Lesson 6
Introduction to Inferential Statistics
Confidence Intervals

Outline of the Lesson

As the tax-filing deadline approached in early April of 2011, a national polling organization conducted a survey. They asked the question, "Do you think the amount of taxes you pay is fair?" News media reported the results as, "According to a recent survey, 54% of Americans age 18 and above believe that the amount of taxes they pay is fair, with a margin of error ±4.2%." This is an example of ***inferential statistics***: the polling organization certainly did not poll every person age 18 and above in the entire country, but the results of the poll were reported as a fact about the entire country. The pollsters made an *inference* about the entire country, based on the results of the poll. This raises several questions which we will attempt to answer in this lesson. 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 "margin of error ±4.2%" mean? More generally, what is a *margin of error*?

This particular poll asked a categorical question, with two possible answers: yes, no. The result was reported as a proportion. 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 54% figure reported in the news media:

- The proportion of Americans, age 18 or above, who believe the amount of tax they pay is fair, is 54%.
- The probability that a randomly selected American, age 18 or above, would believe the amount of tax they pay is fair, is 0.54.

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. Then we apply what we learn to the polling process.

6.1 – Estimating Probabilities

When we toss a coin the probability of getting *heads* is ½, or 0.5, or 50%. This does not mean, however, that if you toss the coin twice, you must get exactly one head and one tail. It does not even mean that you will get 500 heads out of 1000 tosses. What it does mean 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.*

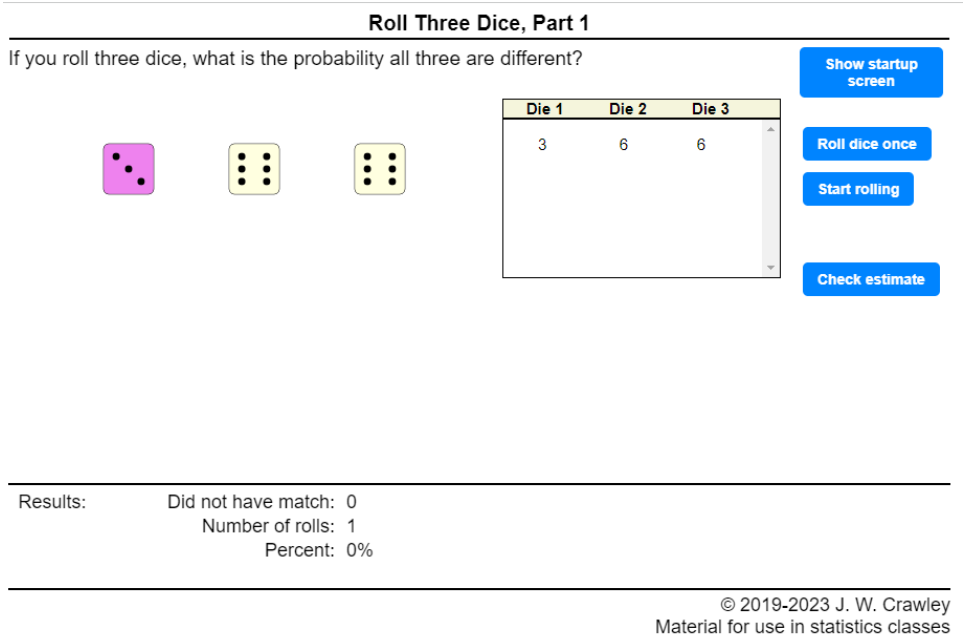
Here is an important consequence of this long-term behavior: *if we were unable to calculate that the theoretical probability is 0.5, we could use the long-term proportion to estimate that probability*. The more times we toss the coin, the closer we can expect our proportion to be to the actual probability of 0.5. This fact is crucial to understanding how opinion polls work, so we examine it in more detail in the next section.

An experiment with three dice

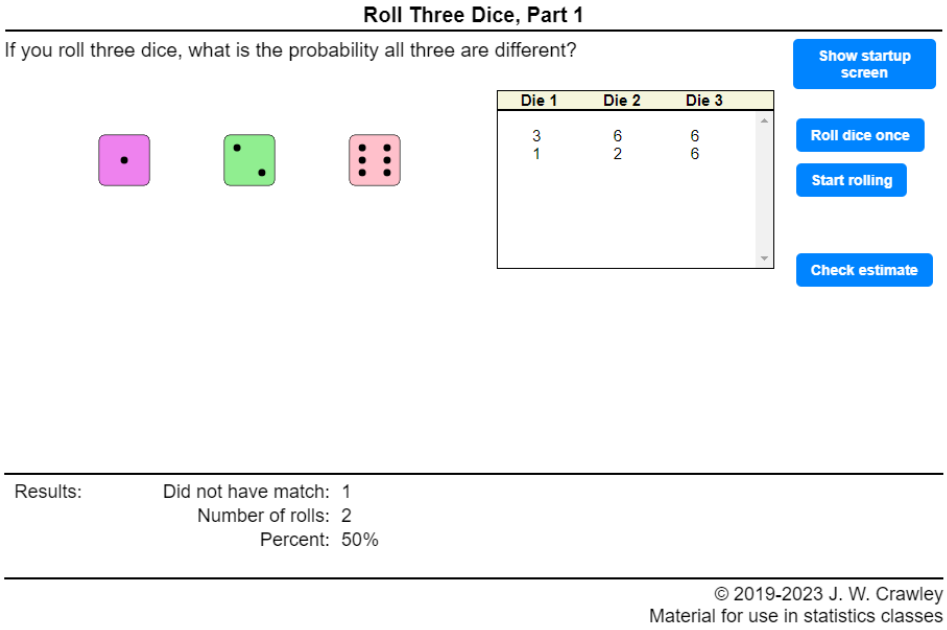
For our next example, consider the following: roll three dice, and determine the probability that all three are different. Although it is not impossible to calculate the theoretical probability, this calculation is not the subject matter of this course. Our approach is to estimate the probability by experimentation. We can do this one of two ways: roll some actual physical dice, or use the applet "Roll Three Dice, Part 1" which is supplied with this material. Your instructor may ask you to do both – use actual dice, and then use the applet at this link:

> Roll 3 dice, part 1

Here is what you should do. Roll the dice approximately 50 times, keeping track of two things: how many times you rolled the dice, and how many times all three were different. For example, here is what happened when the author used the applet to roll the dice once:



This roll happens to have had a match – when that happens the applet colors both dice light yellow to highlight the match. Look at the results near the bottom of the screen. Out of 1 roll so far, there have been 0 times that did not have a match, for a percent of 0 / 1 = 0%. Now here is a second roll.



