Lesson 6
Introduction to Inferential Statistics
Hypothesis Tests

Outline of the Lesson

In late May of 2011, the World Health Organization (WHO) reversed its previous stand, and announced its conclusion that radiation from cell phones can possibly cause cancer. This conclusion was based on a number of studies, one of which was a large international study released in 2010. That study showed that participants in the study who used a cell phone for 10 years or more had double the rate of a particular type of brain cancer. This is an example of ***inferential statistics***. The international study included a large number of people, but it certainly did not include everyone in the world. Yet the WHO statement was a statement that applied to everyone. The WHO made an *inference* about the entire population of the world, based on the results of the study. In general, *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.

The WHO study is an example of a very common form of inference, in which the researchers examine the association between two variables. In this example, the variables are cell phone usage and contracting cancer. The researchers concluded that there is an association between the variables, that using a cell phone can increase the probability of contracting cancer. Other examples of this type of study abound. For example, is there a connection between gender and political party preference? Does a certain

test used in connection with hiring firefighters for the city discriminate against minorities? Is there a relationship between gender and the likelihood of receiving a promotion in this company? We will study this type of inferential statistics in detail in Lesson 10.

For now, we begin with a simpler form of inference. This will allow us to develop the tools and skills that underlie all forms of inferential reasoning, but within a context that is easier to understand. In this lesson and the next two, we will concentrate on a single population, and a categorical variable with only two possible values. Here is a simple example of the type of study we have in mind.

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)?

To answer this question, one possibility would be to ask every adult in the entire town whether or not they are a smoker. However, that would be very time-consuming and very expensive. The approach taken by statisticians to answer this type of question can be summarized as follows:

- The population being studied consists of all the adults in that particular town. The variable is the categorical variable indicated by this question: Are you a smoker (YES/NO)?
- We want to examine the claim that the proportion of smokers in that town actually is 16.8%.
- Take a random sample of adults in the town. Measure the proportion of smokers in this sample.
- If the proportion in the sample is not very close to 16.8%, we conclude that the proportion in the entire town is not 16.8%; otherwise, we acknowledge that the proportion in the entire town could be 16.8%.

This description raises several questions which we will attempt to answer in this lesson and the next. For example:

1. How is it possible to draw conclusions about a group that is larger than the group you actually questioned?
2. Is this process legitimate? If so, what precautions must we take in interpreting the result?
3. What exactly does the phrase "not very close to 16.8%" mean? More generally, exactly how do statisticians make their decision?

In our analysis of the process (whether it works, how well it works, how it works) we will take advantage of the connection between proportions and probabilities. For this poll, there are two ways to interpret the 16.8% figure we are investigating:

- The proportion of all adults in that town who are smokers is 16.8%.
- The probability that a randomly selected adult from that town would be a smoker is 0.168.

Building on this connection between proportions and probabilities, we begin our explanation by thinking about probabilities for random events such as coin tosses and rolls of dice. In the next lesson we return to this example, and apply what we learn to the polling process.

6.1 – 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%. In making this statement, we are assuming the coin is a "fair" coin, not weighted in any way

to make one side more likely to show up than the other. 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.

What does this theoretical probability mean in practical terms? Does it mean that if you toss the coin twice, you must get exactly one head and one tail? Obviously not, our experience indicates that two heads or two tails can easily occur in succession. Does it mean that if you toss the coin 10 times, exactly 5 will be heads; or that you will get 500 heads out of 1000 tosses? Again, no – the randomness of tossing a coin implies that the sequences of heads and tails will be unpredictable, and that the exact number of heads obtained will also be unpredictable. So, what are the implications of stating that the probability of obtaining heads is 0.5? The answer is this:

> *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:

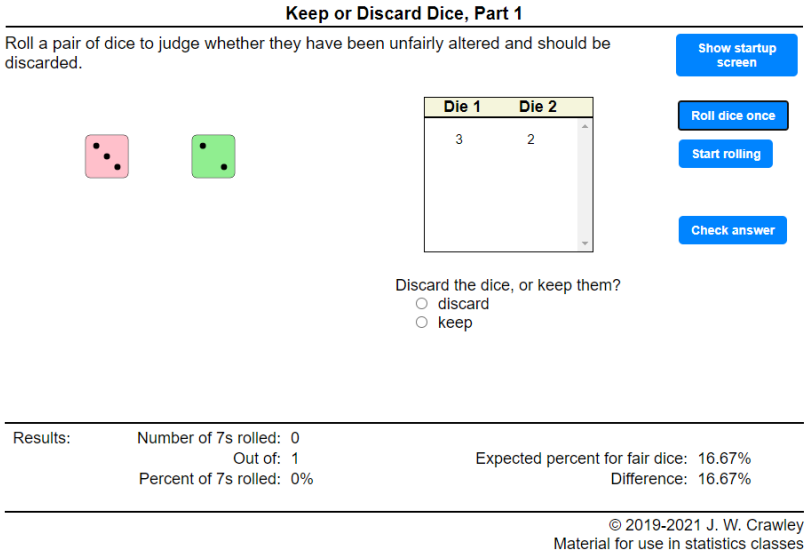
- Roll the dice many times.
    - If the proportion of 7s is not close to 1/6, we have evidence that the probability is not 1/6. We will **reject** the claim, and conclude that the dice are loaded. We will describe our decision as "discard the dice."
    - If the proportion of 7s is close to 1/6, we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

