps most importantly, the results are consistent. In particular, the percentage of Type 1 error (discarding fair dice) was very nearly the same for you as it was for the author, and for the others in your class who did the exercise.

As we stated earlier in this lesson, Type 1 error can be controlled in a predictable manner by our choice of sample size and measure of closeness. The results of Exercise 2 illustrate this fact. In this section, we will examine this phenomenon a bit further.

One shortcoming of the applet we have been running is that the choices for sample size are relatively limited (50, 100, 500, or 1000). Another is that the choices for closeness measure are also limited (1%, 2%, 3%, or 4%). A third limitation is that we must wait while the applet runs sample after sample, so that seeing what happens for perhaps 900-1000 samples takes some time. To allow for more rapid and more detailed experimentation, the author has developed the "Keep or Discard Dice, Part 3" applet, which overcomes all three of these difficulties. Here is a screen shot from that applet:



The slider bars can be used to choose any sample size from 100 to 1500 (in increments of 100), and any measure of closeness from 0.5% to 10.0% (in increments of 0.1%). Moreover, a single click on the "2000 samples" button will:

- generate 2000 samples, each of the indicated sample size
- use the indicated measure of closeness to make a decision about each of the 2000 samples
- record the results in a form similar to the previous applet.

So we can very quickly generate a large number of samples and get a reasonable feel for the Type 1 (and Type 2) error for that combination of sample size and closeness measure. For example, here are the results obtained for 10,000 samples using sample size 500 and closeness measure 3%. (The author simply clicked on the "2000 samples" button five times.)

Results:	Decision made			
Actual status of dice	Keep	Discard	Percent correct decision	Error rates
"Fair" dice:	5554	428	For "fair" dice: 92.85%	Type 1: 7.15%
"Loaded" dice:	2349	1669	For "loaded" dice: 41.54%	Type 2: 58.46%

You should use this link to run the applet yourself; your results should be quite similar to these.

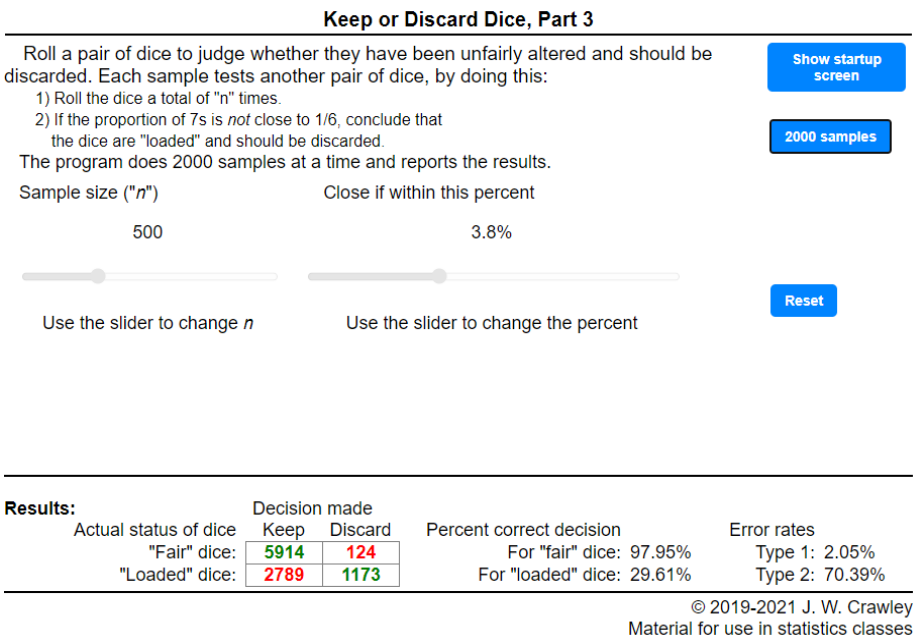
Keep or Discard Dice Part 3

In the author's results recorded above, we can summarize what happened as follows:

- By clicking on the "2000 samples" 5 times, the author repeated the same experiment 10,000 times.
- Each experiment tested a pair of dice. Sometimes the dice being tested were "fair," sometimes they were "loaded."
- To test a pair of dice, the experiment consisted of 500 rolls of the dice. The percentage of 7s was recorded, and compared to the theoretical proportion that should occur for fair dice (1/6, or approximately 16.67%).
- Based on this percentage of 7s, one of two possible decisions was made:
    - *Discard the dice*. If the percentage was below 13.67% or above 19.67% (that is, *not within* the chosen 3% measure of closeness), the decision was to discard the dice.
    - *Keep the dice*. If the percentage was between 13.67% and 19.67% (that is, *within* the chosen 3% measure of closeness), the decision was to keep the dice.
- For fair dice, this resulted in a correct decision 92.85% of the time; therefore, it resulted in an incorrect decision for fair dice 7.15% of the time.
- Discarding fair dice is referred to as Type 1 error. Thus, the strategy of using sample size 500 and closeness measure 3% led to having Type 1 error for 7.15% of the fair dice.

In general, statisticians like to keep Type 1 error a little lower than this. For most statistical studies, the person conducting the study will have chosen, in advance, to use a strategy that restricts the Type 1 error rate to one of these two values: either to 5%, or to 1%. Which is chosen depends on the nature of the study.

We can use this third applet to experiment with strategies to achieve these two goals. In the first example, let's stick with a sample size of 500, and try to find a closeness measure that will keep the Type 1 error rate to approximately 5%. As we found in Exercise 2, for a given sample size we can decrease the Type 1 error by using a larger measure of closeness. Here are the results using 3.8% as the closeness measure (again, the author ran 10,000 trials by clicking the "2000 samples" button five times).



Our goal is to keep Type 1 error very close to 5%. It appears we have overshot this goal, so we try a closeness measure somewhere between 3% and 3.8%. After a little experimentation we settled on 3.3% as our measure of closeness, with a resulting Type 1 error rate of 4.96%. You should try this same experiment; your results will not be identical to ours, but they should be similar.

As we indicated, sometimes statistical studies are designed with the goal of reducing the Type 1 error rate to 1% instead of 5%. Still using a sample size of 500, what choice for measure of closeness could we use to achieve this goal? In the results shown above, using 3.8% led to a Type 1 error rate of 2.05%, so we try something a bit larger than 3.8%. Again, a little experimentation allows us to find a choice which seems to work. We tried 4.1% (Type 1 error rate was 1.59%), then 4.5% (Type 1 error rate was 0.75%), then finally settled on 4.2% (Type 1 error rate was 0.95%). Again, if you try the same experiment, your results should be similar but certainly not identical.

Exercise 3. Use the applet to fill in the following table. The first row is already filled in based on the author's experimentation described above.

Sample size	**Measure of closeness to achieve indicated Type 1 error rate**	
	5%	1%
500	3.3%	4.2%
800		
1000		
1400		

Note: Refer to the answer keys at the end of the section to see what happened when the author did this exercise. Your results will not match those exactly, but will likely be similar.

8.2 – The Logic of Hypothesis Testing

The procedure we used to decide whether to keep or discard the dice is an example of what statisticians call **hypothesis testing**. Carrying out that procedure is referred to as a **hypothesis test.** In this example, the hypothesis we are testing can be described as, "This pair of dice is fair," or more precisely as, "The long-term proportion of 7s for this pair of dice is 1/6." We test that hypothesis using the procedure outlined in the previous section.

The general logic of hypothesis testing for population proportions is identical to these procedures for examining the fairness of dice. It can be summarized as follows:

There is a claim, or hypothesis to be investigated, involving a probability / long-term proportion. For the dice example we can write this claim as follows:

- The (long-term) proportion of 7s for this pair of dice is 1/6.

Procedure:

- Take a sample of size *n*. (For the dice example, we did this by rolling the dice *n* times.)
- Measure the proportion for that sample (the *sample proportion*). (For the dice example, we measured the proportion of 7s obtained when we rolled the dice *n* times.)

Decision:

- If the proportion in the sample *is not* close to the proportion in the claim, reject the original claim. (For the dice example, we referred to this as "discard the dice.")
- If the proportion in the sample *is* close to the proportion in the claim, acknowledge that the original claim *could* be true; that is, *do not* reject the claim. (For the dice example, we referred to this as "keep the dice.")

Possible errors you could make using this procedure

- Type 1 error: Rejecting a true claim. (For the dice example, this meant discarding a pair of fair dice.)
- Type 2 error: Failing to reject a false claim. (For the dice example, this meant keeping a loaded pair of dice.)

Notation: The notation we use is identical to what we have encountered in our coverage of confidence intervals.

1. We use the variable *n* to stand for the sample size. In the dice example, this is the number of times we roll the dice.
2. The sample proportion is denoted by the variable $\hat{p}$ (read as *p*-hat). In the dice example, this is the proportion of 7s we obtain when we roll the dice *n* times.
3. The variable *p*, without the "hat," indicates the probability being examined, that is the long-term proportion being examined. The original claim is a claim about the value of this variable. In the dice example, *p* stands for the long-term proportion of 7s for that pair of dice, and the claim can be written as

$$p = 1/6$$

We make our decision by comparing $\hat{p}$ (the proportion in the sample) to *p* (the proportion in the claim). The only question is this: how should we decide whether the proportion in the sample is or is not close to the proportion in the claim? As we have stated, and as you have explored using the applets, it is possible to adopt a decision strategy which controls the likelihood of making a Type 1 error. Based on our exploration, the decision strategy seems to depend on two things:

Sample size
Type 1 error rate that is desired

It turns out that the details of the decision strategy also depend on the particular proportion being investigated. We will explore this further in the subsection titled "Testing dice using other proportions."

The basis for decision-making (the sampling distribution)

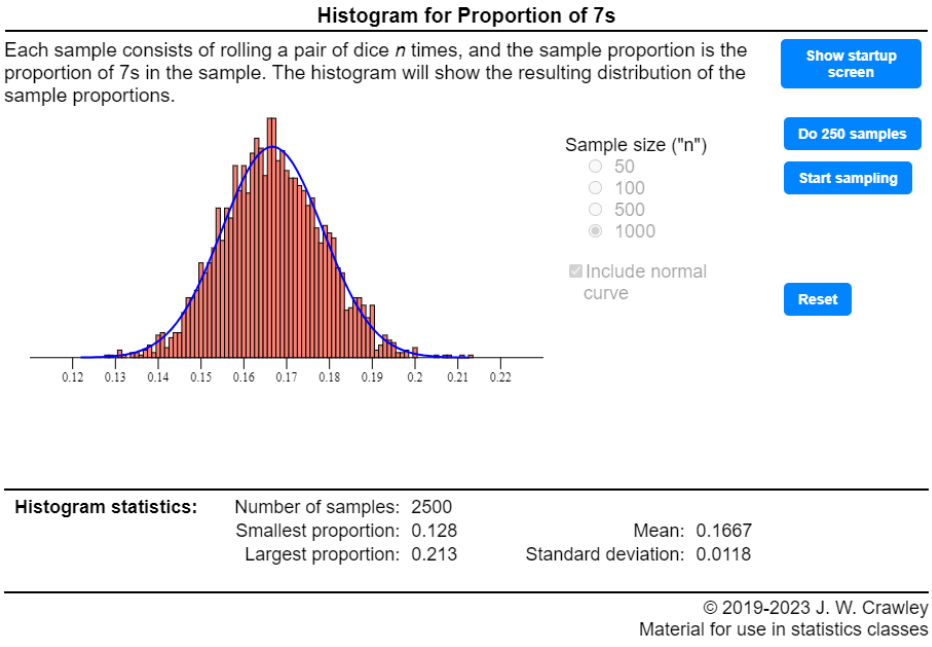
In Lesson 6, we learned that the logic of confidence intervals for proportions is based on the so-called *sampling distribution of the sample proportions*. It turns out that our calculations for hypothesis testing are based on that same concept. We will therefore quickly review a few key ideas from that lesson, and apply them to our current situation.

In order to control the occurrence of Type 1 error (discarding a fair pair of dice), it will be useful to consider what "should" happen when we roll a fair pair of dice n times and measure the sample proportion $\hat{p}$ for that sample. The answer (provided n is large enough) is that we expect the proportion of 7s to be close to 1/6 – probably not exactly 1/6, but pretty close. If we repeat the experiment over and over, sometimes it will be very close, sometimes not so close. In addition, it will be very close more often than it will be not so close.

The link below runs an applet that illustrates this. (We have used a similar applet earlier, in Lesson 6). When you click on the link, the applet will take 250 samples, each consisting of 1000 rolls of a fair pair of dice. It will calculate the sample proportion, $\hat{p}$, for each sample, and it will make a histogram of all these $\hat{p}$ values. Once you are in the applet, you can add more samples to the graph, start over with the same sample size, or change the sample size to 50, 100, or 500 while starting over.

Histogram of p-hat values

For example, here are the results obtained by the author using sample size 1000 and generating 2500 samples, using the option to overlay the histogram with a normal curve.



From the graph you can see that most of the values cluster around the probability / long-term proportion *p* (in this case 1/6, since the dice are fair and the probability of obtaining a 7 is 1/6). In addition, the histogram shape is similar to that of the normal curve. The following exercise will examine this phenomenon further.

Exercise 4. Use the applet[1] to experiment. In particular, for each of the sample sizes (50, 100, 500, 1000) use the start sampling / stop sampling buttons to generate about 100,000 samples with an overlaid normal curve. For each sample size, record your answers to these questions:

a. Does the histogram match the normal curve fairly closely?
b. What is the mean for the histogram?
c. What is the standard deviation?