This time all three dice were different (no match). So far, this has happened 1 time in 2 rolls, or 50% of the time. If we were to stop now, we would estimate the probability of it happening as 50%. But of course no one would expect accurate estimates based on only two trials. So we continue until we have about 50 rolls.

Note: You can either press the "Roll dice once" repeatedly, or use the "Start rolling" and "Stop rolling" to carry out the experiment more quickly.

Here are the results obtained by the author:

Roll Three Dice, Part 1

If you roll three dice, what is the probability all three are different?

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

Show startup screen

Roll dice once

Start rolling

Check estimate

Results: Did not have match: 24
Number of rolls: 50
Percent: 48%

In this experiment, for 24 of the 50 rolls all three dice were different, for a proportion of 24/50 = 0.48 = 48%. Based on this experiment, we would estimate the probability that all three dice will be different as 0.48 or 48%.

When we click the "Check estimate" button in the applet, the applet lets us know how good the estimate was. If the estimate was "close to" the actual probability, we get a message printed in green, otherwise a message printed in red. Any estimate within 3% of the actual probability is considered close in this applet. For our estimate of 48%, here is the result:

Roll Three Dice, Part 1

If you roll three dice, what is the probability all three are different?

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

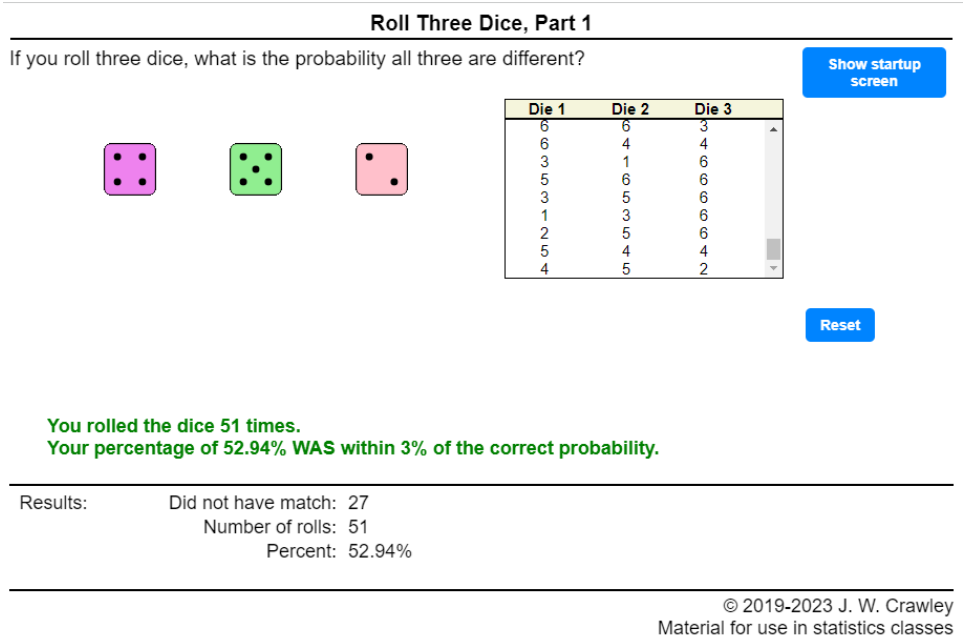
Show startup screen

Reset

You rolled the dice 50 times.
Your percentage of 48% was NOT within 3% of the correct probability.

Results: Did not have match: 24
Number of rolls: 50
Percent: 48%

When we roll the dice 50 times, as you can see, our estimate may not be very close to the actual probability. However, sometimes it is; here is another experiment using the applet again:



Exercise 1: Use the applet several times, each time rolling the dice about 50 times. Record your results here: the estimate and the answer to the question, "Was the estimate close – within 3% of the actual probability?" The first two are filled in based on the results shown above.

Percent (estimate)	Close?
48.00%	no
52.94%	yes

Margin of error (closeness, correctness, and interval estimates)

In the original example about taxes, the polling organization reported the population proportion as 54%, "with a margin of error $\pm 4.2\%$." This raises the question:

> What exactly does the phrase "margin of error ±4.2%" mean? More generally, what is a *margin of error*?

We are now in a position to answer this question, by thinking about what happened with the dice.

Before we talk about the dice, however, think again about tossing a coin. If you toss a coin 1000 times, do you expect to get exactly 500 heads? Not really – but you do expect that the percentage of heads will be close to 50%. This illustrates the following important point: *When we obtain an estimate of a probability by experimentation, we do not expect the estimate to be absolutely correct*.

Because of this fact, estimates obtained by experimentation are often given as ***interval estimates*** rather than just ***point estimates*****.** Let's explain the terminology just a bit. Using the applet, our first estimate for the probability was 62.00%. This is called a *point estimate*, because it consists of a single point on the number line. Now let's use the point estimate to create an interval estimate, as follows:

> For the point estimate of 48.00%, we are really saying, "We think the actual probability is close to 48.00%. By 'close' we mean within 3%. That is, we think the probability is between 45.00% and 51.00%.

An interval on the number line consists of numbers between two given numbers. The phrase, "between 45.00% and 51.00%" describes an interval, so it is called an *interval estimate*. This estimate is sometimes written in mathematical interval notation: (45.00%, 51.00%).

What does this have to do with "margin of error"? The answer is simple – when statisticians use the terminology margin of error, they are describing what they mean by "close." Here are three different ways to describe the same estimate:

- 48.00%, with a margin of error of ±3%.
- Between 45.00% and 51.00%.
- In the interval (45.00%, 51.00%).

To calculate the interval estimate, we subtract the margin of error from the point estimate, and add that same margin of error to the point estimate.

Now to a crucial question – should the estimate earn a message printed in green, or one printed in red? There are two equivalent ways to answer this question. To earn a green message, the point estimate must be close (within the margin of error ±3%). This is the same as saying that the interval estimate must be correct, it must contain the actual probability.

The first point estimate, 48.00%, earned a red message. The actual probability *is not* close to (within ±3% of) this estimate. The actual probability *is not* in the interval between 45.00% and 51.00%.

On the other hand, the second point estimate, 52.94%, earned a green message. The actual probability *is* close to (within ±3% of) this estimate. The actual probability *is* in the interval between 49.94% and 55.94%.