An applet for experimenting

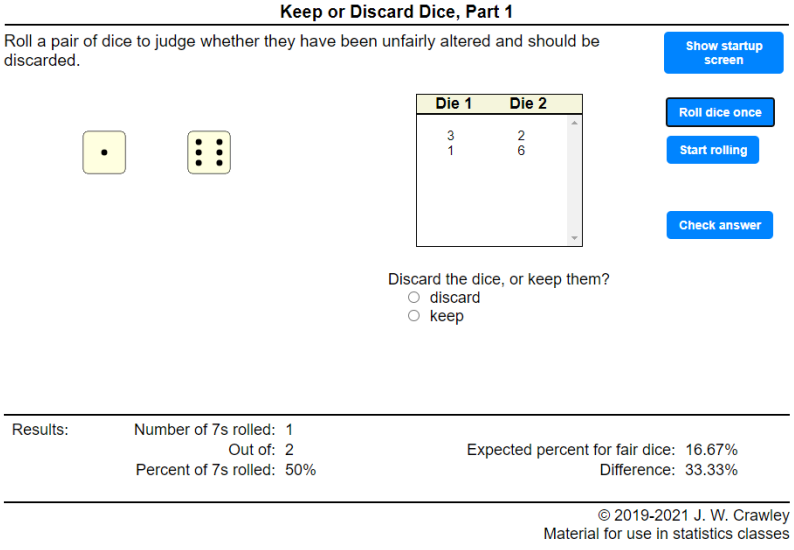
To allow you to experiment with this example, we have created the applet "Keep or Discard Dice, Part 1," at this link:

Keep or Discard Dice Part 1

Here is what you should do, using the applet. Roll the dice as many times as you wish. The applet will keep track of how many 7s are rolled, and the proportion of 7s rolled. For example, here is what happened when the author used the applet to roll the dice once (by using the "Roll dice once" button):

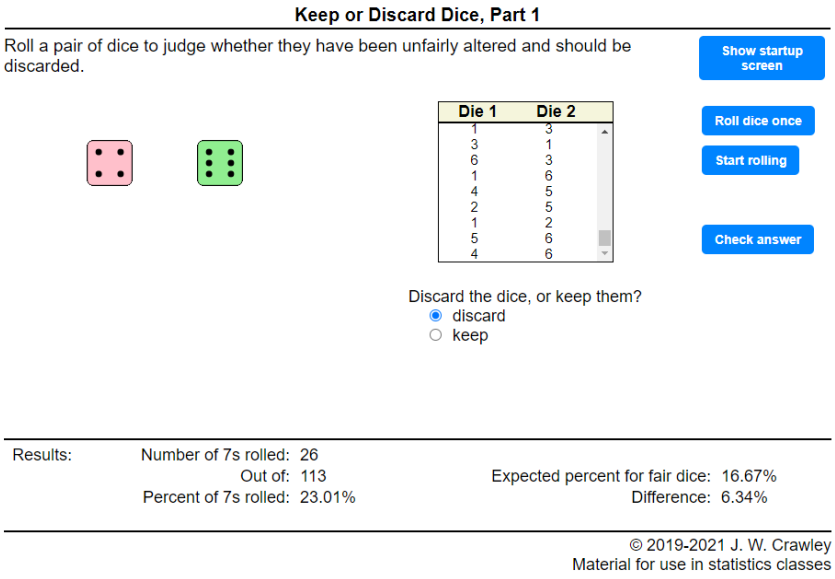


This roll happens to have *not* been a seven. Out of 1 roll so far, there have been 0 times that the result was a seven, for a percent of 0 / 1 = 0.00%. Now here is a second roll.



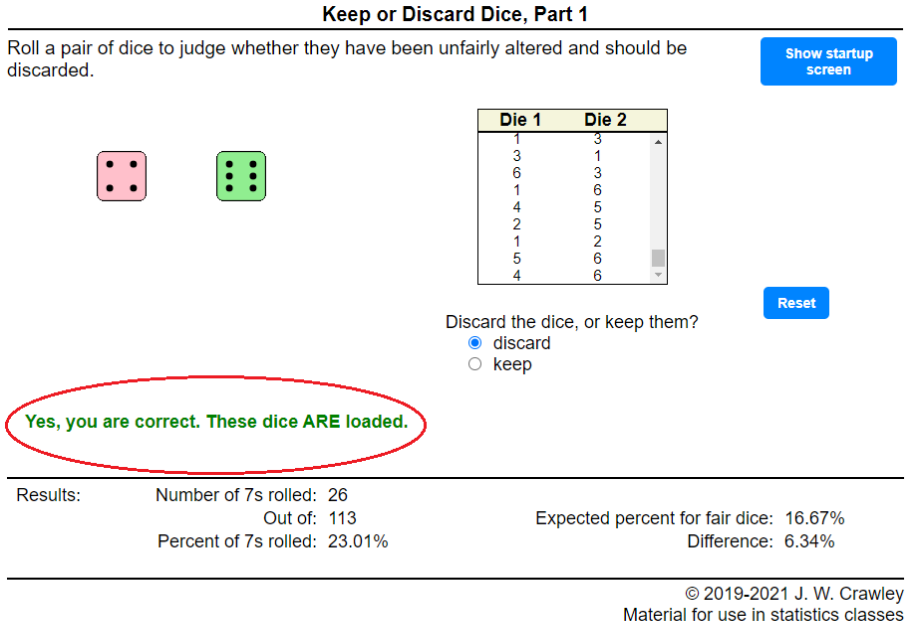
This time the total on the dice was seven – when that happens the applet colors both dice light yellow to highlight the match. Notice that both rolls are shown in the table to the right of the dice, and that the results near the bottom of the screen now show one seven out of two rolls, that is 50%.