Terminology: As we learned in Lesson 7, statisticians call the histograms you have viewed, or more precisely the theoretical histogram containing the $\hat{p}$ values for *all possible* samples of the chosen size, the **sampling distribution of the sample proportions.**

In general, we can make three statements about this sampling distribution:

- The mean is equal to the population proportion *p*.
- The standard deviation for a sampling distribution is referred to as the **standard error**. It can be calculated by the formula $\sqrt{\frac{p(1-p)}{n}}$.
- It is mound-shaped. In fact, provided *n* is large enough, it is approximately normal. (Note that this is consistent with your, and the author's, results in Exercise 4.)

Comment: The histogram the author obtained for 2500 samples closely matches the theoretical sampling distribution: 1) it is definitely mound-shaped; 2) its mean, 0.1668, is quite close to the population proportion $p = 1/6 = 0.1667$ when rounded to four places; 3) its standard deviation 0.012 is also quite close to $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(\frac{1}{6})(1-\frac{1}{6})}{1000}} = 0.0118$ when rounded to four places.

[1] Just as for the applet in the previous two lessons, we actually have two versions of the applet. In the first, the scale for the histograms is based on the range of values in the histogram, so the graphs for $n = 50$, 100, and so on, look quite similar. In the second, at the following link, all the graphs are drawn on the same scale. Using the second applet will emphasize how much closer the various sample proportions are to the mean when the sample size is larger. Here is the link:

Histogram of p-hat values

Exercise 5. Use your results from Exercise 4 to answer the following. Note that for each sample size, p is 1/6.

a. For $n = 50$, calculate the mean p and the standard error $\sqrt{\frac{p(1-p)}{n}}$ for the sampling distribution, rounded to 4 places. Compare the mean and standard deviation you obtained in Exercise 4 to the mean and standard error for the theoretical sampling distribution.

b. Do the same for $n = 100$.

c. Do the same for $n = 500$.

d. Do the same for $n = 1000$.

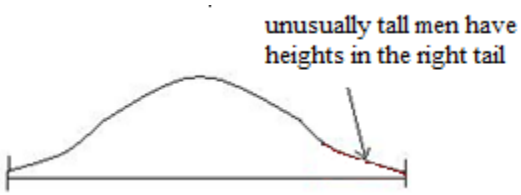
We can use the fact that the sampling distribution is approximately normal to build a viable strategy for deciding whether to discard or keep the dice. Put very simply, and in terms of the dice example, the decision strategy can be summarized as follows:

- For fair dice, very few pairs of dice have proportions of 7s which fall out in the tails of the sampling distribution.
- Therefore, if we discard dice whose proportion of 7s would fall out in the tails of the sampling distribution, we will not discard very many pairs of fair dice.
- That is, we will not have very much Type 1 error.

Thus, our decision strategy will be based on just how far out in the tails of the sampling distribution a particular sample proportion would lie. Since the sampling distribution is approximately normal, we can use the concept of the two-tail P-value to provide a numerical measure of how far out in the tail the sample proportion lies. This concept was first developed in Lesson 4, and the following section provides a very quick review of the topic.

8.3 – A Brief Review: P-Value Calculations

When you see an adult male walk into the room, you can instinctively judge his height, perhaps identifying him as "about average height" or as "unusually tall" or perhaps as "unusually short." If we think about the (approximately normal) distribution of adult male heights, the statement that a particular person is "unusually tall" can be rephrased as saying that his height is in the right tail of the distribution, as pictured here.



Similarly, an "unusually short" person would have a height lying in the left tail of the distribution.

We can quantify this rather vague notion of "unusual," using the fact that the distribution of adult male heights is approximately normal (with mean 70 and standard deviation 4). The concepts we develop can be applied to any situation involving a normal distribution, and indeed they can be generalized to

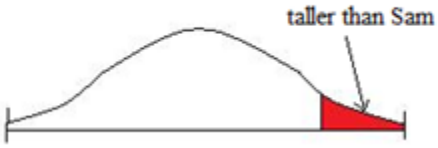
apply to many other distributions. These concepts are crucial to an understanding of the calculations and logic involved in inferential statistics.

A note on the use of technology. We will use technology to calculate probabilities/areas, with *z* scores calculated to four decimal places. (This was covered in section 4.4 of Lesson 4.) If you are using Table A, with *z* scores calculated to two decimal places (as covered in section 4.3), your answers will be slightly different.

One-tail P-values

Example. Sam is 79 inches tall. Notice that Sam is taller than the average height of 70 inches. How unusually tall is Sam?

Solution: We answer this question by calculating the probability that a randomly chosen adult male will be as tall as, or taller than, Sam, as illustrated in this figure:



To do this, we calculate a *z* score for Sam's height, $z = \frac{79-70}{4} = 2.2500$. Using the methods we learned about in Lesson 4, we calculate that the area to the right of this *z* score in the standard normal distribution is 0.0122. In the graph, the shaded area is 0.0122. The probability that a randomly selected adult male will be this tall or taller is 0.0122. Put another way, only 1.22% of adult males are as tall as, or taller than, Sam.

Terminology. We have calculated what we might call a "right tail probability" – the probability that a randomly chosen person's height is at least as far out in the right tail as Sam's height. Put another way, we have calculated the probability that a randomly chosen person's height is at least as far away from the mean as Sam's height, *in the "taller" direction.* In inferential statistics, a right tail probability such as is generally referred to as a **one-tail (right tail) P-value**.

Sometimes "P-value" is written as "*p*-value" or even more simply as just "*p*."

Example. Joe is 59 inches tall. Notice that Joe is shorter than the average height of 70 inches. How unusually short is Joe?

Solution: We answer this question by calculating the probability that a randomly chosen adult male will be as short as, or shorter than, Joe, as illustrated in this figure:

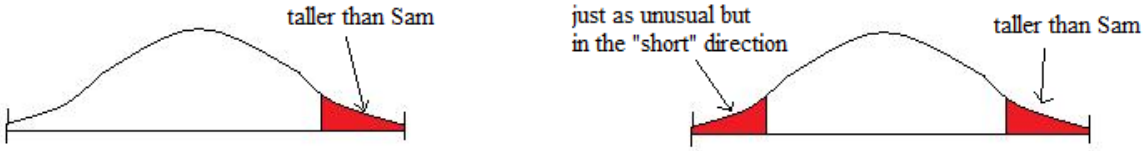
Again, we calculate a *z* score for Joe's height, $z = \frac{59-70}{4} = -2.7500$, then use the methods of Lesson 4 to calculate that the area to the left of this *z* score in the standard normal distribution is 0.0030. In the graph, the shaded area is 0.0030; this means that 0.30% of adult males are as

short as, or shorter than, Joe. The probability that a randomly selected adult male will be this short or shorter is 0.0030. The probability that a randomly selected adult male's will be this far away from the mean height, *in the "shorter" direction*, is 0.0030.

Terminology. A left tail probability such as this is generally referred to as a **one-tail (left tail) P-value**.

Two-tail P-values

The one-tail P-values we have calculated give a measure of how unusual a particular height is, in the sense of "unusually tall" or in the sense of "unusually short." A **two-tail P-value**, on the other hand, answers the more generic question of "how unusual is this height?" To answer this question for Sam, for example, consider the following diagram:



We have already seen the graph on the left, illustrating the 1.22% of adult men who are taller than Sam. Now examine the graph on the right, and recall that the normal distribution is symmetric. In addition to the 1.22% who are taller than Sam, there are an additional 1.22% whose height is just as unusual but in the "short" direction. So a total of 2.44% of all adult males have heights that are at least as unusual as Sam's height. Put another way, a total of 2.44% of all adult males have height as far away from, or further away from, the mean as Sam's height.

Terminology. A two tail probability such as this is generally referred to as a **two-tail P-value**.

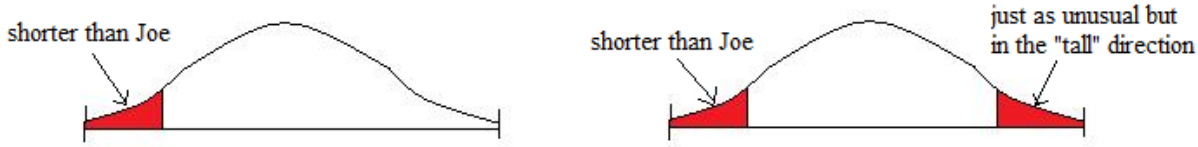
For a particular observation, the two-tail P-value measures the probability of being this far away from, or further away from, the mean (or average) of the distribution – *in either direction.*

Note: For data above the mean (such as Sam's height), we calculate the two-tail P-value by first calculating the one-tail (right tail) P-value. Because of the symmetry of the normal distribution, the two-tail P-value will be just double the value of the right tail P-value.

Similarly, for data below the mean, the two-tail P-value will be twice as large as the one-tail P-value, but in this case we begin by calculating the left tail P-value.

Problem. Joe is 59 inches tall. How unusual is this height; that is, calculate the two-tail P-value for this height.

Solution: We have already calculated the one-tail (left tail) P-value, obtaining 0.0030, as shown in the graph on the left below. This means that 0.30% of adult males are shorter than Joe.



Now examine the graph on the right. In addition to the 0.30% who are shorter than Joe, there are an additional 0.30% whose height is just as unusual but in the "tall" direction. So a total of 0.60% of all adult males have heights that are at least as unusual as Joe's height. The two-tail P-value is 0.0060, twice as large as the one-tail (left tail) P-value.

Exercise 6: Give the one-tail and two-tail P-values for these individuals:

a. Bill, 81 inches tall
b. Ted, 58.5 inches tall

Exercise 7: In this exercise, we imagine we have already done the first step in calculating a P-value, namely calculating the *z*-score. For each of the indicated *z*-scores, calculate the one-tail and two-tail P-values. (For positive *z*-scores, the one-tail P-value to calculate is the right tail; for negative *z*-scores, the left tail.)

a. $z = 2.03$
b. $z = 1.27$
c. $z = -2.65$
d. $z = -0.17$

Comment: Observe that the more unusual the data, the smaller the P-value. Put another way, the further the data is from the mean, the smaller the P-value. Visually, this is true because the P-value measures the area even further out in the tail or tails than the given piece of data. The more unusual the data (the further the data is from the mean), the smaller that area is. This is important enough to state it again:

Small P-values indicate data items far away from the mean.

The smaller the P-value, the further the data item is from the mean.

The methods of this section can be applied to any data that is approximately normally distributed, as illustrated by the following example.

Example. The verbal SAT scores at a particular college are approximately normal, with mean 493 and standard deviation 104. A student at that college scores 370 on the verbal SAT. Find the corresponding left-tail and two-tail P-values

Solution. The left-tail P-value is the probability of a random student's score being as low as, or lower than, this score. This is the same as the area to the left of this score in the normal distribution representing the scores at that college. As usual, we calculate a *z*-score and use that *z*-score to determine the corresponding area.

The *z*-score is calculated as $\frac{370-493}{104} = -1.1827$. Using technology, the area to the left of −1.1827 in the standard normal distribution is 0.1185. So the left-tail P-value is 0.1185 and the two-tail P-value is twice as large, 2(0.1185) = 0.2370.

The applet at the following link provides additional practice.

Calculating P-values

8.4 – Hypothesis Testing Calculations for Dice

The decision-making strategy using (two-tail) P-values

We are now ready to develop a formal strategy for deciding whether to discard or keep a pair of dice, by considering the proportion of 7s obtained in a sample of *n* rolls of the dice. By now, all the steps should be familiar to you, except of course the details of the decision-making step.

1. Write down the claim to be investigated. For this example, we claim that the probability / long-term proportion of 7s for the pair of dice is 1/6 (approximately 0.1667 or 16.67%). We will use 0.1667 in our calculations in the examples that follow. Symbolically, we can write this as

   $p = 0.1667$

2. We plan to take a sample of *n* rolls of the dice. Our decision process will be based on what we know should be true if that claim is true – namely, that the sampling distribution will be approximately normal, with

   Mean = probability / long-term proportion $p = 0.1667$

   Standard error = $\sqrt{\frac{p(1-p)}{n}}$, where p is 0.1667 and n is the size of the sample.

   Remember that "standard error" means standard deviation. As in the earlier lessons, we generally use the notation ***s.e.*** or just ***se*** to stand for the standard error.

3. Take the sample of *n* rolls of the dice, and calculate the sample proportion $\hat{p}$.

4. We will discard the dice if the sample proportion lies out in either tail of the sampling distribution. So we measure how far out in the tail this sample proportion lies by calculating its two-tail P-value. We do this in two steps:
   a. Calculate the *z* score for the sample, using the usual formula $z = \frac{x-\mu}{\sigma}$. In this formula:
      - x stands for the data value, which in this case is the sample proportion
      - μ stands for the mean of the distribution. In step 2 we observed that the mean of the distribution is 0.1667
      - σ stands for the standard deviation of the distribution, given by the standard error formula in step 2.
   b. Use the *z*-score to calculate the (two-tail) P-value.

5. Make the decision to discard or keep the dice by comparing the P-value to the desired Type 1 error rate. Remember that small P-values imply that the sample is far out in the tail of the sampling distribution, and therefore small P-values will cause us to discard the dice. Thus:
   a. If the P-value is less than the desired Type 1 error rate, discard the dice.
   b. Otherwise, keep the dice.