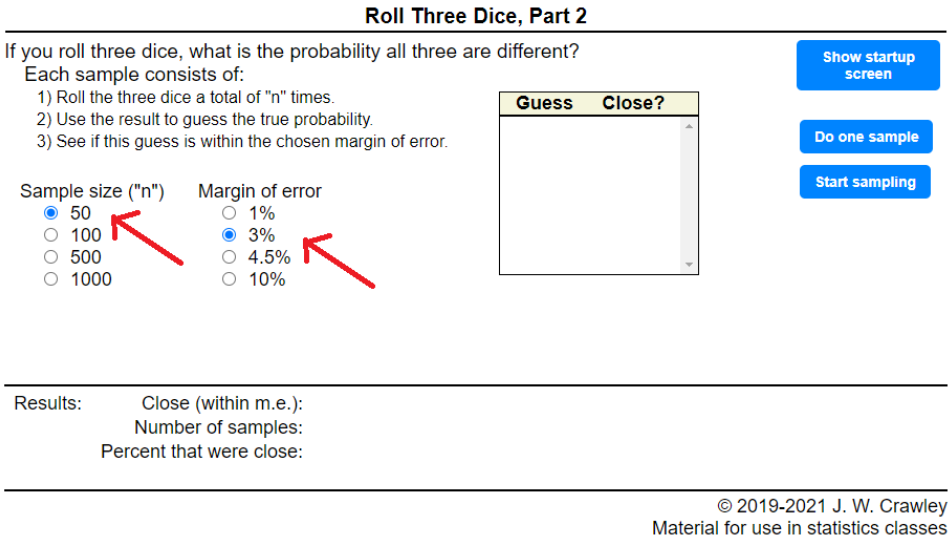
> *You have heard it said that close only counts in horseshoes. However, close also counts in estimating probabilities using experimentation (long-term proportions). An estimate is considered correct if the corresponding interval estimate contains the actual probability. This means that the point estimate is close (within the margin of error).*
>
> *When the interval estimate is correct, instead of saying that the point estimate is close, we sometimes say that the point estimate is "correct within the margin of error." But that's just a fancy way of saying that it is close.*

Confidence level (do you think your estimate is correct??)

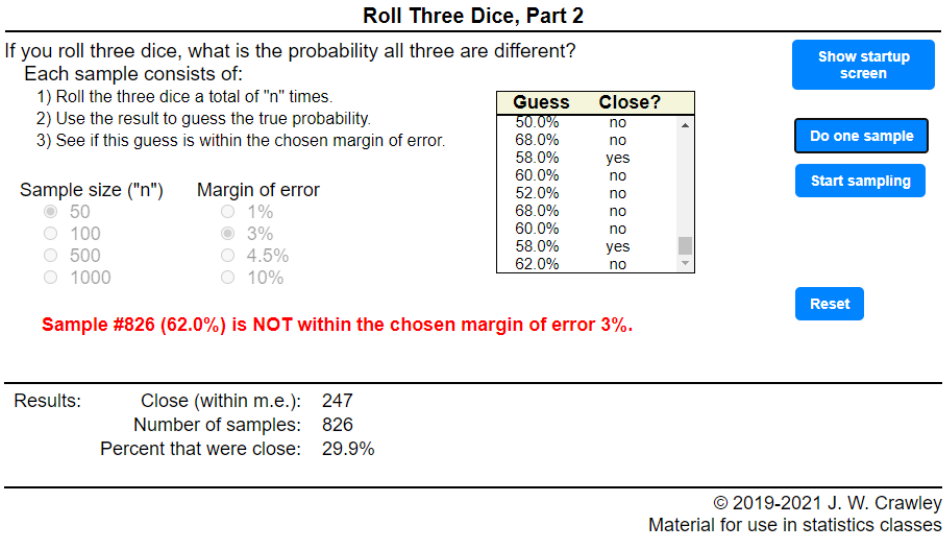
When we roll the dice 50 times, sometimes the point estimate is close – that is, the interval estimate is correct – and sometimes it is not. If you were asked to perform the experiment one more time, again rolling the dice 50 times, how confident would you be that the resulting point estimate would be close? To get a preliminary idea, you might look back at the table above where you recorded your own results. Are you almost certain to be correct? Is it more likely correct than not? Is it about a 50-50 proposition? To get a better idea, you might decide to use the applet several more times. To make that easier for you, we have created a second applet that automatically runs the first applet multiple times.

Roll 3 dice, part 2

As you can see highlighted with the arrows in the screen shot below, this applet is initially set up to repeat exactly the experiment you did – 50 rolls of the dice, and a margin of error of 3%. But you can change those numbers if you wish.



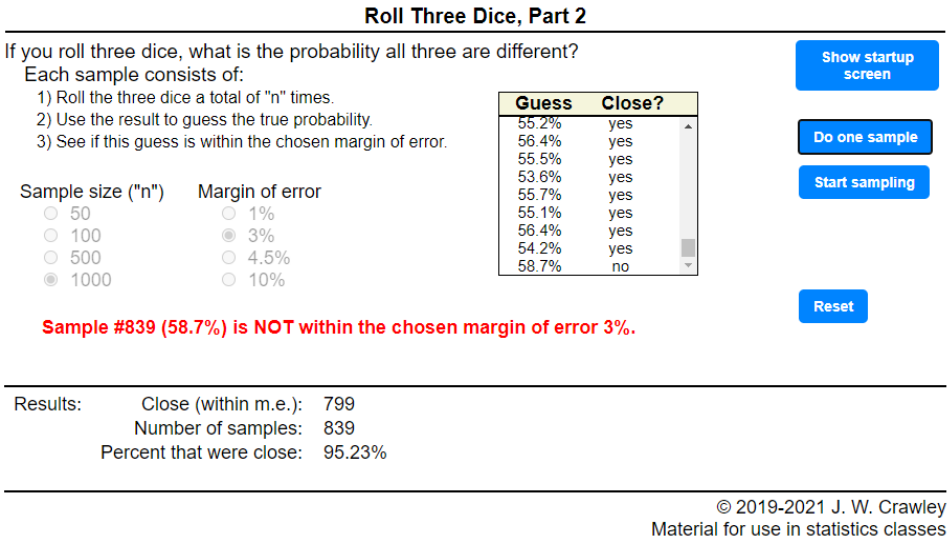
When we ran the applet with these values, here is what we saw after a large number of samples (we did 826 samples, where each sample consisted of 50 rolls of the dice):



It happened that our final sample's proportion (62.0%) was not close, it was not within the margin of error. Put another way, our final interval estimate of (59.0%, 65.0%) was incorrect. In fact, only about 30% of the samples yielded results that were close. In answer to the question, "How confident are you that the next sample will yield a correct answer?" we would say, "Not very confident – it looks like only about 30% of the samples yield correct answers."