Obviously, it is way too early to make any kind of intelligent decision about these dice. Using either the "Roll dice once" button repeatedly, or else the "Start rolling" button followed by the "Stop rolling" button, we can roll the dice as many times as we wish. Here is what happened when the author rolled the dice a total of about 100 rolls.



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. He has chosen the "discard" radio button to answer the "Discard the dice, or keep them?" question.

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:



Using the "Reset" button, 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:

Keep or Discard Dice, Part 1

Roll a pair of dice to judge whether they have been unfairly altered and should be discarded.

Show startup screen

Die 1	Die 2
2	2
4	2
3	2
5	5
3	3
6	5
1	6
4	1
3	3

Reset

Discard the dice, or keep them?
- ○ discard
- ◉ keep

Yes, you are correct. These dice are NOT loaded.

Results:			
	Number of 7s rolled:	21	
	Out of:	105	
	Percent of 7s rolled:	20%	
		Expected percent for fair dice:	16.67%
		Difference:	3.33%

Keep or Discard Dice, Part 1

Roll a pair of dice to judge whether they have been unfairly altered and should be discarded.

Show startup screen

Die 1	Die 2
3	3
1	3
2	5
3	5
1	3
3	5
1	5
1	1
2	2

Reset

Discard the dice, or keep them?
- ○ discard
- ◉ keep

No, you are not correct. These dice ARE loaded and should be discarded.

Results:			
	Number of 7s rolled:	17	
	Out of:	110	
	Percent of 7s rolled:	15.45%	
		Expected percent for fair dice:	16.67%
		Difference:	1.22%

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?
6.34%	yes	correct
3.33%	no	correct
1.22%	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 of closeness, 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

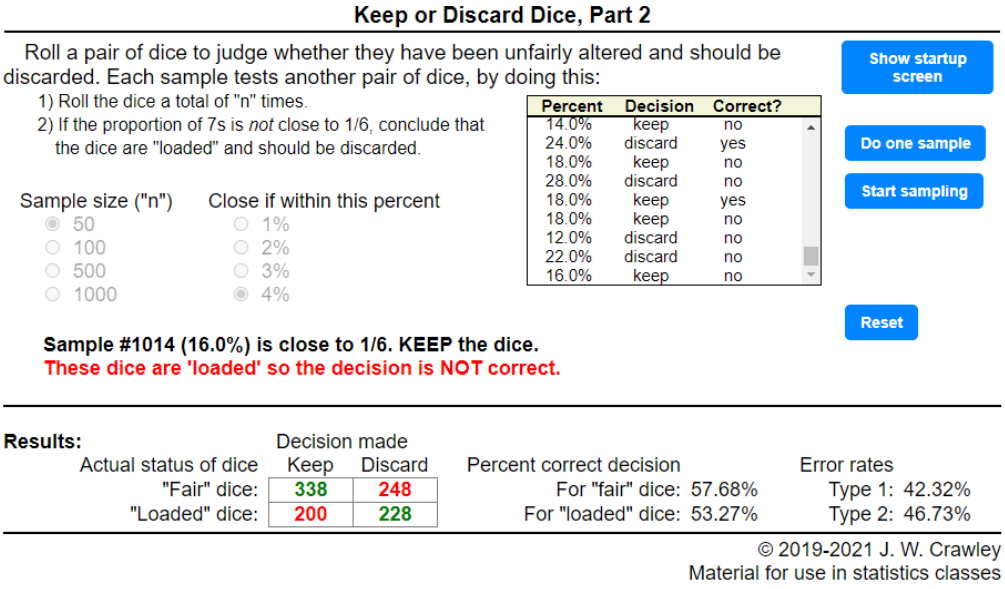
In the first type of error, we have incorrectly concluded that the manufacturer has provided loaded dice. The manufacturer's claim, "The proportion of 7s is 16.67%" was true, but we have concluded that it was false. This is called a Type 1 error, and it can be controlled in a predictable manner by our 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:



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.