An example: revisiting the first applet

When we first introduced the loaded dice applet, we used what might be termed a "seat of the pants" approach to deciding whether or not to conclude the dice were loaded. We now have the tools to adopt a more systematic approach. Let us revisit the process in the first applet, at this link:

Keep or Discard Dice Part 1

We will follow the steps of a standard hypothesis test, using 0.05 as our criterion for deciding if our P-value is small enough to reject the claim. Using that criterion means that we are using a strategy that limits the Type 1 error probability to 0.05 (that is, to 5%).

Claim to be investigated. The probability / long-term proportion is 1/6 (approximately 0.1667 or 16.67%). We will use 0.1667 in our calculations. Written symbolically, the claim is

$$p = 0.1667$$

The sampling distribution. We will take a sample of size 1000 (that is, *n* is 1000). Therefore, the sampling distribution is approximately normal with mean and standard error (that is, standard deviation) given as follows:

mean = p = 0.1667

s.e (standard error) = $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.1667(1-0.1667)}{1000}} = 0.0118.$

Obtain sample, calculate z-score and P-value. We used the applet to obtain the sample, with these results:

Results:	Number of 7s rolled:	195		
	Out of:	1000	Expected percent for fair dice:	16.67%
	Percent of 7s rolled:	19.50%	Difference:	2.83%

The sample proportion, $\hat{p}$, is 19.50% or 0.1950 (calculated as 195/1000, that is 195 rolls of a 7 occurred in 1000 rolls of the dice).

The *z*-score for this particular sample is $z = \frac{x-\mu}{\sigma} = \frac{0.1950-0.1667}{0.0118} = 2.3983.$

Using technology, the area to the right of this *z*-score is 0.00823, giving a two-tail P-value of 0.0165. (If you use Table A, or do less rounding, you may obtain a slightly different answer.)

Conclusion. This P-value is smaller than 0.05, so we reject the claim, and discard the dice.

Was the conclusion correct? When a claim is rejected, there is always the possibility that we have committed a Type 1 error. The two-tail P-value calculated as 0.0165 tells us that 1.65% of samples from fair dice would have their sample proportion $\hat{p}$ this far, or further, out in the tail of the sampling distribution. So perhaps these dice *are* fair, and this just happened to be one of the very unusual samples that might arise from a set of fair dice.

In a real-world study, we never know for sure. However, in this dice applet the program itself knows what type of dice it was rolling, and the program informed the author that the dice were in fact loaded dice.

Comments:

1. Statisticians frequently use 0.05 as the criterion for making a decision, rejecting the original claim if the calculated P-value is less than 0.05. This strategy is used if the goal is to ensure that the probability of a Type 1 error (rejecting a valid claim) is below 0.05 or 5%.
2. Using 0.05 as the criterion is analogous to calculating a 95% confidence interval.
3. Another common strategy used by statisticians is to keep the probability of a Type 1 error below 1%, by using 0.01 instead of 0.05 as the criterion for the decision. (This is analogous to calculating a 99% confidence interval.)
4. In this example, if we had used 0.01 as our criterion, we would have kept the dice, since 0.0165 is *not* less than 0.01. As it turns out, this would have been a Type 2 error for this particular example.
5. Remember that in general reducing the likelihood of making a Type 1 error increases the possibility of having a Type 2 error.

Exercise 8: Using the applet, the author tested several additional pairs of dice. The claim being tested is the same as in the discussion: The probability / long-term proportion is 1/6 (approximately 0.1667 or 16.67%). Use 0.1667 in your calculations.

In each case, the sample size was $n = 1000$. The sample proportion of 7s, that is $\hat{p}$, is reported below. Calculate the corresponding *z*-score and (two-tail) *p*-value. Using 0.05 as your criterion for the decision, state your decision (discard or keep the dice).

a. $\hat{p} = 17.10\%$
b. $\hat{p} = 13.10\%$
c. $\hat{p} = 15.30\%$

Exercise 9: Using the applet, the author tested a pair of dice using a sample of size 500, obtaining $\hat{p} = 15.20\%$, and another pair of dice with $\hat{p} = 12.80\%$. For each pair of dice, calculate the *z*-score and (two-tail) *p*-value, and state your conclusion. Use 0.05 as the criterion for the decision.

Hint: You will need to recalculate the standard error, $s.e. = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.1667(1-0.1667)}{500}}$, because *n* is no longer 1000.

Exercise 10: **a.** For which, if any, of the tests carried out in Exercises 8 and 9 would you have reached a different decision if you had used 0.01 rather than 0.05 as your criterion?

b. True or false. Using 0.01 rather than 0.05 will reduce the likelihood of Type 1 errors (discarding fair dice).

The applet at the following link provides additional practice in the calculations for this type of hypothesis test. You will also practice forming a conclusion – keep or discard the dice – based on a specified criterion. For some of the problems, the specified criterion is 0.05 (that is, you wish to keep the Type 1

error rate below 5%). For others, you will used 0.01 as the criterion (to keep the Type 1 error rate below 1%).

Hypothesis tests for dice (calculations and conclusions)

This applet is identical except that it does the calculations for you, providing you with the opportunity to interpret the results of the calculations.

Hypothesis tests for dice (conclusions only)

Testing dice using other proportions

In these examples, we tested dice for fairness by examining the claim: The proportion of 7s for this pair of dice is 0.1667 (that is, 1/6). We could instead examine many other known probabilities for fair dice. For example, the probability of rolling a 2 is 1/36, the probability of rolling a 10 is 3/36, and so on. The important point to note is that for *any* probability (proportion) *p*, the sampling distribution has the same crucial three properties[2]:

- It is approximately normal
- The mean is equal to *p*
- The standard error (standard deviation) is given by the formula $se = \sqrt{\frac{p(1-p)}{n}}$

No matter the proportion being tested, the steps are the same. Compare the procedure in the following examples to the procedure in the preceding examples and exercises.

Example: For fair dice, the probability of rolling an even number total on the two dice (that is, either 2, 4, 6, 8, 10, or 12) is 50%, or 1/2.

Suppose you have rolled a pair of dice 800 times and have counted that the result was even 434 times. Should you discard or keep the dice? Answer the question using both 0.01 and 0.05 as the criterion for the decision. Show all the steps of the process.

Claim being investigated: The probability / long-term proportion of even rolls is 0.5 for this pair of dice. Symbolically,

$p = 0.5$

The sampling distribution, assuming the claim is true.

approximately normal
mean = 0.5

[2] By now you have experimented quite a bit with sampling distributions – in this lesson for counting 7s when dice are rolled, and in Lessons 6 and 7 for sampling from a population. If you wish to do additional empirical experimentation, these applets create histograms of the sampling distribution for the situation of rolling an even total (i.e., 2, 4, 6, 8, 10, or 12) with a pair of dice.

Histogram of p-hat values (counting even rolls)
Histogram of p-hat values (counting even rolls), version 2

s.e. (standard error) = $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.5(1-0.5)}{800}} = 0.0177$

Obtain sample, calculate z-score and P-value.

the sample proportion is $\hat{p} = \frac{434}{800} = 0.5425$

$z = \frac{0.5425-0.5}{0.0177} = 2.4011$

using technology, the two-tail P-value is 0.0163

Decision. Using 0.05 as the criterion, the P-value is less than 0.05, so we reject the claim and discard the dice.

Using 0.01 as the criterion, the P-value is *not* less than 0.01, so we acknowledge that the claim could be true and keep the dice.

Reminders:

- The criterion you use is the probability of having a Type 1 error (discarding fair dice).
- If the P-value is less than the criterion, reject the claim and discard the dice.
- Otherwise, acknowledge the claim could be true and keep the dice.

Exercise 11: Do an analysis similar to this example, for each of the following situations.

a. There were 723 even rolls in a sample of 1500 rolls.
b. There were 653 even rolls in a sample of 1200 rolls.
c. There were 585 even rolls in a sample of 1097 rolls.

Example. For fair dice, the probability of rolling a 10 is 3/36 ≈ 0.0833. You roll a pair of dice 1043 times, obtaining a 10 on 108 of those rolls. Test the dice for fairness using 0.01 as the criterion. (That is, you want to keep the probability of a Type 1 error below 1%.) Show all the steps of the process.

Claim being investigated: The probability / long-term proportion of 10s is 0.0833 for this pair of dice. Symbolically,

$p = 0.0833$

The sampling distribution, assuming the claim is true.

approximately normal
mean = 0.0833

s.e. (standard error) = $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.0833(1-0.0833)}{1043}} = 0.0086$

Obtain sample, calculate z-score and P-value.

the sample proportion is $\hat{p} = \frac{108}{1043} = 0.1035$

$z = \frac{0.1035-0.0833}{0.0086} = 2.3488$

using technology, the two-tail P-value is 0.0188

Decision. We are using 0.01 as the criterion. Since the P-value is *not* less than 0.01, we acknowledge that the claim could be true and keep the dice.

The applet at the following link provides additional practice:

Hypothesis tests for dice, general situation (calculations and conclusions)

This applet is identical except that it does the calculations for you, providing you with the opportunity to interpret the results of the calculations.

Hypothesis tests for dice, general situation (conclusions only)

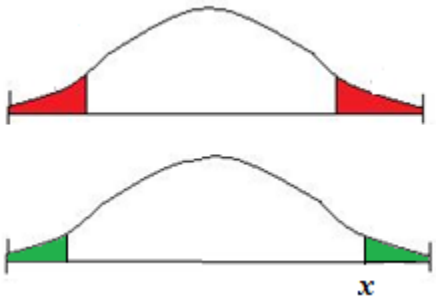
Why does this decision strategy work?

The decision strategy we have outlined is simple enough. If your desired probability of making a Type 1 error is 5% (that is 0.05), for example, you simply compare the P-value to 0.05. If the P-value is less than 0.05, you discard the dice, otherwise you keep the dice. Since the proportion in the claim is the mean of the distribution, and since small P-values belong to samples whose proportion is far from the mean, it intuitively makes sense to reject the claim if the sample proportion has a small P-value. But why does this result in a Type 1 error rate of 5%?

To answer this question, we recall that a Type 1 error is rejecting a true claim, that is discarding fair dice. So the probability of having a Type 1 error is based on what happens when you sample with fair dice – that is, the sampling distribution. That sampling distribution is approximately normal, as illustrated in the first part of the diagram below. In that diagram, we have shaded in red the 5% of the data that is furthest away from the mean. In order to have a 5% Type 1 error rate, we can adopt this strategy:

Discard the dice if the sample proportion falls in the red shaded area of the sampling distribution.

Since 5% of the results for fair dice will fall in the red shaded area, we will discard 5% of the fair dice – that is, we will have a 5% Type 1 error rate.



Red shading indicates the 5% of data furthest from the mean.

Green shading indicates the P-value for the data item labeled x.

Now, what characteristic do the samples that fall in the red-shaded area share? They are precisely those samples whose P-value is less than 0.05. For example, in the lower part of the diagram we consider a sample which falls at the value labeled *x*. The P-value for that sample will be the green shaded area, which is obviously less than the 5% red shaded area shown in the top part of the diagram. Thus, we can rephrase our strategy as:

Discard the dice if the sample proportion has a P-value less than 0.05.

A similar discussion applies when we want to limit the Type 1 error rate to 1% - we discard those dice whose sample proportion has a P-value less than 0.01.

8.5 – Hypothesis Test Calculations for Samples from a Population

Lessons 6 and 7 developed strategies for calculating confidence intervals. In those lessons, we saw that ideas which were developed in the context of dice rolls carried over quite nicely to the more general context of sampling from a population. The same is true for hypothesis testing strategies. In this section, we will see how to extend what we have learned about hypothesis tests for dice to more general situations.

We will continue to refer to the dice examples, and in addition, we will revisit this example first described in Lesson 7:

> A recent report by the Centers for Disease Control states that 16.8% of American adults are smokers. The author was recently taking care of family business in the town where his in-laws live, and felt that he was seeing more people smoking than he usually did in his own home town. This raised the question: Is the proportion of smokers in his in-laws' town 16.8%, or is it higher (or possibly lower)?

Just as we discussed in Lesson 7, to answer this question statisticians would conduct a poll, that is: choose a random sample, measure the proportion of smokers in that sample, and use the result to make a judgement about the proportion in the entire town.

Connecting population proportions to probabilities

As was the case for confidence intervals, the key to our discussion is the connection between proportion and probability. For example, consider the question about the proportion of smokers in a certain town. The question the statisticians would ask the people in the sample is this: "Are you a smoker?" For that question, the *population* the pollsters are interested in consists of all adults in that town. The pollsters want to investigate the proportion of that population that would answer "yes" to that question. The basis for what they do is the fact that these two numbers are the same:

- The proportion of adults in that town who would answer "yes."
- The probability that a randomly selected adult from that town would answer "yes."

As a result, the claim the pollsters are investigating can be thought of in two equivalent ways:

- The proportion of smokers in that town is 16.8% (that is, 0.168).
- The probability that a randomly selected adult from that town is a smoker is 0.168.

The pollsters are going to investigate a claim about a probability, just as we investigated claims about probabilities for dice rolls in Lesson 6. The methods we developed for dice probabilities in Lesson 6 carry over quite nicely to investigating general claims about population proportions. Here is a table summarizing the connections between the first dice example and the question about smoking proportion.