Exercise 2. Use the applet to run between 800 and 1000 samples. Record the results here. Are your results fairly consistent with ours?

Based on these results, it looks like rolling the dice 50 times, and using a margin of error of 3%, isn't a very good strategy, if we would like to get a correct answer most of the time. The proportion obtained in 50 rolls is not very likely to be within 3% of the actual probability. If we think about it, this really shouldn't come as a shock. We might expect to have to roll the dice more than 50 times to begin to get an accurate estimate. The applet allows us to do this. For example, here is what we obtained when we increased the sample size from 50 rolls to 1000 rolls, and ran 839 samples:



It happened that our final sample was not close (correct within the margin of error), but over 95% of the samples were. If we did one more sample, we could be pretty confident that it would be correct. Our ***confidence level*** would be about 95%. We are 95% confident of getting a correct answer because it appears that about 95% of the samples do give a correct answer.

Note: Just as each sample yields only an approximate indication of the probability for the dice, this process yields only an approximate indication of how confident we should be (the confidence level). But since we did a large number of samples, namely 839, it is a pretty good indication.

Exercise 3. Use the applet to run between 800 and 1000 samples, with the sample size set to 1000 and the margin of error still 3%. Record the results here. Are your results fairly consistent with ours?

There is obviously a connection between sample size and confidence level. This should not come as a surprise. Thinking back to the coin-tossing thought experiment, our own experience suggests that when you do just a few tosses results can be more random, but when you do lots of tosses the proportion is more likely to be close to the 50% theoretical probability. The more we toss the coin, the more confident we are that our proportion will be close to the actual probability. Paraphrased, "increasing the sample size increases the confidence level."

There is also obviously a connection between margin of error and confidence level. The margin of error indicates how close the estimate has to be to be considered correct. If our margin of error is reduced from 3% to 1%, we are less confident that our sample will satisfy the closeness criteria. Conversely, increasing the margin of error to 10% raises our confidence level. In short, "increasing the margin of error increases the confidence level."

Exercise 4: Use the applet several times, experimenting with changes in the sample size and the margin of error. Each time do at least 800 samples. Record your results here. The first two are filled in based on the results shown above.

n	m.e.	# samples	% correct (approximate confidence level)
50	3%	826	29.9%
1000	3%	839	95.23%

Do the results support the discussion about the relationships among sample size, margin of error, and confidence level?

Terminology: We have been discussing interval estimates, and the confidence level associated with those estimates. Because there is always an associated confidence level, these interval estimates are generally called **confidence intervals.**

6.2 – Estimating Population Proportions

We are now ready to apply what we have learned to the process of sampling from a large population. Although there are a few subtle differences, with proper sampling methods these differences are very minor and do not affect our final results.

The key to our discussion is the connection between proportion and probability. For example, consider the question which began this lesson, "Do you think the amount of taxes you pay is fair?" For that question, the population the pollsters were interested in consisted of all Americans age 18 and above. The pollsters wanted to estimate the proportion of that population that would answer "yes" to that question. The basis for what they did is the fact that these two numbers are the same:

- The proportion of Americans, age 18 or above, who would answer "yes."
- The probability that a randomly selected American, age 18 or above, would answer "yes."

As a result, the pollsters need to estimate a probability, just as we estimated a probability for the roll of three dice.

How should they proceed? Just as we did for the dice, the method consists of estimating the probability experimentally. To get an estimate for the dice-rolling problem, we repeated the following process 50 times: roll three dice and record the result. To get an estimate for the tax question, the pollsters could repeat this process 50 times: select an American age 18 or above, ask the question, and record the answer. 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 survey to have any meaning, the selection must be random.

To illustrate the similarity to the dice rolling situation, we have two applets which are quite similar to the two applets you used to investigate dice rolling. In the next two subsections we will discuss these applets and analyze what they show us about sampling.

Notation and terminology: When pollsters sample from a large population, as described above, they are trying to estimate a proportion for the entire population (the ***population proportion***). They measure the corresponding proportion for their sample (the ***sample proportion***). Statisticians frequently use the letter p to stand for a population proportion, and the variable $\hat{p}$ (pronounced as "p-hat") to stand for a sample proportion. To help you get used to this notation, we will employ it in the remainder of our discussion.

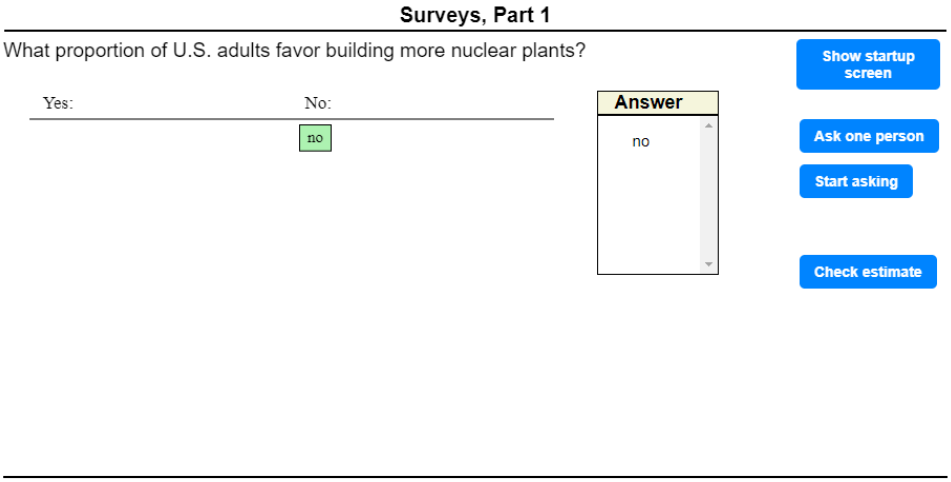
Surveys, Part 1 – taking a sample to estimate a population proportion *p*

The applet at the following link deals with asking the question, "Should additional nuclear plants be built?"