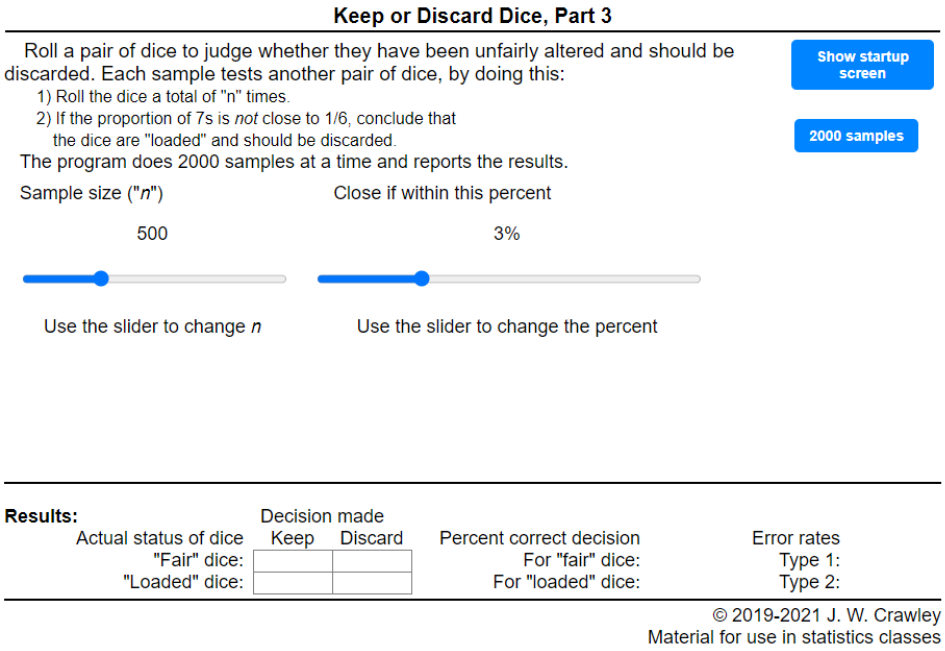
Controlling Type 1 error

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:

Actual status of dice	Decision made: Keep	Decision made: 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

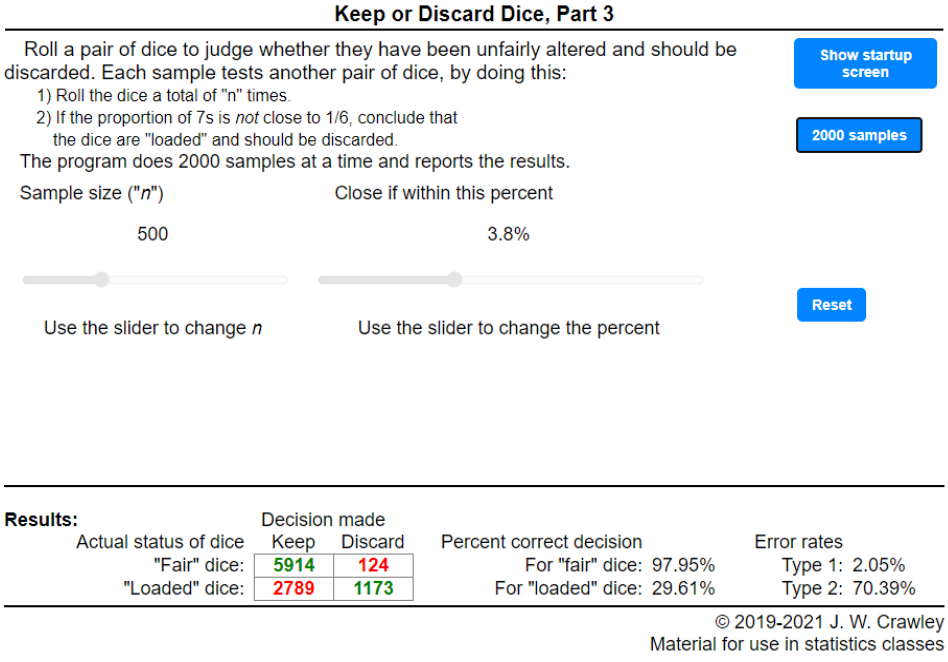
In the author's results recorded above, we can summarize what happened as follows:

- 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.

6.2 – 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, about a probability / long-term proportion to be investigated. 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; this is called 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:

1. As noted earlier, 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 a variable called *p*-hat, written $\hat{p}$. In the dice example, this is the proportion of 7s we obtain when we roll the dice *n* times.
3. We use the variable *p*, without the "hat," to indicate 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
$$p = 1/6$$

Using this notation, 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? Based on the results we have seen in the applets, 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 in Section 6.5.

The basis for decision-making (the sampling distribution)

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. In this subsection we will learn about the calculations which are used by statisticians in their decision strategy.