	Dice example	Smoking example
Claim being investigated	This pair of dice is fair (the probability of rolling 7 is 1/6).	The proportion of smokers in the town is 16.68%.
Conclusions possible	Discard: The dice are loaded. We found evidence the probability *is not* 1/6. Keep: The dice *might be* fair (the evidence isn't enough to conclude that they are loaded).	Reject the claim: We found evidence the proportion *is not* 16.68%. Do not reject the claim: The proportion *might be* 16.68% (the evidence isn't enough to conclude the claim was incorrect).
Type 1 error	Discard fair dice.	Reject a true claim.
Type 2 error	Keep loaded dice.	"Keep" a false claim (that is, fail to reject that false claim).

Important note. Recall from Lesson 6 that when we keep the dice, we aren't claiming that they *are* fair; we are only acknowledging that there wasn't enough evidence to cause us to discard them – they *might be* fair. Similarly, in the general hypothesis testing methodology of this lesson, we have these possible conclusions: we either reject the original claim, or we acknowledge that the original claim *might be* true. We never assert that the original claim *is* true.

Another important note: For our process of rolling the dice to have any meaning, the rolling of the dice had to be random. Similarly, for the process of selecting individuals for the sample to have any meaning, the selection must be random. Provided the pollsters use a simple random sample for the study, the mathematical results are quite predictable and reliable, just as we saw for the dice rolling examples.

The logic of hypothesis testing (two-tail tests)

We are now ready to develop the formal strategy used to carry out a hypothesis test involving a population proportion. That strategy has already been introduced informally earlier in this lesson, in the context of testing dice for fairness. In this section, we concentrate on the overall logic of the strategy, using the dice examples and the smoking example described above as illustration. In the next section we apply the strategy to the smoking example mentioned above. In Section 8.6 we fill in the details, including terminology, notation, and assumptions.

Claim to be investigated

The purpose of the hypothesis test is to investigate the claim that the population proportion is equal to some particular value. We will reject this claim if we obtain evidence that the population proportion is not equal to this value – evidence either that it is smaller than claimed or that it is larger than claimed.

For example, we first focused on the proportion of 7s, which for fair dice should be 1/6 (which is approximately 16.67%, or 0.1667). The manufacturer's claim that the dice are fair translated to this claim to be investigated:

$$p = 0.1667$$

In later examples, we also considered other proportions, including the proportion of even rolls, which should be 1/2 for fair dice. The corresponding claim to be investigated was written as

$$p = 0.5$$

In general, the claim is written in a form similar to these two examples: population proportion is equal to some particular value. This particular value is commonly written as p_0, so our claim takes the form

$$p = p_0$$

where p_0 is some particular proportion.

The sampling distribution

The decision process is based on considering what should happen if the claim is true, the so-called "sampling distribution" for sample proportions. If we take many, many simple random samples of the same size, calculate the sample proportion for each sample, and create a histogram of these sample proportions, the resulting distribution is approximately normal. Moreover, the mean for this distribution is the population proportion in the claim (designated p_0), and the standard deviation (standard error) is given by the formula $\sqrt{\frac{p_0(1-p_0)}{n}}$. We use these formulas to calculate the mean and standard error for the sampling distribution.

Comment: We have encountered these formulas many times in the past few lessons, but without the subscripts on the variable *p*. In this context of hypothesis testing, we will use p_0 to emphasize that the calculations are based on the particular number in the claim being investigated.

For example, if we take a sample of size 1000 to investigate the claim that the proportion of even rolls is 0.5, the sampling distribution has these properties:

- Approximately normal
- Mean is the population proportion in the claim, so the mean is 1/2 or 0.5.
- The standard error is calculated as $se = \sqrt{\frac{0.5(1-0.5)}{1000}}$

Calculations based on a sample

To investigate the claim, we obtain a simple random sample from the given population, and we calculate the proportion for that sample, written as $\hat{p}$ (read "p-hat").

The sampling distribution tells us what we should expect to see if the claim is true, so we can use the sampling distribution to find out if what we see in the sample supports or contradicts the claim. Samples that fall reasonably close to the proportion in the claim will support that claim. Samples which fall out in the tails of the sampling distribution will contradict the claim. We therefore need some measure of how far out in the tails a particular sample proportion lies.

Because the sampling distribution is approximately normal, we can quantify this by first using the mean and standard error for the sampling distribution to calculate a *z*-score for our sample, then using that *z*-score to calculate a *P*-value. The *z*-score calculation uses the usual formula for a *z*-score, which we can express in words as

$$z = \frac{data\ value - mean}{standard\ deviation}$$

or in symbols as

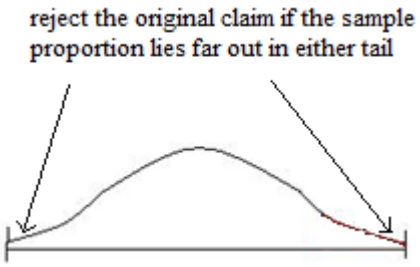
$$z = \frac{\hat{p} - p_o}{se}$$

We then use either Table A or technology to calculate a tail probability (a *P*-value) for this *z*-score. We use a two-tail *P*-value because both unusually small and unusually large sample proportions provide evidence that the claim is incorrect.

The decision: Is there enough evidence to reject the original claim?

In words, the strategy for making a decision is very simple. We will reject the claim if the sample yields a proportion which is nowhere close to the claimed proportion. On the other hand, if the sample's proportion is reasonably close to the claimed proportion, we will acknowledge that that claim might be correct.

In terms of the sampling distribution, this means that we should reject the original claim when we obtain a sample proportion which lies far out in one of the tails of the sampling distribution.



As we have seen, the *P*-value provides a measure of how far out in the tail a particular piece of data lies – the further out in the tail, the smaller the *P*-value. This leads to this summary of the decision-making strategy:

***P*-value**	**Conclusion**	**For dice examples**
Small	Reject claim	Discard dice
Otherwise	Do not reject claim	Keep dice

But, how small is small? As for the dice examples, our criterion for answering this question is tied to our desire to limit the occurrence of Type 1 errors. Using "less than 0.05" as our criterion for "small" will result in a 5% occurrence of Type 1 errors; similarly, using "less than 0.01" results in a 1% occurrence of Type 1 errors. Although other criteria are certainly possible, these two are the most frequently used by statisticians.

Notes:

1. Which criterion (boundary value) to use – whether 0.05, or 0.01, or perhaps some other choice – must be decided *before* carrying out the study on the sample. It is unethical for a statistician to wait until after calculating the *P*-value to decide on the criterion for smallness.
2. For purposes of this course, you should use "less than 0.05" for your decision if we do not specify some other boundary value.
3. The criterion we use is related to the desire to limit Type 1 errors.

Summary:

- If the calculated *P*-value is less than the chosen boundary value, reject the claim.
- Otherwise, do not reject the claim (acknowledge that the claim *could be* correct). Notice that this conclusion does *not* state that the claim *is* correct, only that it *could be* correct.

An example: smoking in a small town

To illustrate how similar the methods are to those we have already seen, we will carry out a hypothesis test for the example mentioned earlier. You should compare this example to the examples in the previous section. Those examples started with the claim that the proportion for this particular pair of dice matches what it should be for fair dice. In this example, we begin with the claim that the proportion of smokers in this particular town matches that reported by the CDC.

Example. A recent report by the Centers for Disease Control states that 16.8% of American adults are smokers. Suppose that researchers take a random sample of adults from a particular town, and find that 215 out of 1100 adults sampled report being smokers[3]. They will use 0.05 as the criterion / boundary value (that is, they want to keep the probability of a Type 1 error below 5%). What conclusion would they draw?

The Form of the Claim to be Investigated

No matter how the situation might be described in the statement of the problem, for a hypothesis test about population proportions the claim to be investigated will always be written as a claim that the population proportion *is equal to* some particular value, that is in the form

$$p = p_0$$

where p_0 is some particular proportion. In this example, the study was triggered by the author's suspicion that the proportion of smokers in that town *is not equal to* 16.8%. Nevertheless, the claim to be investigated is that the proportion *is* 16.8%.

In general, we can describe the claim as some way of saying "no difference" or "no change" or "just as it should be." In this case, the claim states that there is no difference between the smoking rate for this particular town and the rate reported by the CDC for the entire country. Frequently, as in this example, studies are motivated by a researcher's opinion that things are *not* the same, or *not* as they should be. In the dice examples, we started out with a worry/suspicion that some casinos might be using loaded dice, but the claim we investigated was "these dice are fair." The hypothesis test methodology always begins by investigating the claim that the population proportion *is* equal to some specific value. This is true no matter what the researcher's suspicion might be.

[3] As we noted in Lesson 7, the examples in these notes generally fall into one of several categories. Some may be taken completely from actual studies reported in the literature. Others are adaptations of actual studies, with the data modified to allow for practice with the methods being studied. Still others are situations which are similar to studies that statisticians do but which are made up by the author. This example falls into the latter category (although the 16.8% figure reported by the CDC is real).

Solution:

Claim being investigated: The study was triggered by a suspicion that the proportion of smokers in that particular town might not match the 16.8% reported by the CDC for the entire country. We will investigate the claim written symbolically as:

$p = 0.168$ (that is, $p = 16.8\%$)

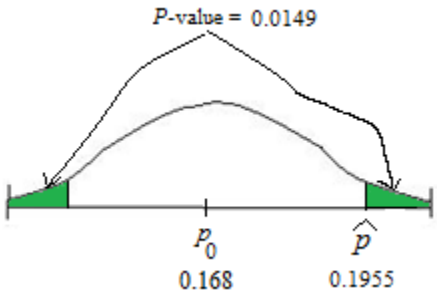
The sampling distribution, assuming the claim is true.

approximately normal
mean: $p = p_0$, that is $p = 0.168$
standard error: $se = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.168(1-0.168)}{1100}} = 0.0113$

Obtain sample, calculate z-score and P-value.

the sample proportion is $\hat{p} = \frac{215}{1100} = 0.1955$
$z = \frac{proportion\ in\ sample - mean}{standard\ error} = \frac{0.1955-0.168}{0.0113} = 2.4336$
using technology, the two-tail P-value is 0.0149

This figure illustrates the situation:



Decision. We are using 0.05 as the criterion. Since the P-value is less than 0.05, we reject the claim. We conclude that the proportion of smokers in this town is not 16.8%. (Based on the sample, we believe it is larger than 16.8%.)

Comment: One way to think of the *P*-value is this: It is the probability, *if the claim is correct*, of obtaining a sample whose sample proportion is at least this far away from the proportion in the claim. That probability is less than 5% (0.05), so we reject the claim.

Comment: We might ask, *Was the conclusion correct*? When a claim is rejected, there is always the possibility that we have committed a Type 1 error. Perhaps the proportion in the town actually *is* 16.8%, and this just happened to be one of the very unusual samples that might arise from a population proportion of 16.8% (Only 1.49% of all such samples would be as unusual as, or more unusual than, this sample.)

In a real-world study, we never know for sure. We can say, however, that we have used a procedure which limits the occurrence of Type 1 error (rejecting a true claim) to 5%.

Exercise 12: In the previous example, suppose the researcher had decided, when planning the study, to use 0.01 as the criterion for the decision. Describe the decision that would be made in that case.

Exercise 13: In this exercise, you will carry out the same hypothesis test but using several different hypothetical sample results. For each, show the calculations, and answer the question "Do you reject the claim" using criterion 0.05 and also criterion 0.01.

a. 99 out of 700 are smokers
b. 266 out of 1365 are smokers
c. 68 out of 520 are smokers

	Standard error	$\hat{p}$	z	***P*-value**	**Do you reject claim?**	
					Using 0.05	**Using 0.01**
a.						
b.						
c.						

Here is another example. Except that the context does not involve tossing dice or checking people's smoking habits, the problem is essentially identical to those you have already seen. The steps are the same, using the data supplied in the description of the problem.

Example. Should we believe that a certain population proportion is 0.35? Based on a sample in which 355 people answer yes to the question posed, do an appropriate two-tail test. (The sample consisted of 927 people.) Use 0.05 as the criterion. (That is, you want to keep the probability of a Type 1 error below 5%.) Show all the steps of the process.

Claim being investigated: The population proportion is 0.35. Symbolically,
$p = 0.35$

The sampling distribution, assuming the claim is true.
Approximately normal
mean = p_0 (the proportion in the claim) = 0.35

se (standard error) = $\sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.35(1-0.35)}{927}} = 0.0157$

Obtain sample, calculate z-score and P-value.
the sample proportion is $\hat{p} = \frac{355}{927} = 0.3830$
$z = \frac{0.3830-0.35}{0.0157} = 2.1019$
using technology, the two-tail P-value is 0.0356

Decision. We are using 0.05 as the criterion. Since the P-value is less than 0.05, we reject the claim. (Note that a criterion of 0.01 would lead to a different decision in this example.)

The applet at the following link provides additional practice in the calculations for this type of hypothesis test. You will also practice forming a conclusion – reject the claim, or do not reject the claim – based on a specified criterion. For some of the problems, the specified criterion is 0.05 (that is, you wish to keep the Type 1 error rate below 5%). For others, you will use 0.01 as the criterion (to keep the Type 1 error rate below 1%).