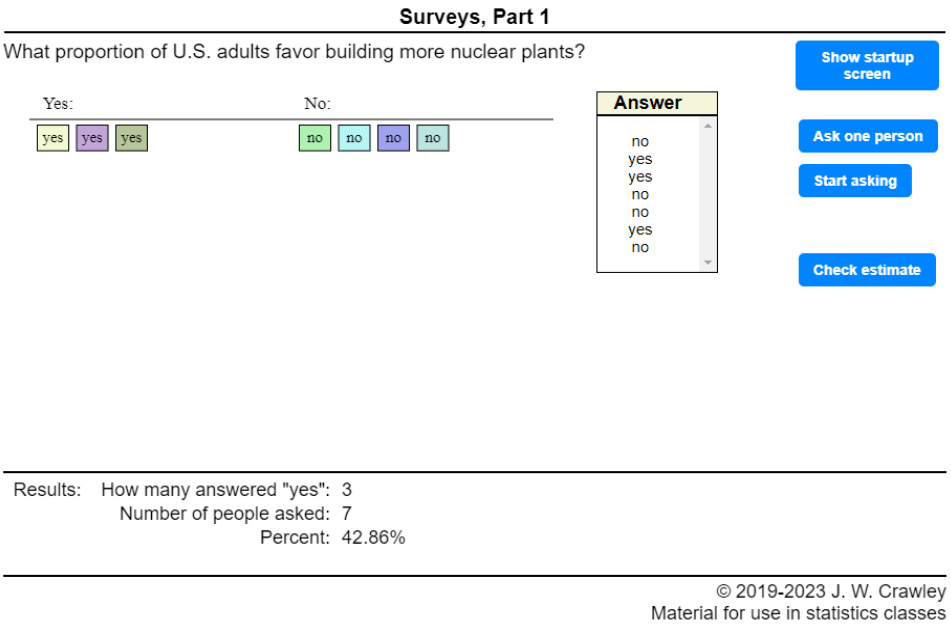
Surveys, part 1

Each time the *Ask one person* button is pressed, the person's answer is shown. You can use that button repeatedly, or the *Start asking* and *Stop asking* buttons, to obtain a sample of whatever size you wish. The following screen shots show the author's results after 1, 7, and 50 persons have been asked.

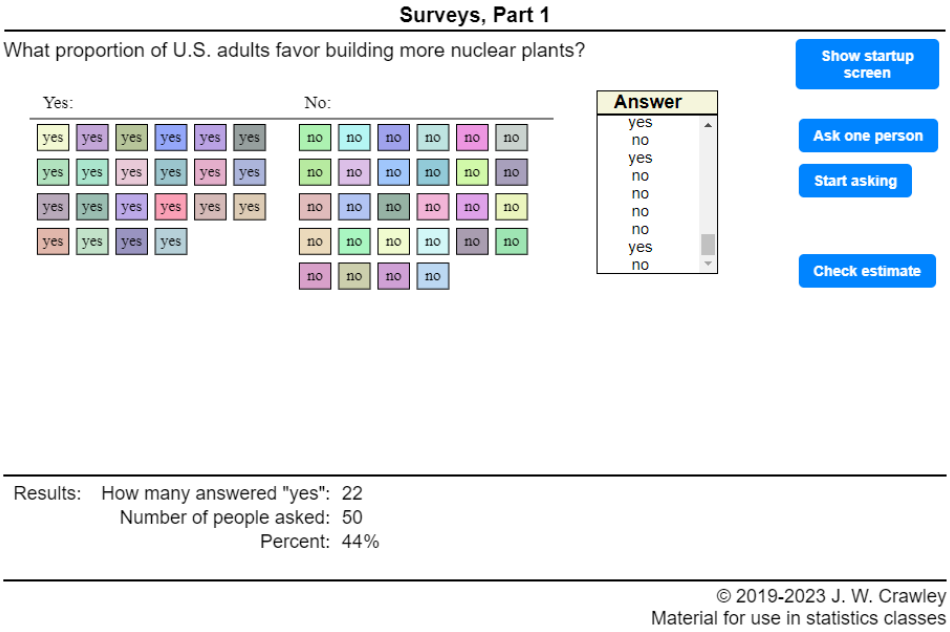
After 1:



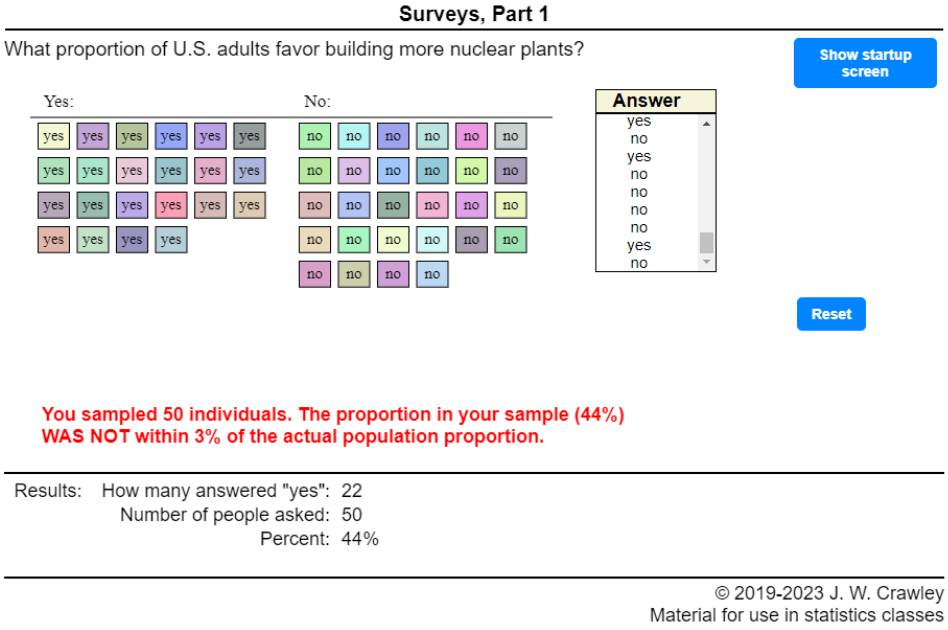
After 7:



After 50:



After 50 people have been asked, 22 of those 50 people (44.00%) have answered "yes"; the sample proportion $\hat{p}$ is 44.00%. We could stop now, and use that percentage to estimate the proportion p of the ***entire population*** that would answer yes. If we do, the estimate might be close or it might not. Just as in the dice applet, when we choose the *Check estimate* button, this applet will tell us whether or not the estimate is close, as illustrated below.



Here is the result for another run of the applet, which again asks 50 randomly chosen individuals:

You sampled 50 individuals. The proportion in your sample (40%)
WAS within 3% of the actual population proportion.