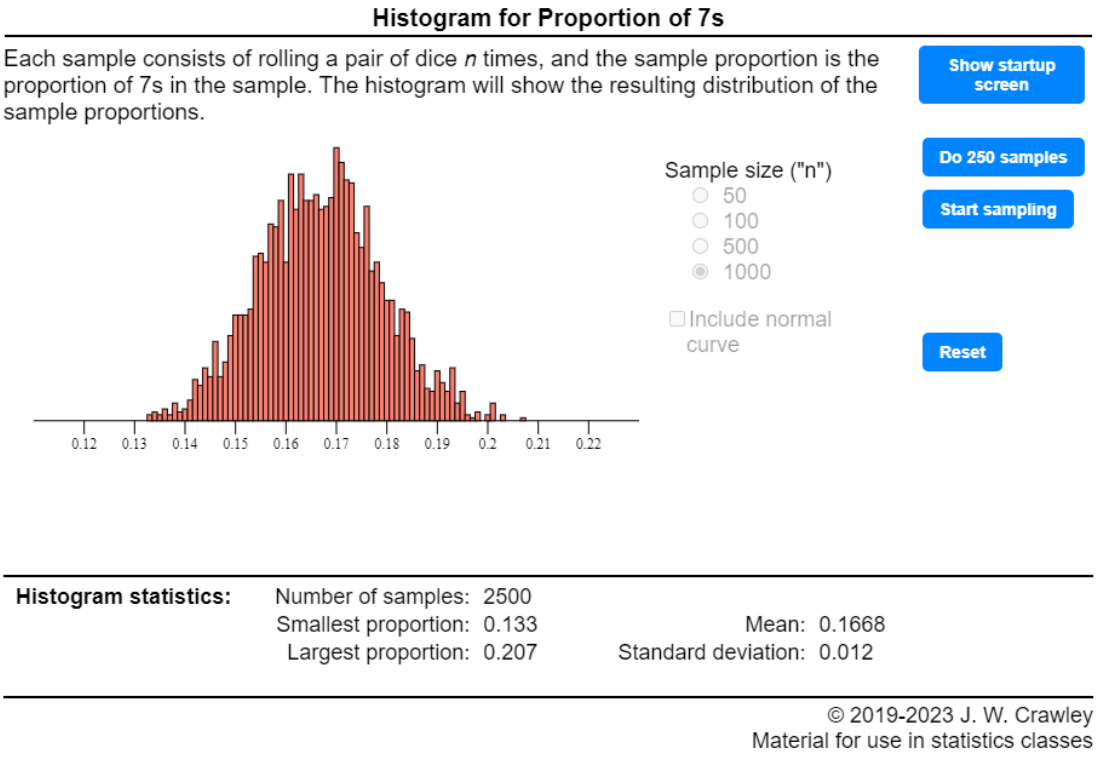
In order to control the occurrence of Type 1 error (discarding a fair pair of dice), it will be useful to consider what "should" happen when we roll a fair pair of dice *n* times and measure the sample proportion $\hat{p}$ for that sample. The answer (provided *n* is large enough) is that we expect the 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, although this may not be obvious, it will be very close more often than it will be not so close.

The link below runs an applet that illustrates this. In the applet you can select a sample size *n*, with the default being 1000. When you click on the "Do 250 samples" button to take 250 samples, each

consisting of n rolls of a fair pair of dice. For each sample it will calculate the sample proportion, $\hat{p}$, that is the proportion of 7s that occur. Finally, it will make a histogram of all these $\hat{p}$ values. From the graph you will be able to 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). You can use the controls to 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.