Hypothesis tests for population proportions

8.6 – Terminology, Notation, and Assumptions

Assumptions

In this section we will do a quick review of the logic of a two-tail hypothesis test, along with an introduction to additional terminology and notation that is used. Before we begin, however, here are a few "assumptions," that is, conditions which must hold if we wish to use these methods.

- It goes without saying, perhaps, that we are studying a categorical variable – otherwise, it makes no sense to even be talking about proportions.
- The population must be much larger than the sample (typical guidelines vary from specifying at least 10 times as large to at least 20 times as large). For this course, we will generally just assume this to be true – after all, if the entire population is only a little larger than our sample, it would be possible to just ask the question to the entire population.
- The theory is based on taking a simple random sample from the population, or performing a random experiment on a sample taken from the population.
- We will be using the normal distribution in our calculations. For this to be valid, the sample size must be large enough that both these conditions hold, where p_0 is the particular proportion in the claim you are investigating:
  - $np_0 \geq 15$
  - $n(1-p_0) \geq 15$

Example: For the example of Section 8.5, the sample size was *n* = 1100, and the proportion in the claim we investigated was p_0 = 0.168.

- The categorical variable is the answer to the question, "Are you a smoker?"
- The sample was size 1100, and we are assuming the town has at least 10 to 20 times the sample size. This would be true for a town of size 25,000 or larger, for example.
- The problem description described using a random sample.
- Since *n* is 1100, and p_0 is 0.168, we have these calculations:
  - $1100(0.168) = 184.8$
  - $1100(1-0.168) = 915.2$

  Both numbers are certainly ≥ 15

Hypothesis test logic with terminology and notation (two-tail tests)

At this point we have completed our description of the logic and calculations used for testing hypotheses about population proportions. In the current subsection we will enhance that coverage by introducing some of the terminology and notation that statisticians use when carrying out a hypothesis test

about a population proportion. The calculations and logic are the same – the only thing new is the terminology and notation.

Claim to be investigated

The purpose of the hypothesis test is to investigate the claim that the population proportion is equal to some particular value commonly written as p_0. This claim is referred to as the ***null hypothesis***, and the symbol H_0 is used to represent the claim. The null hypothesis is written symbolically in this form:

$H_0: p = p_0$

Along with the null hypothesis, we also write what is called an ***alternative hypothesis***, written as H_a. The alternative hypothesis describes the conclusion we will reach if we reject the null hypothesis. In general, for the two-tail tests we are studying, the alternative hypothesis is of this form:

$H_a: p \neq p_0$.

We will reject the null hypothesis (and conclude that the alternative hypothesis is true) if we obtain evidence that the population proportion is not equal to p_0 – evidence either that it is smaller than claimed or that it is larger than claimed.

The purpose of a hypothesis test is to examine whether we believe a claim that a proportion is equal to a particular value. This particular value is written p_0 and the null and alternative hypotheses are written:

$H_0: p = p_0$
$H_a: p \neq p_0$

This particular value p_0 can have a variety of sources.

- In the dice example, it is the theoretical probability for fair dice.
- In the smoking example, it is the nationwide proportion reported by the CDC. We were testing whether that particular town's proportion is the same as the nationwide proportion.
- Another common situation involves testing whether the proportion is still the same as it was at some time in the past.

Note that how a problem is stated does not change the nature of the null hypothesis – it always makes the claim that the proportion *is* equal to some stated value. For the dice example, one person might say the Gaming Commission is investigating whether the dice are fair; another might describe it as testing whether the dice are loaded. In either case, the null hypothesis is that $p = 0.5$.

Examples: For the dice example based on the proportion of even rolls, the null and alternative hypotheses would be

$H_0: p = 0.5$
$H_a: p \neq 0.5$

For the smokers example, we have

$H_0: p = 0.168$
$H_a: p \neq 0.168$

The sampling distribution

The logic of hypothesis testing always relies on the sampling distribution. If the null hypothesis *is* true, what should we expect when we take a sample? If we know what to expect, we can decide whether our sample falls into the "this is what we expected to happen" category or into the "wow, that doesn't seem consistent with the null hypothesis" category.

For proportions, the sampling distribution is approximately $N\left(p_0, \sqrt{\frac{p_0(1-p_0)}{n}}\right)$. That is, it has these three properties:

- It is approximately normal
- The mean is equal to the claimed proportion p_0
- The standard deviation (standard error) is given by the formula $se = \sqrt{\frac{p_0(1-p_0)}{n}}$.

If we obtain a sample whose proportion lies far out in either tail of this normal distribution, we will reject the null hypothesis.

Comment. The confidence interval calculations from Lessons 6 and 7 also used the formula for standard error. There is one major difference, however:

- For a hypothesis test, we use p_0, the claimed proportion for the entire population.
- When calculating a confidence interval, we have no value – claimed or otherwise – for the entire population. We are forced to do the best we can, using the sample proportion $\hat{p}$ as the proportion in the formula.

Calculations based on a sample

After we take our sample, there are three steps, although of course they can be combined.

1. Calculate the sample proportion $\hat{p}$.
2. Calculate the *z*-score. The *z*-score calculation uses the usual formula for a *z*-score, which we can express as

$$z = \frac{data\ value - mean}{standard\ deviation} = \frac{x-\mu}{\sigma}$$

In this setting, the "data value" is the sample proportion, the "mean" is the mean of the sampling distribution, and the "standard deviation" is the standard error of the sampling distribution, so we can also write the formula as

$$z = \frac{\hat{p}-p_o}{se}$$

In words, the formula can be read as, "what the sample says, minus what the null hypothesis claimed, divided by the standard error." (This *z*-score is frequently referred to as a ***test statistic***. It is a statistic calculated from the sample, and which is used to test the null

hypothesis. In what follows we will frequently refer to this calculation as calculating the test statistic.)

3. Use the test statistic (that is, the *z*-score) to measure how far out in the tails of the sampling distribution this sample lies. This is given by the two-tail *P*-value.

The decision: Is there enough evidence to reject the original claim?

Samples that lie far out in the tails of the sampling distribution have small *P*-values. So if the *P*-value is small, we reject the null hypothesis. How small is small enough? The criterion we use is based on our desire to limit the occurrence of Type 1 error. The most common criteria used are 0.05 and 0.01.

This criterion is referred to as the ***significance level***, and it is symbolized using the Greek letter alpha, written α. So a common value is $\alpha = 0.05$, and another common value is $\alpha = 0.01$. The significance level indicates the probability of making a Type 1 error.

- If *P*-value is less than α, **reject** the null hypothesis, and conclude that the alternative hypothesis is true.

- Otherwise, **fail to reject** the null hypothesis – it might be true, there is not enough evidence to conclude that the alternative hypothesis is true.

Example: A researcher investigating a claim about a population proportion samples 983 people from the population, and 380 of those people answer yes to the question she poses. Evaluate the claim that the population proportion is 37.2%, using a two tail hypothesis test. At a 0.01 significance level, what does the researcher conclude?

Solution: The first step is to write the null hypothesis and the alternative hypothesis:

$$H_0: p = 0.372$$
$$H_a: p \neq 0.372$$

The decision will be based on the sampling distribution, which is approximately normal with:

Mean = $p_0 = 0.372$

Standard error = $\sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{.372(1-.372)}{983}} = 0.0154$

We calculate the sample proportion – that is, the proportion obtained by asking the question to the sample:

$$\hat{p} = \frac{380}{983} = 0.3866$$

Notice that the sample proportion is definitely NOT exactly equal to the 37.2% (0.372) in the null hypothesis. However, our job is to decide whether this difference is due to just the randomness of the sampling process, or whether it indicates that the null hypothesis should be rejected. To make this decision, we calculate two additional values: the *z* score (test statistic) for this particular sample proportion, and the two-tail *P*-value for that *z* score. A small *P*-value will indicate that this sample lies out in the tail of the sampling distribution,

and we will reject the null hypothesis. If the *P*-value is not small, we will not reject the null hypothesis. Here are the calculations (using technology to determine the *P*-value):

$z = \frac{\hat{p} - p_0}{se} = \frac{0.3866 - 0.372}{0.0154} = 0.9481$
P-value = 0.3431

Because our *P*-value (0.3431) is **not** less than the significance level (0.01), we **do not** reject the null hypothesis. We conclude that the population proportion **could be** 37.2%.

Comment: We used the normal distribution in our calculations. For this to be valid, the sample size must be large enough that both these conditions hold:

- $np_0 \geq 15$
- $n(1 - p_0) \geq 15$

Here are the calculations that justify this use of the normal distribution:

- $np_0 = 983(0.372) = 365.676$
- $n(1 - p_0) = 983(1 - 0.372) = 617.324$

Both these numbers are definitely ≥ 15, so our use of the normal distribution is justified.

Exercise 14: Should we believe that a certain population proportion is 0.695? Based on a sample in which 755 people answer yes to the question posed, carry out an appropriate two-tail test. (The sample consisted of 1149 people.)

Note: The problem as stated does not specify the significance level you should use. As noted above, when this happens you should use $\alpha = 0.05$.

The applet at the following link provides additional practice similar to Exercise 4.

Hypothesis tests (calculations and conclusions)

The next applet is similar to the previous applet, but the calculations are done for you – it is up to you to write the null and alternative hypotheses, and to identify the correct conclusion.

Hypothesis tests (conclusions only)

8.7 – Stating the Conclusion, in the Context of the Problem

Our conclusion in general takes one of two forms: (1) We *reject* the null hypothesis; or (2) We *fail to reject* the null hypothesis. One caution we must always keep in mind is that *fail to reject* does not mean the same thing as *accept*. When we fail to reject the null hypothesis, we acknowledge that it *might be* true, which is different from saying that it *is* true. The conclusion we write frequently includes the significance level. For example, we might write sentences such as these:

- At significance level $\alpha = 0.05$, we reject the null hypothesis; or
- We are unable to reject the null hypothesis at significance level 0.01.

This way of writing the conclusion is generic; it applies to any problem. However, we also want to learn how to express the conclusion in terms of the original problem – that is, in terms of dice, or in terms of smokers, etc. The easiest way to do this is to write the conclusion in terms of the alternative hypothesis, as in these sample templates. First, if we reject the null hypothesis:

- At significance level $\alpha = 0.05$ we are able to conclude that ___verbal description of what the alternative hypothesis states_____.
- There was enough evidence, at significance level 0.05, to show that ___verbal description of what the alternative hypothesis states_____.

When we fail to reject the null hypothesis, we simply negate the statement we would make if we rejected the null hypothesis, as in these templates.

- At significance level $\alpha = 0.05$ we are **un**able to conclude that ___verbal description of what the alternative hypothesis states_____.
- There was **not** enough evidence, at significance level 0.05, to show that ___verbal description of what the alternative hypothesis states_____.

Example. In our smoking example of this section and the previous section, the researcher is studying the smoking rate in a certain small town, with the following null and alternative hypotheses:

$H_0: p = 0.168$

$H_a: p \neq 0.168$

In the sample, 215 out of 1100 adults were smokers, giving a sample proportion of 0.1955, a z-score is of 2.4336, and P-value of 0.0149. Write the conclusion, in the context of the study, for these situations:

a. Using significance level 0.05.

Solution: Since 0.0149 is less than 0.05, we reject the null hypothesis, and write:

At significance level $\alpha = 0.05$ we are able to conclude that the proportion of smokers in that town is not 16.8%.

In fact, since the sample proportion was 19.55%, we can go one step further and conclude that the proportion of smokers in that town is *greater than* 16.8%.

b. Using significance level 0.01.

Solution: Since 0.0149 is **not** less than 0.05, we **do not** reject the null hypothesis, and write:

There was not enough evidence, at significance level 0.01, to show that the proportion of smokers in that town is not 16.8%.

That is, the proportion **might be** 16.8%, we cannot conclude that it isn't 16.8%.

Exercise 15: Researchers compare the success rate for a new drug to the known success rate (73%) for an existing drug, using the following null and alternative hypotheses:

$H_0: p = 0.73$

$H_a: p \neq 0.73$

a. If they reject the null hypothesis, how will they write the conclusion in the context of the study?

b. How will they write the conclusion if they do not reject the null hypothesis?

Comment: There is another way statisticians like to write the conclusion which uses the word *significant* or a similar word. For example, if they reject the null hypothesis in the example involving smokers in a small town, they might write, "The smoking rate in that town is significantly different from the CDC's reported 16.8% for the entire nation." If they did not reject the null hypothesis, they would just insert the word *not* as shown here: "The smoking rate in that town is **not** significantly different from the CDC's reported 16.8% for the entire nation."

In this manner of writing the conclusion, it is important to note that the word "significant" simply implies that the difference between the sample and the claim was large enough to reject the claim for the entire population. That is, *significant* implies the null hypothesis was rejected, and *not significant* implies that the null hypothesis was not rejected.

Exercise 16: For the situation described in exercise 15, use the word "significant" to write the conclusion if:

a. they reject the null hypothesis

b. they do not reject the null hypothesis

The applet at the following link provides additional practice formulating conclusions in terms of the context of the particular study, as illustrated by Exercises 15 and 16.

Formulating conclusions and other interpretation practice

8.8 – More Practice

If you have done all the exercises and used the suggested applets, you have extensive practice with the individual components of hypothesis testing for proportions. In this section, we present complete examples incorporating all the steps of the hypothesis test.

Example: Has the proportion of smokers among adults in the Unites States changed since the 1960s, when it was reported to be 44%? Researchers randomly sample 2075 such adults, and find that 857 are smokers. Carry out a hypothesis test based on this data, using significance level 0.05.