Some observations are in order:

- Just as happened for rolling dice, sometimes the estimate $\hat{p}$ is close to the true proportion p and sometimes it is not.
- This applet, like the first dice-rolling applet, uses a margin of error of ±3% as its measure of closeness.
- The point estimates (sample proportions, $\hat{p}$ values) from the two different surveys are 44.00% for the first, 40.00% for the second.
- The interval estimates (found by adding and subtracting the margin of error from $\hat{p}$) are (41.00%, 47.00%) for the first, and (37.00%, 43.00%) for the second.
- The first interval estimate is incorrect – it does *not* contain the actual probability. The second interval estimate is correct.
- Remember that these interval estimates are called *confidence intervals*, because there is always a confidence level lurking in the background – more on this later.
- Remember that the actual probability is the same as the population proportion p, that is, *the proportion in the entire population that would answer "yes" to the question*.

Exercise 5: Use the applet several times, each time sampling about 50 people. Record your results here: the estimate and the answer to the question, "Was the estimate close – within 3% of the actual population proportion p?"

Percent ($\hat{p}$)	**Close?**

Exercise 6: Use the applet a few more times, with larger samples. Record your results here. Does having a large sample ***guarantee*** that the estimate is close?

Sample size	**Percent ($\hat{p}$)**	**Close?**

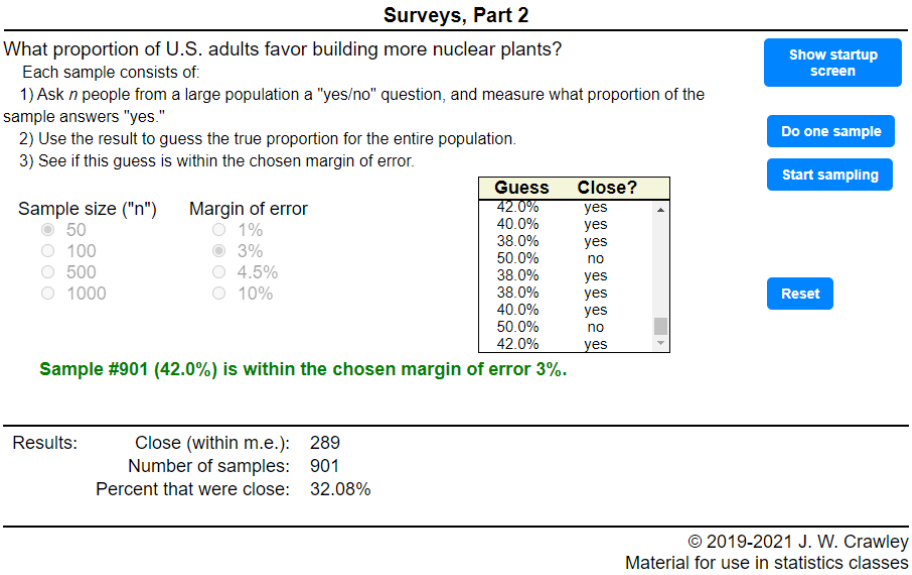
Surveys, Part 2 – confidence level

To get a handle on the relationship between sample size and confidence level (the likelihood that the estimate will be close), you could repeat the previous process indefinitely. However, just as we did

for dice rolling, we provide a second applet which carries out that process a large number of times automatically.

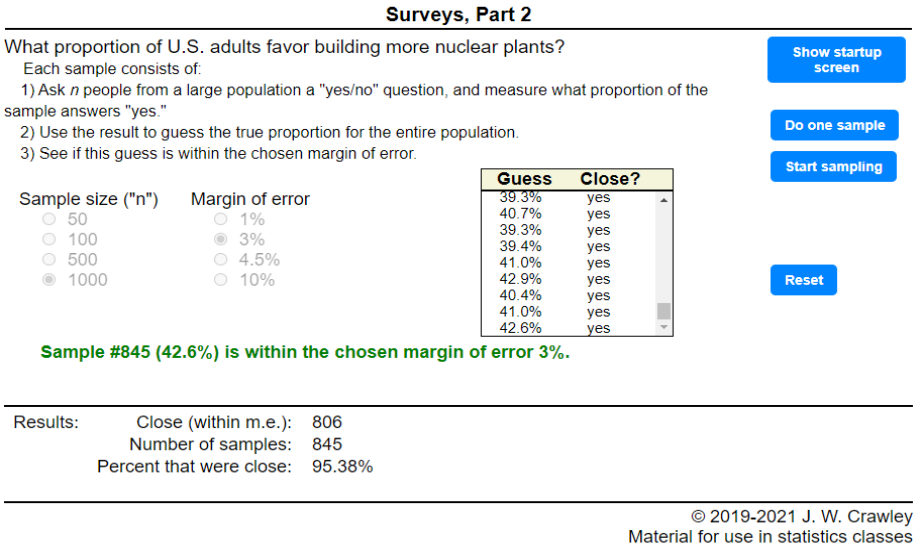
Surveys, part 2

Just as in the dice applet, the default sample size and margin of error are 50 and ±3%, respectively, but you can change those values. The goal is the same – take samples of a given size, and see if the estimate they provide – that is, the sample proportion $\hat{p}$ – is close to the actual proportion p. As for the dice, we let the applet do hundreds of samples (minimum of 800) in order to get a good feel for how confident we should be that we the sample yields a close estimate. Here is the result for one run; out of 901 samples, 289 – that is, 32.08% of them – were close (within the chosen margin of error).



Exercise 7. Use the applet to run between 800 and 1000 samples, with n = 50 and m.e. = 3%. Record the results here. Are your results fairly consistent with ours?

Just as for rolling dice, it appears that using sample size 50 and margin of error 3%, isn't a very good strategy, if we would like to get a correct answer most of the time. The confidence level seems to be well below 50%. Just as for the dice, we can improve the results by using a larger sample. Here is an example:



Surveys, Part 2

What proportion of U.S. adults favor building more nuclear plants?

Each sample consists of:

1) Ask *n* people from a large population a "yes/no" question, and measure what proportion of the sample answers "yes."

2) Use the result to guess the true proportion for the entire population.

3) See if this guess is within the chosen margin of error.

Show startup screen

Do one sample

Start sampling

Reset

Sample size ("n")	Margin of error
○ 50	○ 1%
○ 100	◉ 3%
○ 500	○ 4.5%
◉ 1000	○ 10%

Guess	Close?
39.3%	yes
40.7%	yes
39.3%	yes
39.4%	yes
41.0%	yes
42.9%	yes
40.4%	yes
41.0%	yes
42.6%	yes

Sample #845 (42.6%) is within the chosen margin of error 3%.