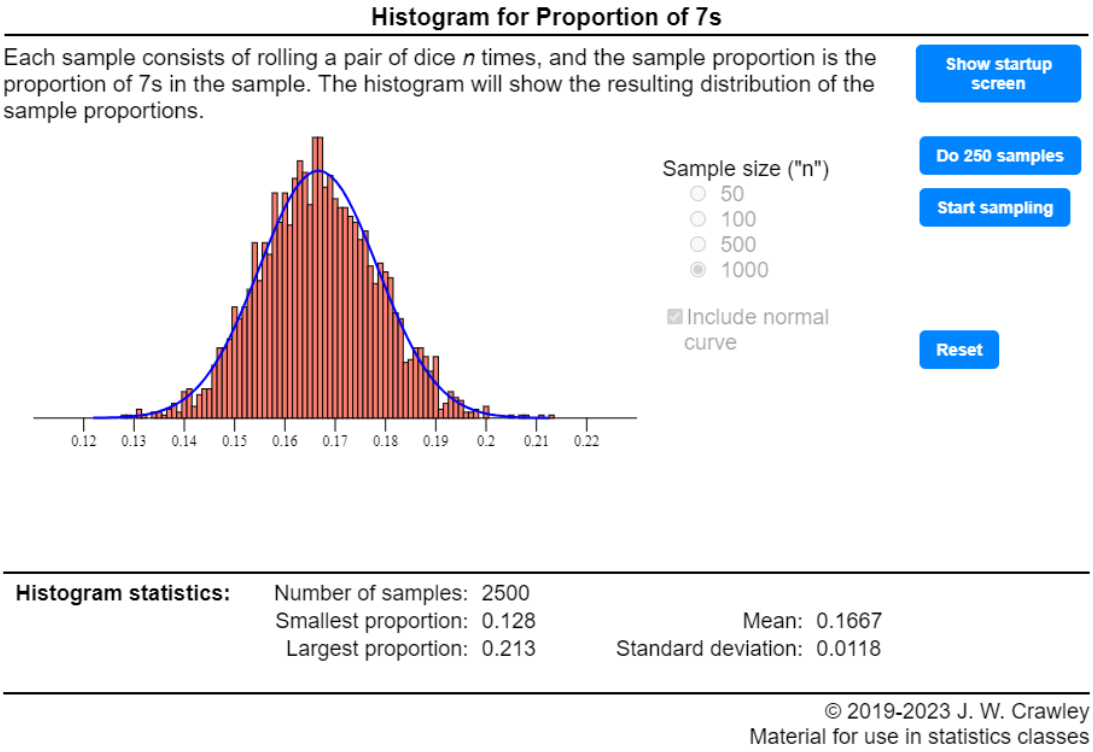
Histogram of p-hat values

For example, here are the results obtained by the author using sample size 1000 and generating 2500 samples.



Exercise 4. Use the applet to reproduce what the author did – that is, generate 2500 samples each with n = 1000. Are your results consistent with those shown above?

When the author worked Exercise 4 several times, each histogram's shape was similar to the last. They were all mound shaped, resembling a normal curve. The applet allows you to overlay a normal curve with the same mean and standard deviation as the histogram to see how closely the histogram matches the normal curve. Here are the author's results using this feature of the app:



As you can see, the histogram shape is similar to that of the normal curve. The following exercise will examine this phenomenon further.

Exercise 5. 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: 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.** We will use this terminology more in the next lesson.

[1] 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

Mathematicians have established three fundamental facts about this theoretical sampling distribution:

- The mean of the distribution is equal to the probability (long-term proportion) – what we have labeled with the variable *p*.
- The standard deviation of the distribution can be calculated using the formula $\sqrt{\frac{p(1-p)}{n}}$
- Provided *n* is large enough, the theoretical sampling distribution not just mound-shaped, it is in fact approximately normal. (Note that this is consistent with your, and the author's, results in Exercise 5.)

Notes:

1. For the situation pictured above, *n* was 1000 and *p* was 1/6. These theoretical formulas yield the following for the mean and standard deviation for the theoretical sampling distribution:

Mean = $p = 1/6 = 0.1667$

Standard deviation = $\sqrt{\frac{p(1-p)}{n}} = 0.0118$

Notice that the results of running the applet (the run without the overload normal curve) are pretty close to this expected result – the mean for the 2500 samples was 0.1668 with a standard deviation of 0.012.

2. For reasons beyond the scope of this course, it is customary to refer to the standard deviation for a sampling distribution as ***standard error***. We will follow that custom in what follows, but whenever you see the term "standard error" you should remind yourself that it is nothing more than a standard deviation.

Exercise 6. Use your results from Exercise 5 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 5 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.

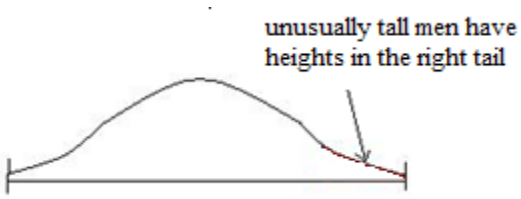
We can use what we know about the sampling distribution to build a viable strategy for deciding whether to discard or keep the dice. Put very simply, and in terms of the dice example, the decision strategy can be summarized as follows:

- When we roll a pair of fair dice *n* times (that is, obtain a sample of rolls for that pair of dice), the result is most likely to be in the middle part of the sampling distribution. Very seldom will rolling a pair of fair dice generate a proportion of 7s which falls out in the tails of the sampling distribution.
- Therefore, if we discard dice when we obtain a proportion of 7s that 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.

6.3 – 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.

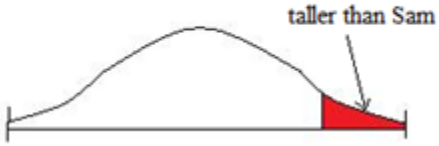
We can quantify this rather vague notion of "unusual," using the fact that the distribution of adult male heights is approximately normal (with mean 70 and standard deviation 4). The concepts we develop can be applied to any situation involving a normal distribution, and indeed they can be generalized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

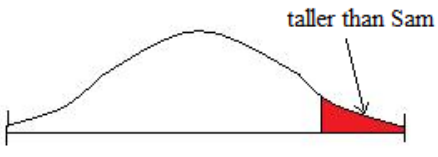
Solution: We answer this question by calculating the probability that a randomly chosen adult male will be as short as, or shorter than, Joe, as illustrated in this figure:

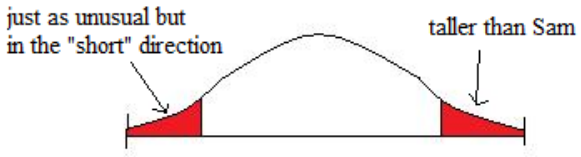
Again, we calculate a *z* score for Joe's height, $z = \frac{59-70}{4} = -2.7500$, then use the methods of Lesson 4 to calculate that the area to the left of this *z* score in the standard normal distribution is 0.0030. In the graph, the shaded area is 0.0030; this means that 0.30% of adult males are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