Solution: The first step is to write the null and alternative hypotheses. The null hypothesis will be of the form $p = p_0$, and will represent the idea "no change" or "no difference." That is, it will represent the claim that nothing has changed since the 1960s. The alternative hypothesis will state that the null hypothesis is false.

$H_0: p = 0.44$
$H_a: p \neq 0.44$

Note that these are claims about the population: *all adults in the United States.* In words, the null hypothesis claims that the proportion of smokers among all adults in the United States is still 44%.

To examine these hypotheses, we calculate the proportion of smokers for our sample. If that proportion is "fairly close" to 44%, we will conclude that the population proportion may still be 44%. If not, we will conclude that the population proportion is no longer 44%.

Sample proportion = $\hat{p} = \frac{857}{2075} = 0.4130$

The notion of "fairly close" is based on the sampling distribution, as described in previous sections. Here are the calculations.

Sampling distribution:

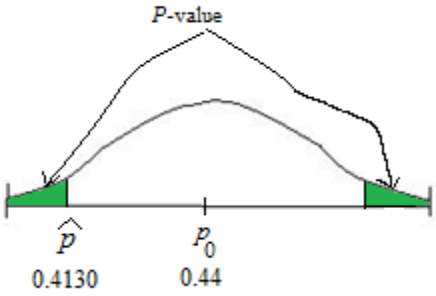
Mean = $p_0 = 0.44$

Standard error = $\sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.44(1-0.44)}{2075}} = 0.0109$

We know the sampling distribution is approximately normal, provided both np_0 and $n(1-p_0)$ are at least 15. Here are the calculations that justify the use of the normal distribution:

$np_0 = 2075(0.44) = 913$
$n(1-p_0) = 2075(1-0.44) = 1162$

This figure illustrates the situation. The calculated sample proportion is certainly less than the 44% in the null hypothesis. To measure how close it is to the null hypothesis we calculate the illustrated two-tail P-value.



The first step in this calculation is to find the *z*-score. We then use technology to find the two-tail P-value corresponding to this *z*-score.

$z = \frac{\hat{p}-p_0}{se} = \frac{0.4130-0.44}{0.0109} = -2.4771$
P-value = 0.0132

Since this P-value is less than the chosen significance level of 0.05, we reject the null hypothesis. We have evidence that the alternative hypothesis is true. We write our conclusion in terms of the alternative hypothesis: ***We have evidence, at $\alpha = 0.05$, that the proportion of smokers among***

U.S. adults is no longer 44%. Or, using the "significant" terminology: ***The proportion of smokers among U.S. adults differs significantly from the 44% that was reported in the 1960s.***

Example: An existing medication for a certain medical condition is known to provide a cure for 73% of patients treated with the medication. In a study of a newly developed medicine, 951 of the 1250 patients in the study are cured. Carry out an appropriate hypothesis test using significance level 0.01.

Solution: As usual, the null hypothesis will be of the form $p = p_0$, and will represent the idea "no change" or "no difference." In this case, "no difference" implies that the cure rate for the new medicine is no different from that of the existing medication.

$$H_0: p = 0.73$$

The alternative hypothesis will state that the null hypothesis is false.

$$H_a: p \neq 0.73$$

Note that these are claims about the population. But what is that population? Well, the sample consists of 1250 patients treated with the new medication. This sample must have been drawn from a population consisting of all patients who have been (or will be in the future) treated with this new medication. In words, the null hypothesis claims that the proportion of cures among all patients treated with the new medication will be 73% (the same as for the existing medication).

The hypothesis testing method we are using is based on a normal sampling distribution; to justify the use of the normal distribution we calculate both np_0 and $n(1 - p_0)$ – they should be at least 15 to justify the methods we use. Here are the pertinent calculations:

$$np_0 = 1250(0.73) = 912.5$$
$$n(1 - p_0) = 1250(1 - 0.73) = 337.5$$

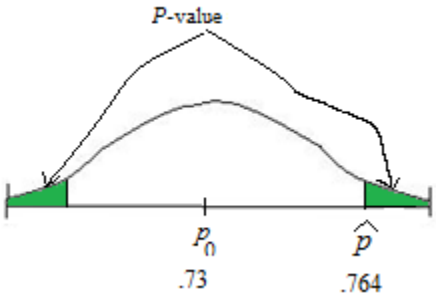
Since both numbers are at least 15, we move on to the calculations for the hypothesis test. First, we calculate the sample proportion – that is, the proportion of cures for the sample:

$$\text{Sample proportion} = \hat{p} = \frac{951}{1250} = 0.7608$$

The test will be based on the sampling distribution whose mean and standard error (that is, standard deviation) are given by these calculations.

$$\text{Mean} = p_0 = 0.73$$
$$\text{Standard error} = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.73(1-0.73)}{1250}} = 0.0126$$

This figure illustrates the situation. The calculated sample proportion is certainly more than the 73% in the null hypothesis. To measure how close it is to the null hypothesis we calculate the *z*-score and use that to find the illustrated two-tail P-value (using technology or Table A).



$z = \frac{\hat{p}-p_0}{se} = \frac{0.7608-0.73}{0.0126} = 2.4444$

P-value = 0.0145

Since this P-value is **not** less than the chosen significance level of 0.01, we are unable to reject the null hypothesis (or, put another way, we don't have enough evidence to support the alternative hypothesis). Although the proportion of cures *in this particular sample* was certainly more than 73%, that fact might be just a feature of the inherent randomness of selecting samples – it is conceivable that within the entire population of **all** persons treated with this new medicine the cure rate is 73%.

As always, we write our conclusion in terms of the alternative hypothesis. ***There is not enough evidence, at $\alpha = 0.01$, to conclude that the proportion of cures for the new medicine is any different from the 73% cure rate for the old medicine.*** Or, using the "significant" terminology: ***The proportion of cures for the new medicine is not significantly different from the 73% cure rate for the old medicine.*** Or, more simply: ***Researchers found no significant difference between the cure rates for the two medicines.***

Summary of hypothesis testing: Although the order in which we describe the calculations may differ somewhat from example to example, in general a hypothesis test for proportions involves the following steps:

- State the null and alternative hypotheses. Remember that the null hypothesis always has the form $H_0: p = p_0$.
- Check assumptions. Some of these are implicit in the statement of the problem, and are therefore not always mentioned – for example, we must be dealing with a categorical variable, and the sample must be randomly chosen from a much larger population. In terms of calculations, this step involves verifying that both np_0 and $n(1-p_0)$ are at least 15.
- Describe the sampling distribution on which the conclusion will be based. The calculations are

  Mean = p_0

  Standard error = $se = \sqrt{\frac{p_0(1-p_0)}{n}}$
- Find the sample proportion:

  $\hat{p} = \frac{count\ for\ sample}{n}$
- Calculate the test statistic (the *z*-score):

  $z = \frac{\hat{p}-p_0}{se}$

- Use the test statistic and technology (or Table A) to calculate a two-tail *P*-value.
- Compare the *P*-value to the significance level α to arrive at a conclusion. If the *P*-value is less than α, we reject the null hypothesis. Otherwise, we fail to reject the null hypothesis and acknowledge that the claim in the null hypothesis ***might*** be correct.

Exercise 17: In a large company, 45% of the non-management employees are female, and 55% are male. Management is starting a class to train future managers, and claims to have randomly chosen the participants from the available pool. It turns out that 47 of the 110 chosen are female. Use a hypothesis test to examine management's claim of having randomly chosen the participants.

Exercise 18: A newspaper article claims that 38% of the population of a large city favors a stricter drug control law. To investigate this claim, a statistics student randomly samples 200 residents of the city and finds that 80 favor the stricter law. Carry out an appropriate hypothesis test.

8.9 – One-Tail Tests

Hypotheses for one-tail tests

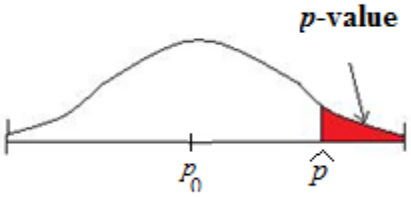
It occasionally happens that the researcher, *prior to taking the sample*, has a preconception of whether the proportion in the null hypothesis is too low or too high. In this case, the researcher may not use the neutral alternative hypothesis

$$H_a: p \neq p_0$$

Instead, the alternative hypothesis will reflect the belief of the researcher. For example, a researcher who believes the stated proportion is too low will write:

$$H_0: p = p_0$$
$$H_a: p > p_0$$

In this case, only samples whose proportion is unusually *large* will be considered to support the alternative hypothesis. Therefore, the *p*-value will be calculated as the probability, if the null hypothesis is true, of obtaining a sample proportion as large as, or larger than, that obtained in the study, as shown in this diagram:

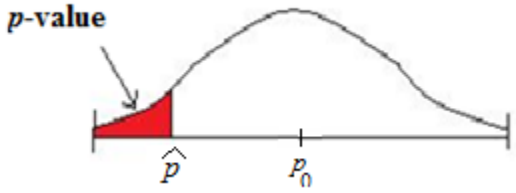


This is a one-tail (right-tail) *p*-value.

On the other hand, if the researcher believes the stated proportion is too high, he or she will write

$$H_0: p = p_0$$
$$H_a: p < p_0$$

and will calculate a one-tail (left-tail) *p*-value as depicted by this diagram:



Comment: We use a one-tail test only if there is some indication that the researcher has a *preconception* that the stated claim is too low or too high. This happens **before** the sample is taken. We do not use the sample's results to decide to do a one-tail test. For a professional statistician, this is never an issue, since the null and alternative hypotheses are established before the actual sampling begins. In this course, the entire study is described in a single paragraph, and if you are not careful you can let the stated results of the survey influence your choice of alternative hypothesis. This is never appropriate: you should use a one-tail test only if the problem statement tells you something about the researcher's *preconceptions*.

Note: There is a movement in the statistical community to do away with, or at least greatly restrict, the use of one-tail tests. The complete discussion of this issue is beyond the scope of this course. In this course, you should use a one-tail test if the problem statement tells you something about the researcher's preconceptions. Otherwise you should use a two-tail test.

Examples: For each of the following situations, describe the null and alternative hypotheses that would be used. Will the researcher calculate a right-tail, left-tail, or two-tail *P*-value?

1. A researcher believes that meditation improves clairvoyance – for example, that immediately following a 30-minute period of meditation your ability to anticipate whether a card from a standard deck of cards is red or black will be improved.
2. A researcher is investigating the claim that 10% of all drivers in a large city fail to wear a seatbelt. In a random sample of drivers in that city, she finds that 8% are not wearing a seatbelt.
3. A new non-addictive pain medication has been developed. The pharmaceutical company developing it hopes that as a side-effect it will reduce the incidence of heart attacks in persons taking the medication. Assume the probability of having a heart attack in the general population is 0.0025.

Solutions:

1. The null hypothesis always takes the form "no difference," or "no change," or "no effect." In this case, the null hypothesis asserts that the probability of guessing correctly will be the same

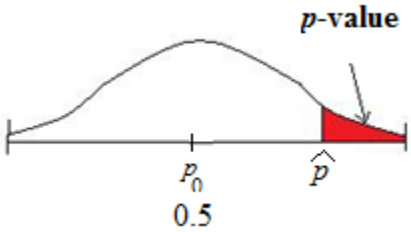
as it is without the meditation. Since half the cards in a standard deck of cards are red and half are black, a random guess has a 50% chance of being correct. Thus the null hypothesis is

$$H_0: p = 0.5$$

The statement of the problem indicates a preconception that the ability to guess correctly will be improved by meditation. That is, the researcher believes the probability of a correct guess will be better than 50%. Thus the alternative hypothesis is

$$H_a: p > 0.5$$

Only samples whose sample proportion is larger than 0.5 will support the alternative hypothesis, so the researcher will calculate a right-tail *P*-value, as illustrated here:



2. The null hypothesis states that there is no difference between the percentage in the city and the percentage that has been claimed:

$$H_0: p = 0.1$$

The statement of the problem does not indicate any preconception on the part of the researcher, so we will use a two-tail test and a two-tail *P*-value. The alternative hypothesis is

$$H_a: p \neq 0.1$$

Caution**:* Students are frequently misled by the reported 8% result from the study; since 8% is less than 10%, they might write $H_a: p < 0.1$ as the alternative hypothesis. The only time you should use a one-tail test is if there is a ***pre-conception (that is, a belief ***before*** the data is collected) on the part of the researcher.

3. The null hypothesis states that there is "no effect," that the incidence of heart attacks for the population of people taking this drug is / will be the same as for the general population:

$$H_0: p = 0.0025$$

The statement of the problem indicates a preconception that taking the drug will reduce the incidence ("the pharmaceutical company … hopes"). For purposes of this course, you should use this preconception to write the alternative hypothesis for a one-tail (left-tail) test using a left-tail *P*-value:

$$H_a: p < 0.0025$$

We note, however, that this is a prime example of the type of situation in which professional statisticians today would instead use a two-tail test. The reason is simple – although the

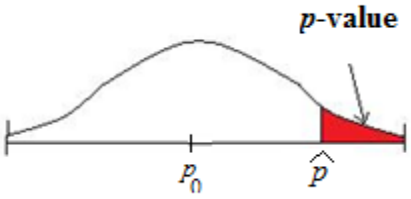
company hopes the percentage will be lowered, it could also be raised. Medical ethics would require the pharmaceutical company to investigate both possibilities. A two-tail test is designed to uncover differences from the proportion in the null hypothesis, *in either direction*, and thus this is the type of test which should be carried out.