Results:		
	Close (within m.e.):	806
	Number of samples:	845
	Percent that were close:	95.38%

Exercise 8. Use the applet to run at least 800 samples, with the sample size set to 1000 and the margin of error still 3%. Record the results here. Are your results fairly consistent with ours?

Just as for rolling dice, increasing the sample size increases the confidence level. And, just as for rolling dice, increasing the margin of error increases the confidence level. And, just as happened for the die, we obtain an apparent confidence level close to 95% by using $n = 1000$ and m.e. = 3%.

Exercise 9: Use the applet several times, experimenting with changes in the sample size and the margin of error. Each time do between 800 and 1000 samples. Record your results here. The first two are filled in based on the results shown above. *Use n = 1000 and m.e. = 3% for several of the runs.*

n	m.e.	# samples	% correct (approximate confidence level)
50	3%	901	32.08%
1000	3%	845	95.38%

Do the results support the discussion about the relationships among sample size, margin of error, and confidence level? Do the results seem to indicate that *n* = 1000 and m.e. = 3% gives a confidence level close to 95%?

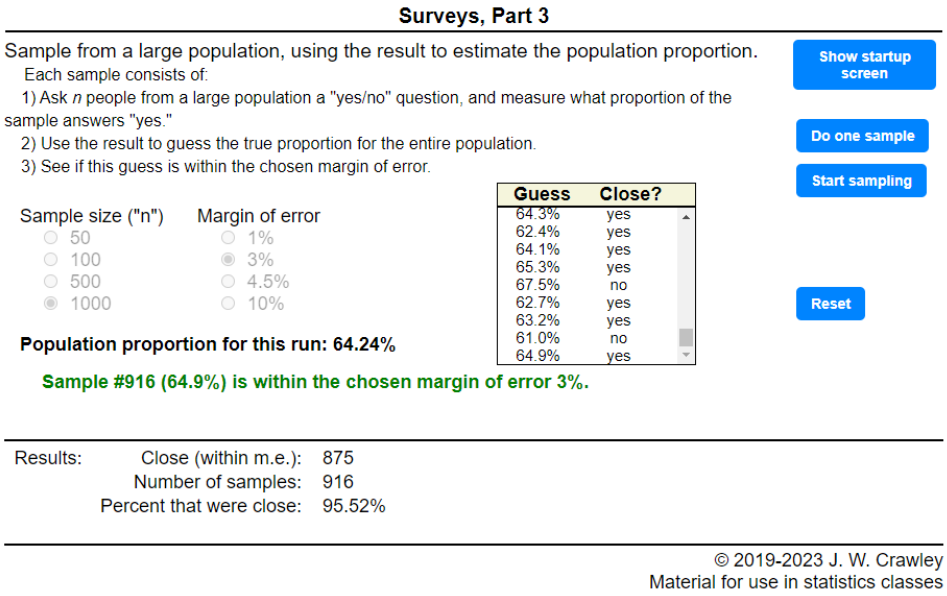
Surveys, Part 3 – do the results depend on the population proportion?

Both survey applets are based on a question for which the population proportion was $p = 0.40$. For this particular population proportion, the table in Exercise 9 provides evidence that using *n* = 1000 and *m.e.* = 3% yields a confidence level of approximately 95%. For example, the experiment reported in the second row reports that 95.4% of the 969 samples yielded a correct answer. If you have used those same values for *n* and *m.e.*, you have obtained similar results.

An important question is this: Would the same thing happen for questions where the population proportion was something other than 40%? To help you see the answer to this question, we provide another version of the surveys part 2 applet; here is the link.

Surveys, part 3

This version uses a different population proportion every time the *Reset* button is chosen. Here is a sample run, with *n* = 1000 and *m.e.* = 3%.



As you can see, in this run the population proportion was $p = 64.24\%$. For this value of p, once again a strategy of using samples of size 1000 with a 3% margin of error seems to suggest a confidence level of approximately 95%.

Exercise 10: Use the applet several times, each time with $n = 1000$ and *m.e.* = 3%. Each time do between 800 and 1000 samples. Record your results here. The first row is filled in based on the results shown above.

Population proportion	**# samples**	**% correct (approximate confidence level)**
64.24%	916	95.52%

Do the results seem to indicate that a strategy of using samples of size 1000 with a 3% margin of error seems to suggest a confidence level of approximately 95%, no matter what the population proportion might have been?

Summary – back to the original example

We began this discussion with an example reported in the news media as, "According to a recent survey, 54% of Americans age 18 and above believe that the amount of taxes they pay is fair, with a margin of error ±4.2%." As we pointed out, the survey did *not* obtain its results by surveying everyone in the country – it only surveyed a *sample* of Americans age 18 and above. We are now in a position to explain exactly what the news media report means, and to answer the questions posed in connection with the example.

1. How is it possible to draw conclusions about a group that is larger than the group you actually questioned?

As we have demonstrating by using the dice applets and the survey applets, there is a certain predictability in carrying out random samples, whether the sampling consists of rolling three dice or of asking people questions. Sometimes the estimates we obtain using $\hat{p}$ are close to the actual value p, and sometimes they are not. With a sample of size 50, they are more often incorrect (not close) than correct (close). However, with a sample of size 1000, the $\hat{p}$ estimates tend to be pretty consistently close to the actual population proportion p.

2. Is this process legitimate? If so, what precautions must we take in interpreting the result?

Yes, the process is legitimate, provided we keep a few things in mind. First and most important, the sampling must be *random*. In the case of the dice, the randomness is implicit in the action of throwing the dice. But if the dice are weighted so that 6 is more likely to come up than any other number, this will destroy the randomness and make our estimates unreliable. In the case of surveying people, it can be very difficult to obtain true randomness. For example, some people refuse to answer surveys, and some people don't have phones – just two of the many obstacles to obtaining a truly random sample. Ethical professional pollsters utilize a variety of methods for reducing the bad effects of these difficulties, 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.)

Even assuming a perfectly random sampling methodology, with the sample taken from the same population we want to report on, we must take care to hear the results correctly. People in general tend to focus on the "54% of Americans age 18 and above believe…" part of the statement, and to ignore the margin of error because they don't know what it means. But there is always a margin of error, even if the news media doesn't report it, and there is also always a confidence level. To understand what we are hearing, we must understand these issues. We will address this further as we answer the next question.