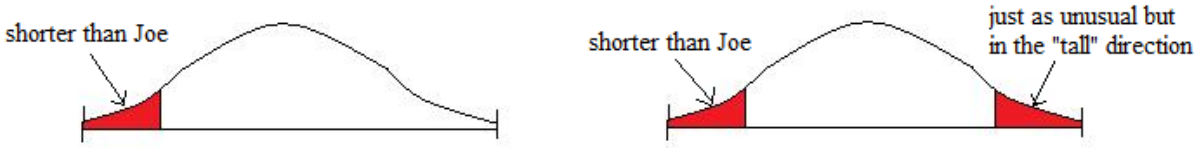
For a particular observation, the two-tail P-value measures the probability of being this far away from, or further away from, the mean (or average) of the distribution – *in either direction.*

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

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

Exercise 8: 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

6.4 – Hypothesis Testing Calculations

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 dice are fair. Put in mathematical terms, 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. We will frequently 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. Measure how far from 0.1667 (the mean of the sampling distribution) this particular sample is, 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 *not close* to 0.1667 (the mean 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. (That is, we are using a strategy that limits the Type 1 error probability to 0.05 which is the same as 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.

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

mean = 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, 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.

Comment: 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%.

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. In this example, if we had used this strategy 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.

Remember that in general reducing the likelihood of making a Type 1 error increases the possibility of having a Type 2 error.

Exercise 9: 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 10: 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 11: **a.** For which, if any, of the tests carried out in Exercises 9 and 10 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)

6.5 – Testing Dice Using Other Proportions

Using even rolls rather than 7s

In the preceding sections, 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. In this section, we will begin by examining a slightly different probability, namely the probability of rolling an even number total on the two dice (that is, either 2, 4, 6, 8, 10, or 12). For a fair pair of dice, that probability is 50%, or 1/2. For a given pair of dice, we will therefore carry out a test using this logic:

Claim being investigated: the proportion of even rolls for this pair of dice is 50%. In symbols, we can write this as: $p = 0.5$.

Procedure:

- Roll the dice *n* time for some chosen value of *n* (the sample size)
- Find the proportion of even rolls obtained.

Decision: either discard or keep the dice, as follows:

- If the proportion of even rolls *is not* close to 50%, reject the claim and discard the dice.
- If the proportion of even rolls *is* close to 50%, acknowledge that the claim could be true and keep the dice.

Possible errors you could make using this procedure

- Discarding a pair of fair dice. This is a Type 1 error.
- Keeping a loaded pair of dice. This is a Type 2 error.

Just as in the preceding section, we want to choose our measure of closeness to control the Type 1 error rate. We begin by considering the sampling distribution for this situation[2].

The sampling distribution

As will be true for every hypothesis test strategy we cover, the starting point is to figure out what *should* happen if the claim being investigated is true. This is the so-called ***sampling distribution***.

So, what should happen when we roll a fair pair of dice *n* times and measure the sample proportion $\hat{p}$ (the proportion of even rolls) for that sample? The answer (provided *n* is large enough) is that we expect the proportion of even rolls to be close to 1/2 – probably not exactly 1/2, but pretty close. If we repeat the experiment over and over, sometimes it will be very close, sometimes not so close. This is illustrated by the applets at this link, which are similar to the ones you used in Section 6.2, except that these applets count the occurrence of even rolls rather than 7s. (As for the previous apps, the only difference between the two is that the second plots all the histograms, no matter the sample size, on the same scale.)

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

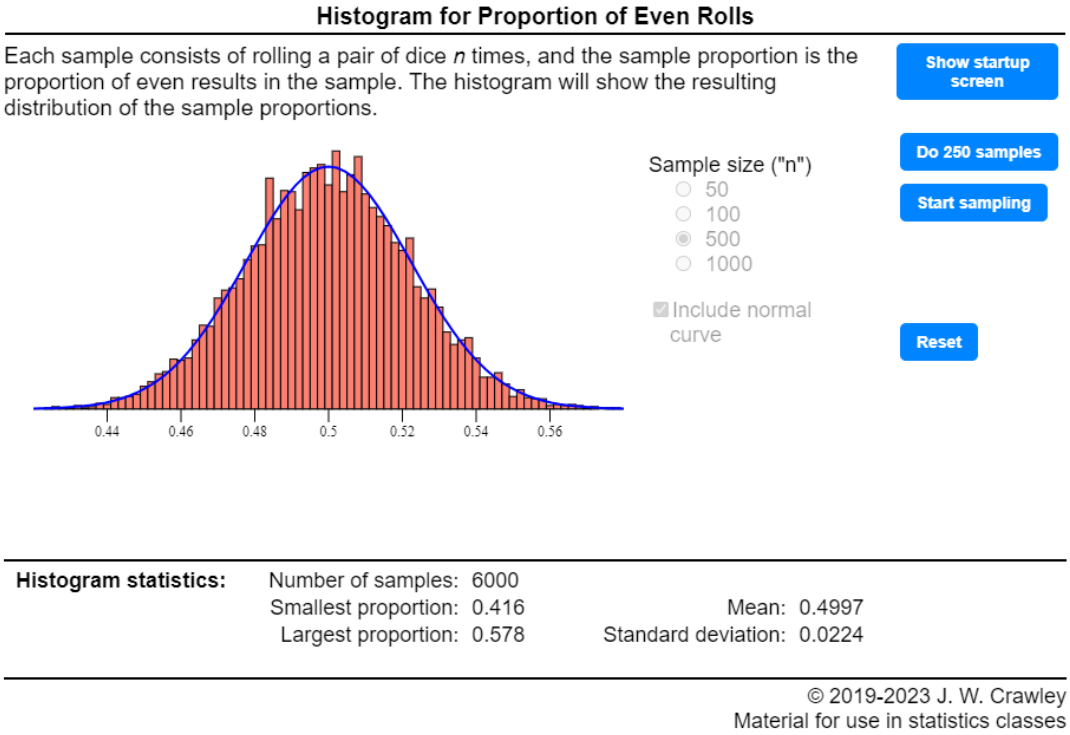
Here is a sample run of the first applet obtained by the author, using a sample size of 500 and repeating the experiment for 6000 samples.