Calculations for one-tail tests

Once you have chosen to do a one-tail test, and have written the null and alternative hypotheses, the remainder of the process is identical to that for a two-tail test, except that you will calculate a one-tail *p*-value rather than a two-tail *p*-value. As we noted earlier, if the alternative hypothesis has the form

$$H_a: p > p_0$$

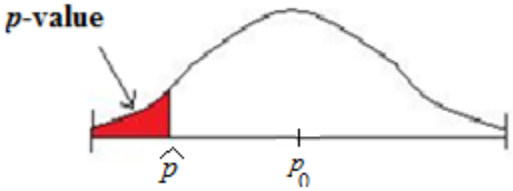
the conclusion will be based on a right-tail *p*-value as shown in this diagram:



Similarly, for alternative hypotheses of the form

$$H_a: p < p_0$$

the decision will be based on a left-tail *p*-value as depicted by this diagram:



Example. Consider the researcher who believes that meditation improves clairvoyance, who has chosen these hypotheses:

$$H_0: p = 0.5$$
$$H_a: p > 0.5$$

He carries out an experiment using a large number of subjects trying to anticipate the color of a large number of cards from standard decks of cards following a period of meditation. Out of 5000 cards, the subjects guess correctly for 2587 of the cards. Carry out the appropriate hypothesis test.

Solution. First, the size of the experiment is certainly large enough to use our methods:

$$np_0 = 5000(0.5) = 2500$$
$$n(1 - p_0) = 5000(1 - 0.5) = 2500$$

Both numbers are larger than 15. We therefore use a sampling distribution which is approximately normal, with

Mean = $p_0 = 0.5$

Standard error = $\sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.5(1-0.5)}{5000}} = 0.0071$

We calculate the sample proportion and the test statistic (*z* score), then use the *z* score to find the right-tail *p*-value using technology:

Sample proportion = $\hat{p} = \frac{2587}{5000} = 0.5174$

$z = \frac{\hat{p}-p_0}{se} = \frac{0.5174-0.5}{0.0071} = 2.4507$

P-value = 0.0071

At either $\alpha = 0.01$ or $\alpha = 0.05$, we would reject the null hypothesis, and conclude that for the population of person who have meditated 30 minutes prior to trying to guess cards, the proportion of correct guesses is indeed more than 50%.[4]

Comments:

1. If the researcher expects the sample proportion to be larger than the null hypothesis claim, and it is actually smaller, we *could* calculate a *p*-value. This diagram illustrates the situation:

However, there is really no need. Clearly the *p*-value will not be small; in fact, it will be greater than 50% (0.50).

A similar comment applies for the opposite situation, where the alternative hypothesis claims that *p* is less than p_0 but the sample proportion is greater than p_0.

2. As always, if we reject the null hypothesis, it means we accept the alternative hypothesis. If we fail to reject the null hypothesis, we acknowledge that the null hypothesis *might* be true.

Exercise 19: For the clairvoyance study, show the calculations and write the conclusion if the study had found that 2048 out of 4000 cards were guessed correctly.

[4] As noted earlier, some examples in these notes are based loosely on the types of studies which statisticians carry out, but contain data which has been created by the author for illustrative purposes. This example falls in that category.

Exercise 20: For the clairvoyance study, what would the researcher conclude if 1493 out of 3000 cards were guessed correctly?

Exercise 21: A researcher is doing a study with the following null and alternative hypotheses:

$H_0: p = 0.27$

$H_a: p < 0.27$

The sample, of size 1027, yields a sample proportion of 240/1027 = 0.2337. What conclusion should the researcher report? Use these steps:

a. Calculate the standard error, $se = \sqrt{\frac{p_0(1-p_0)}{n}}$.

b. Calculate the z score, $z = \frac{\hat{p}-p_0}{se}$.

c. Use technology (or Table A) to find the *p*-value (a left-tail one-tail *p*-value).

d. What is the conclusion, using significance level $\alpha = 0.01$?

The applet at the following link provides additional practice with the calculations for both one-tail and two-tail tests.

Proportion hypothesis test calculations

The next two applets provide additional practice in interpretation for both one-tail and two-tail tests.

Formulating hypotheses and drawing conclusions
Interpretation practice

8.10 – Comments on the Methodology

In Lessons 6 and 7 we discussed the issue of whether a researcher is justified in making a statement about an entire population, based on a survey of only a sample from that population. In this section, we remind you of that discussion, along with some additional ideas developed in the current lesson.

- Mathematically, the method is based on using a normal distribution. To do so, the sample size *n* must be "large enough." In practice, *n* is large enough provided:
  - $np_0 \geq 15$
  - $n(1-p_0) \geq 15$

- The researcher should report the chosen significance level (that is, α). Alternatively, the researcher may report the calculated P-value. This information is almost never reported by the media.

- We must take care to hear the results correctly. For example, consider this statement: "The proportion of smokers in that town is significantly different from the CDC's reported 16.8% for the nation as a whole." Correct interpretation of the statement involves at least the following:
    - The researchers have taken a sample from that town, and are using the results from the sample to make an inference about the entire town. In particular, they have rejected the null hypothesis that the proportion in the town *does match* the 16.8%.
    - In doing so, they have used a method with a known probability (perhaps 5%) of generating a Type 1 error.
    - The word "significant" in the statement does not imply "large." That word is only a statistician's way of stating that a null hypothesis was rejected.

- Similar comments apply when the reported results say something like, "No significant difference was found between the percentage of smokers today as compared to 10 years ago." In this case, the researcher did not reject the null hypothesis, and there is in general an unknown probability of having made a Type 2 error.

- In connection with the two previous bullet points, be aware that the media often will omit the word "significant" in the conclusion. You should be aware that all such conclusions arise from a study which has involved some sort of hypothesis. When you hear "there is a difference" that implies the null hypothesis was rejected; when you hear "there is no difference," the null hypothesis was not rejected.

- The mathematical process is based on the assumption that the sample was a simple random sample. In practice, this may be difficult to achieve. Professional statisticians have developed strategies beyond the scope of this course to help compensate for this difficulty, but there is always the possibility of obtaining a faulty estimate due to lack of randomness in the polling.

- A related problem occurs when a sample is taken from one group, but the results are reported for another group. A recent example occurred when a medical study was done in Singapore, but the results were reported as if the sample had been taken from the entire population of the world. For that particular study, it is possible that the proportion for Singapore and the proportion for the entire world were the same – but it is not obvious on the surface. (Imagine a presidential election poll which samples from only one state but reports the results as valid for the entire country! This would obviously be problematic.)

- Results can be skewed by the wording of the question. This can be unintentional or, in the case of unethical pollsters, intentional.

Solutions to Exercises

Some of the exercises have no specific solutions, since the results from running the applet will vary.

2. a. Use the applet to run between 900 and 1000 samples, with sample size 50 and closeness measure 4%. Record the results for percent correct here. Are your results fairly consistent with ours?

Here are the results obtained by the author. Yes, they are fairly consistent with the results reported in the discussion just before the exercise. The Type 1 and Type 2 error rates are both in the vicinity of 45-50%.

Results:	Decision made					
Actual status of dice	Keep	Discard	Percent correct decision		Error rates	
"Fair" dice:	294	244	For "fair" dice:	54.65%	Type 1:	45.35%
"Loaded" dice:	199	213	For "loaded" dice:	51.70%	Type 2:	48.30%

b. Do the same, but using sample size 1000 and closeness measure 4%. Record the results, and comment.

Here are the author's results; yours should be similar. It appears that using a much larger sample size (1000 rather than 50) while keeping the measure of closeness at 4% has greatly improved the percentage of correct decisions for fair dice, but the percentage of correct decisions for loaded dice has gone down quite a bit. (Type 1 error is 0% for this example, but Type 2 error is 88.56%.)

Results:	Decision made					
Actual status of dice	Keep	Discard	Percent correct decision		Error rates	
"Fair" dice:	564	0	For "fair" dice:	100.00%	Type 1:	0.00%
"Loaded" dice:	333	43	For "loaded" dice:	11.44%	Type 2:	88.56%

c. Do the same, but using sample size 1000 and closeness measure 2%. Record the results, and comment.

Here are the author's results; yours should be similar. Keeping the sample size at 1000 but making the measure of closeness smaller has, not surprisingly, led to more Type 1 error (discarding fair dice), but it has also greatly reduced the incidence of Type 2 error (keeping loaded dice).

Results:	Decision made					
Actual status of dice	Keep	Discard	Percent correct decision		Error rates	
"Fair" dice:	522	51	For "fair" dice:	91.10%	Type 1:	8.90%
"Loaded" dice:	95	280	For "loaded" dice:	74.67%	Type 2:	25.33%

d. Experiment with other combinations of sample size and closeness measure. Suppose you want to keep the Type 1 error rate near or below 5%. Are there any combinations that seem to meet this goal?

Based on the author's test runs, sample size 500 with closeness measure 4% seems to work. Similarly, sample size 1000 with either 3% or 4% as the closeness measure seem to work, with a lower incidence of Type 2 error using 3% as the closeness measure.

3. Use the applet to fill in the following table. The first row is already filled in based on the author's experimentation described above.

Sample size	Measure of closeness to achieve indicated Type 1 error rate	
	5%	1%
500	3.3%	4.2%
800	2.5%	3.5%
1000	2.3%	3.1%
1400	2.0%	2.6%

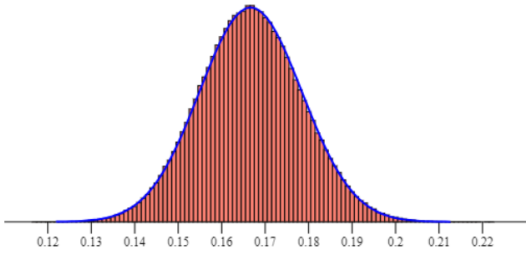
4. Use the applet to experiment. In particular, for each of the sample sizes (50, 100, 500, 1000) use the start sampling / stop sampling buttons to generate about 100,000 samples with an overload normal curve. For each sample size, record your answers to these questions:

- **a.** Does the histogram match the normal curve fairly closely?
- **b.** What is the mean for the histogram?
- **c.** What is the standard deviation?

Here are the author's results, yours should be similar:

n = 50: Pretty well, 0.1667, 0.0527
n = 100: Pretty well, 0.1668, 0.0373
n = 500: Very well, 0.1667, 0.0167
n = 1000: Very well, 0.1667, 0.0118

NOTE: For *n* = 1000, the author kept generating samples until over half a million samples had been generated, with this histogram as the result. The histogram matches the normal curve extremely closely.



5. Use your results from Exercise 4 to answer the following. Note that for each sample size, *p* is 1/6.

- **a.** For *n* = 50, calculate the mean *p* and the standard error $\sqrt{\frac{p(1-p)}{n}}$ for the sampling distribution, rounded to 4 places. Compare the mean and standard deviation you obtained in Exercise 4 to the mean and standard error for the theoretical sampling distribution.

Theoretical mean 0.1667, standard error = 0.0527
Author's results for histogram: mean 0.1667, standard deviation 0.0527 – yours may vary but should be similar

b. Do the same for n = 100.
Theoretical mean 0.1667, standard error = 0.0373
Author's results for histogram: mean 0.1668, standard deviation 0.0373

c. Do the same for n = 500.
Theoretical mean 0.1667, standard error = 0.0167
Author's results for histogram: mean 0.1667, standard deviation 0.0167

d. Do the same for n = 1000.
Theoretical mean 0.1667, standard error = 0.0118
Author's results for histogram: mean 0.1667, standard deviation 0.0118

6. Give the one-tail and two-tail P-values for these individuals:

a. Bill, 81 inches tall $z = \frac{81-70}{4} = 2.7500$. Right tail one-tail p-value is 0.0030, so two-tail p-value is 0.0060.

b. Ted, 58.5 inches tall $z = \frac{58.5-70}{4} = -2.8750$. Left tail one-tail p-value is 0.0020, so two-tail p-value is 0.0040.

7. In this exercise, we imagine we have already done the first step in calculating a P-value, namely calculating the z-score. For each of the indicated z-scores, calculate the one-tail and two-tail P-values. (For positive z-scores, the one-tail P-value to calculate is the right tail; for negative z-scores, the left tail.)

Note: The two-tail answers are found by doubling the "rounded-to-four-places" one-tail answers. If you do the doubling for the un-rounded one-tail answers, your two-tail answers may vary slightly.

a. z = 2.03 0.0212, 0.0424
b. z = 1.27 0.1020, 0.2040
c. z = –2.65 0.0040, 0.0080
d. z = –0.17 0.4325, 0.8650

8. Using the applet, the author tested several additional pairs of dice. The claim being tested is the same as in the discussion: The probability / long-term proportion is 1/6 (approximately 0.1667 or 16.67%). Use 0.1667 in your calculations.

In each case, the sample size was n = 1000. The proportion of 7s, that is $\hat{p}$, is reported below. Calculate the corresponding z-score and (two-tail) p-value. Using 0.05 as your criterion for the decision, state your decision (discard or keep the dice).