---

[2] If you wish to do some empirical experimentation prior to moving on to the detailed explanation of the procedures, you may use the applet at this link:

Keep or Discard Dice Part 4

The applet is identical to the Part 3 applet you used in Section 6.1, except that it uses the proportion of even rolls rather than the proportion of 7s.



As was the case for sample proportions of 7s, the sampling distribution for sampling proportions of even rolls appears to be mound-shaped. And, similar to the case for the proportion of 7s, mathematicians have established three facts about this sampling distribution:

- It is approximately normal.
- The mean is the proportion in the claim, which we have labeled with the variable p. In this case, that proportion is 0.5.
- The standard error (standard deviation) can be calculated using the formula $s.e. = \sqrt{\frac{p(1-p)}{n}}$

This allows us to carry out hypothesis tests using identically the same steps as those we have already learned about.

An example

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.

mean = 0.5

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 12: 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.

Other proportions

Using these same methods, we can test dice for fairness by examining other sample proportions. For example, the probability for fair dice of rolling a 10 is 3/36 ≈ 0.0833. To test a pair of dice for fairness, we could obtain a sample proportion $\hat{p}$ of 10s by rolling the dice *n* times and counting the number of 10s. We would then compare that sample proportion to 0.0833, using the concept of P-value just as in the earlier examples.

Example. 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.

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)

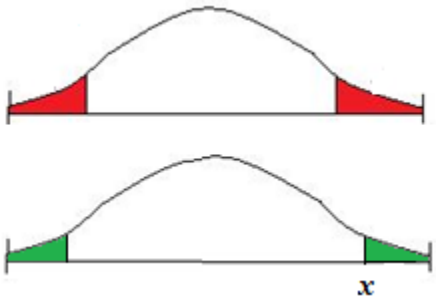
6.6 – Why Does This Decision Strategy Work?

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.

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:

Actual status of dice	Decision made: Keep	Decision made: Discard	Percent correct decision	Error rates
"Fair" dice:	304	283	For "fair" dice: 51.79%	Type 1: 48.21%
"Loaded" dice:	180	195	For "loaded" dice: 52.00%	Type 2: 48.00%

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 73.42%.)

Results:

Actual status of dice	Decision made: Keep	Decision made: Discard	Percent correct decision	Error rates
"Fair" dice:	587	0	For "fair" dice: 100.00%	Type 1: 0.00%
"Loaded" dice:	279	101	For "loaded" dice: 26.58%	Type 2: 73.42%

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:	525	38	For "fair" dice: 93.25%	Type 1: 6.75%
"Loaded" dice:	178	215	For "loaded" dice: 54.71%	Type 2: 45.29%

---

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	2.6%	3.5%
1000	2.3%	3.1%
1400	2.0%	2.6%

4. Use the applet to reproduce what the author did – that is, generate 2500 samples each with $n = 1000$. Are your results consistent with those shown above?

When the author repeated the activity, the results were fairly consistent – the histogram shape was similar, the mean was 0.1667, and the standard deviation was 0.0119. A third repetition again yielded a similar shape, with mean 0.1665 and standard deviation 0.0118.

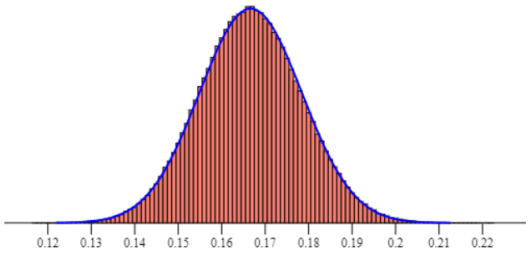
5. Use the applet to experiment. In particular, for each of the sample sizes (50, 100, 500, 1000) use the start sampling / stop sampling buttons to generate about 100,000 samples with an overload normal curve. For each sample size, record your answers to these questions:

- **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.



6. Use your results from Exercise 5 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 5 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

7. 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.

8. Find the one-tail and two-tail P-values for these *z*-scores. (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

9. 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.

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

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

Exercise 7(b), and the second pair of dice in Exercise 8.

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).

12: 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