Recall we are using s.e.(standard error) =0.0118, which was calculated as $\sqrt{\frac{.1667(1-.1667)}{1000}}$.
We use technology, with z-scores rounded to four places and by doubling the un-rounded one-tail p-value; your answers may vary slightly if you round the z-score and/or the one-tail p-value differently/

a. $\hat{p} = 17.10\%$ z = 0.3644, p-value = .7156. Since this is not less than 0.05, we keep the dice.

b. $\hat{p} = 13.10\%$ z = –3.0254, p-value = .0025. Since this is less than 0.05, we reject the claim and discard the dice.

c. $\hat{p} = 15.30\%$ z = –1.1610, *p*-value = .2456. Since this is not less than 0.05, we keep the dice.

Comment. Because the applet knows whether the dice were fair or not, we can tell you that the conclusion was correct for (a) and (b), and incorrect – a Type 2 error – for (c), for the particular dice tested.

9. Using the applet, the author tested a pair of dice using a sample of size 500, obtaining $\hat{p} = 15.20\%$, and another pair of dice with $\hat{p} = 12.80\%$. For each pair of dice, calculate the *z*-score and (two-tail) *p*-value, and state your conclusion. Use 0.05 as the criterion for the decision.

Hint: You will need to recalculate the standard error, $s.e. = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.1667(1-0.1667)}{500}}$, because *n* is no longer 1000.

s.e. = $\sqrt{\frac{.1667(1-.1667)}{500}} = 0.0167$

First pair of dice: $z = \frac{.1520-.1667}{0.0167} = -0.8802$, *p*-value = .3788. Since this is not less than 0.05, we keep the dice.

Second pair: $z = \frac{.1280-.1667}{0.0167} = -2.3174$, *p*-value = .0205. Since this is less than 0.05 we reject the claim, and discard the dice.

10: **a.** For which, if any, of the tests carried out in Exercises 8 and 9 would you have reached a different decision if you had used 0.01 rather than 0.05 as your criterion?

Exercise 8(b), and the second pair of dice in Exercise 9.

b. True or false. Using 0.01 rather than 0.05 will reduce the likelihood of Type 1 errors (discarding fair dice).

True, and it will increase the likelihood of Type 2 errors (keeping loaded dice).

11: Do an analysis similar to this example, for each of the following situations.

a. There were 723 even rolls in a sample of 1500 rolls.

s.e. = $\sqrt{\frac{0.5(1-0.5)}{1500}} = 0.0129$

$\hat{p} = \frac{723}{1500} = 0.482$

$z = \frac{0.482-0.5}{0.0129} = -1.3953$

using technology, the two-tail P-value is 0.1629

for both 0.05 and 0.01 the answer is the same: keep the dice

b. There were 653 even rolls in a sample of 1200 rolls.

s.e. = $\sqrt{\frac{0.5(1-0.5)}{1200}} = 0.0144$

$\hat{p} = \frac{653}{1200} = 0.5442$

$z = \frac{0.5442-0.5}{0.0144} = 3.0694$

using technology, the two-tail P-value is 0.0021

for both 0.05 and 0.01 the answer is the same: discard the dice

c. There were 585 even rolls in a sample of 1097 rolls.

s.e. = $\sqrt{\frac{0.5(1-0.5)}{1097}} = 0.0151$
$\hat{p} = \frac{585}{1097} = 0.5333$
$z = \frac{0.5333-0.5}{0.0151} = 2.2053$
using technology, the two-tail P-value is 0.0274
using 0.05 we discard the dice; using 0.01 we keep the dice

12: In the previous example, suppose the researcher had decided, when planning the study, to use 0.01 as the criterion for the decision. Describe the decision that would be made in that case.

The P-value is not smaller than 0.01, so we do not reject the claim. We are unable to conclude that the proportion of smokers in this town is not 16.8%. Notice that this is not the same as saying the proportion *is* 16.8%; we simply acknowledge that it *could be* 16.8%.

13: In this exercise, you will carry out the same hypothesis test but using several different hypothetical sample results. For each, write the claim being investigated, show the calculations, and give the final decision (reject the claim, or fail to reject the claim).

a. 99 out of 700 are smokers
b. 266 out of 1365 are smokers
c. 68 out of 520 are smokers

	Standard error	$\hat{p}$	***z***	***P*-value**	**Do you reject claim?**	
					Using 0.05	**Using 0.01**
a.	$\sqrt{\frac{.168(1-.168)}{700}} = 0.0141$	$\frac{99}{700} = 0.1414$	$\frac{.1414-.168}{.0141} = -1.8865$	**0.0592**	**No**	**No**
b.	$\sqrt{\frac{.168(1-.168)}{1365}} = 0.0101$	$\frac{266}{1365} = 0.1949$	$\frac{.1949-.168}{.0101} = 2.6634$	**0.0077**	**Yes**	**Yes**
c.	$\sqrt{\frac{.168(1-.168)}{520}} = 0.0164$	$\frac{68}{520} = 0.1308$	$\frac{.1308-.168}{.0164} = -2.2683$	**0.0233**	**Yes**	**No**

14: Should we believe that a certain population proportion is 0.695? Based on a sample in which 755 people answer yes to the question posed, carry out an appropriate two-tail test. (The sample consisted of 1149 people.)

First, we state the null and alternative hypotheses:
$H_0: p = 0.695$
$H_a: p \neq 0.695$
The decision will be based on the sampling distribution, which is normal with:
Mean = p_0 = 0.695
Standard error = $\sqrt{\frac{.695(1-.695)}{1149}} = 0.0136$
Here are the calculations based on the sample:
$\hat{p} = \frac{755}{1149} = 0.6571$
$z = \frac{0.6571-0.695}{0.0136} = -2.7868$
P-value = 0.0054

Conclusion:

Because 0.0054 < 0.05, we reject the null hypothesis, and conclude that the population proportion is not 0.695. (Notice that we would have reached the same conclusion if we had used $\alpha = 0.01$.)

Note: Here are the calculations that justify the use of the normal distribution; both numbers calculated are definitely ≥ 15.

1149(0.965) = 798.555

1149(1 – 0.965) = 350.445

15: Researchers compare the success rate for a new drug to the known success rate (73%) for an existing drug, using the following null and alternative hypotheses:

$H_0: p = 0.73$

$H_a: p \neq 0.73$

a. If they reject the null hypothesis, how will they write the conclusion in the context of the study? We are able to conclude that the success rate for the new drug is not 73%. Or perhaps: The success rate for the new drug differs from the 73% success rate of the established drug.

b. How will they write the conclusion if they do not reject the null hypothesis? There was not enough evidence to show that the success rate for the new drug is not 73%. Or perhaps: We could not conclude that the success rate for the new drug is any different from that of the established drug.

16: For the situation described in exercise 15, use the word "significant" to write the conclusion if:

a. they reject the null hypothesis. The success rate for the new drug differs significantly from the 73% rate for the existing drug. (Or perhaps: Researchers found a significant difference between the new drug and the existing drug.)

b. they do not reject the null hypothesis. The success rate for the new drug does not differ significantly from the 73% rate for the existing drug. (Or perhaps: Researchers found no significant difference between the new drug and the existing drug.)

17: In a large company, 45% of the non-management employees are female, and 55% are male. Management is starting a class to train future managers, and claims to have randomly chosen the participants from the available pool. It turns out that 47 of the 110 chosen are female. Use a hypothesis test to examine management's claim of having randomly chosen the participants.

- State the null and alternative hypotheses.

  $H_0: p = 0.45$ (In words: in the population of all persons that would ever be chosen for the class, 45% would be female – that is, the probability of a female being chosen is 45%.)

  $H_a: p \neq 0.45$
- Check assumptions. Verify that both np_0 and $n(1 - p_0)$ are at least 15.

  np_0= 110(0.45) = 49.5

  $n(1 - p_0) = 110(1 - 0.45) = 60.5$
- Describe the sampling distribution on which the conclusion will be based. The calculations are

  Mean = $p_0 = 0.45$

  Standard error = $\sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.45(1-0.45)}{110}} = 0.0474$
- Find the sample proportion:

  $\hat{p} = \frac{count\ for\ sample}{n} = \frac{47}{110} = 0.4273$

- Calculate the test statistic (the *z*-score), and use the test-statistic to calculate a two-tail P-value.
  $z = \frac{\hat{p}-p_0}{se} = \frac{0.4273-0.45}{0.0474} = -0.4789$
  Using technology, P-value = 0.6320
- Compare the *P*-value to the significance level α to arrive at a conclusion.
  The problem does not state a significance level, so we use $\alpha = 0.05$. Since 0.6320 is *not* less than 0.05, we fail to reject the null hypothesis and acknowledge that the claim in the null hypothesis ***might*** be correct. (Note that the conclusion would be the same at $\alpha = 0.01$.)
  In the context of the problem: At $\alpha = 0.05$, there is no evidence that the long-term proportion of females that would be chosen differs from the 45% of females in the company.

18: A newspaper article claims that 38% of the population of a large city favor a stricter drug control law. To investigate this claim, a statistics student randomly samples 200 residents of the city and finds that 90 favor the stricter law. Carry out an appropriate hypothesis test.

- State the null and alternative hypotheses.
  $H_0: p = 0.38$ (In words: in the entire city, 38% favor the stricter drug control law)
  $H_a: p \neq 0.38$
- Check assumptions. Verify that both np_0 and $n(1-p_0)$ are at least 15.
  np_0= 200(0.38) = 76
  $n(1-p_0) = 200(1-0.38) = 124$
- Describe the sampling distribution on which the conclusion will be based. The calculations are
  Mean = $p_0 = 0.38$
  Standard error = $\sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.38(1-0.38)}{200}} = 0.0343$
- Find the sample proportion:
  $\hat{p} = \frac{count\ for\ sample}{n} = \frac{90}{200} = 0.4500$
- Calculate the test statistic (the *z*-score), and use the test-statistic to calculate a two-tail P-value.
  $z = \frac{\hat{p}-p_0}{se} = \frac{0.4500-0.38}{0.0343} = 2.0408$
  Using technology, P-value = 0.0413
- Compare the *P*-value to the significance level α to arrive at a conclusion.
  The problem does not state a significance level, so we use $\alpha = 0.05$. Since 0.0413 is less than 0.05, we reject the null hypothesis. (Note that the conclusion would be different at $\alpha = 0.01$.)
  In the context of the problem: At $\alpha = 0.05$, there is evidence that the proportion of persons in the city who favor stricter drug control laws is *not* the 38% claimed in the newspaper article. (Based on the sample, we believe it is higher than 38%.)

19: For the clairvoyance study, show the calculations and write the conclusion if the study had found that 2048 out of 4000 cards were guessed correctly.

- Check assumptions. Verify that both np_0 and $n(1-p_0)$ are at least 15.
  np_0= 4000(0.5) = 2000
  $n(1-p_0) = 4000(1-0.5) = 2000$
- Describe the sampling distribution on which the conclusion will be based. The calculations are
  Mean = $p_0 = 0.5$
  Standard error = $\sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.5(1-0.5)}{4000}} = 0.0079$
- Find the sample proportion:
  $\hat{p} = \frac{count\ for\ sample}{n} = \frac{2048}{4000} = 0.5120$

- Calculate the test statistic (the *z*-score), and use the test-statistic to calculate a right-tail P-value.
  $z = \frac{\hat{p}-p_0}{se} = \frac{0.5120-0.5}{0.0079} = 1.5190$
  Using technology, P-value = 0.0644
- Compare the *P*-value to the significance level α to arrive at a conclusion.
  The problem does not state a significance level, so we use α = 0.05. Since 0.0644 is not less than 0.05, we do not reject the null hypothesis.
  In the context of the problem: At α = 0.05, there is not enough evidence to conclude that the proportion of correct guesses for person who have meditated exceeds 50%. More briefly, we conclude that meditation does not significantly improve clairvoyance.

20: For the clairvoyance study, what would the researcher conclude if 1493 out of 3000 cards were guessed correctly?

We could carry out calculations similar to those in Exercise 19, but there really is no need to. Since 1493 is less than half of 3000, this study does not support the alternative hypothesis. The right-tail *p*-value, if calculated, would be large – in fact, more than 0.5 = 50%. We fail to reject the null hypothesis.

21. A researcher is doing a study with the following null and alternative hypotheses:

$H_0: p = 0.27$

$H_a: p < 0.27$

The sample, of size 1027, yields a sample proportion of 240/1027 = 0.2337. What conclusion should the researcher report? Use these steps:

a. Calculate the standard error, $se = \sqrt{\frac{p_0(1-p_0)}{n}}$. $se = \sqrt{\frac{0.27(1-0.27)}{1027}} = 0.0139.$

b. Calculate the z score, $z = \frac{\hat{p}-p_0}{se}$. $z = \frac{0.2337-0.27}{0.0139} = -2.6115$

c. Use technology (or Table A) to find the *p*-value (a left-tail one-tail *p*-value). 0.0045

d. What is the conclusion, using significance level $\alpha = 0.01$? Since the *p*-value is less than 0.01, the researcher rejects the null hypothesis, and writes something like this: "At significance level α = 0.01, we found evidence that the population's proportion is less than 27%." Another way to state this would be to state that the population proportion is significantly less than 27%.

Note: if the P-value had been, for example, 0.0317, the conclusion would have been something like this: "At α = 0.01 we were unable to conclude that the population's proportion is less than 27%." Another way to state this would be to state that the population proportion is **not** significantly less than 27%.