3. What exactly does the phrase "margin of error ±4.2%" mean? More generally, what is a *margin of error*?

We now know that the margin of error can be viewed as a measure of closeness, or as a way of turning a point estimate into an interval estimate – two ways of saying the same thing. To say "54%

of Americans age 18 and above believe that the amount of taxes they pay is fair, with a margin of error ±4.2%" means a lot more than what is stated. Specifically, it means all the following:

- 54% *of the people surveyed* answered yes ($\hat{p}$ is 54%).
- Although I didn't ask everyone in the entire country, I'm using the result from my survey to *estimate* that for the entire country 54% would also answer yes (that is, that p is 54%).
- I *think* the actual value is close to this estimate – no further away than ±4.2%. Put another way, I *think* the actual value is somewhere between 49.8% and 58.2%.

There is even more to the story, although the news media almost never reports this part. There is the confidence level. For this particular example, if you go back to the original report you find a confidence level of 95% reported, although that didn't make it into the news media's version. So there is also this:

- When I say I *think* the actual value is somewhere between 49.8% and 58.2%, I am 95% confident that my result is correct. This is because I used a method that gives a correct answer 95% of the time.
- Since the method gives a correct answer 95% of the time, this means that 5% of the time the method yields an incorrect answer. So 5% of the time the method yields an estimate that is *not* within ±4.2% of the actual population proportion.

To the extent that the sampling method was indeed random, and provided we correctly interpret the results, the methodology is sound.

6.3 – One Final Question: Why does it work?

From running the applets, we have gathered quite a bit of evidence that the following statement is true:

> *No matter what the population proportion p is, when we take samples of about size 1000 and use ±3% as the margin of error, the resulting interval estimate is correct about 95% of the time. This means that the confidence level is about 95% for this combination of sample size and margin of error.*

To think about why this is true, we begin by revisiting the coin-tossing mind experiment. If you toss a coin 1000 times, you expect the proportion of heads to be close to 50% – maybe not exactly 50%, but pretty close. If you 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 same is true for sampling from a large population. If the population proportion is p, and we sample 1000 randomly chosen individuals, we expect the proportion in the sample to be close to p – maybe not exactly p, but pretty close. If we repeat the random sampling process 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.

Reminder: The proportion we measure in the sample is called p-hat, written $\hat{p}$. The variable p, without the "hat," indicates the proportion in the *entire population*.

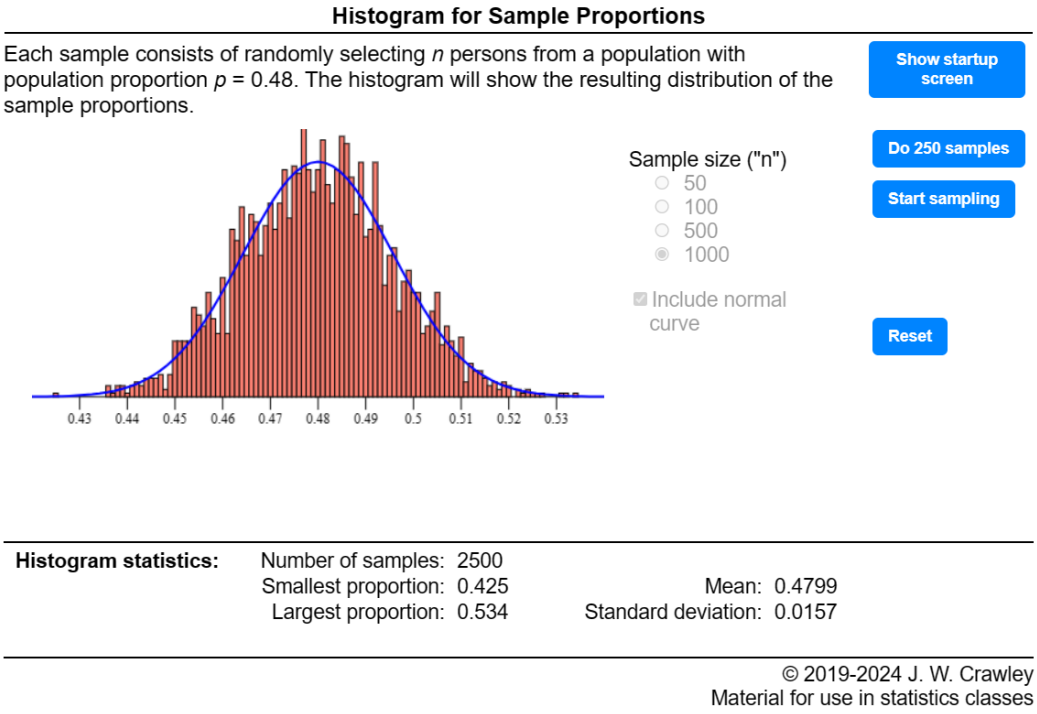
The link below runs an applet that illustrates what will happen if you repeatedly take samples from a population with population proportion p and record the sample proportion $\hat{p}$. In the applet you can select a sample size n, with the default being 1000. When you click on the "Do 250 samples" button the applet will take 250 samples of size n. For each sample it will calculate the sample proportion, $\hat{p}$, and it will

make a histogram of all these $\hat{p}$ values. You can use the controls to add more samples to the graph, or to start the whole process over. If you wish, you can use the check box to overlay a normal curve on the resulting histogram. (You can also change the sample size to 50, 100, or 500 while starting over, but since our discussion is about samples of size 1000 we will not do so in this lesson. What happens for different sample sizes will be discussed in the next lesson.)

Histogram of p-hat values

NOTE: When you click on the link, the app will randomly choose a population proportion for that particular run. For the author's output shown below, the choice was 0.48, but for you it will be different. It will also be different every time you restart the app by clicking on the link. Running the app multiple times will allow you to see that the results are similar, no matter what the population proportion might be.

Here is the result of a run by the author, in which he generated 2500 samples from a population whose population proportion p is 0.48, with a normal curve overlaid. (He simply pressed the "Do 250 samples" button 10 times.)



We will use this applet for further exploration in this lesson and the next. For now, we simply call your attention to two characteristics of the histogram. First, it is what we have called *mound shaped*. Second, the mean for the histogram (0.4799) very closely matches the population proportion (0.48).

Exercise 11. a. Use the applet to reproduce what the author did – that is, generate 2500 samples each with $n = 1000$ and with a normal curve overlaid. Because the app chooses a different population proportion p each time it is run, your results will not be identical, but they should be similar to those above. Is the resulting histogram mound shaped? Is the mean for the histogram close to the population proportion for your run of the applet?

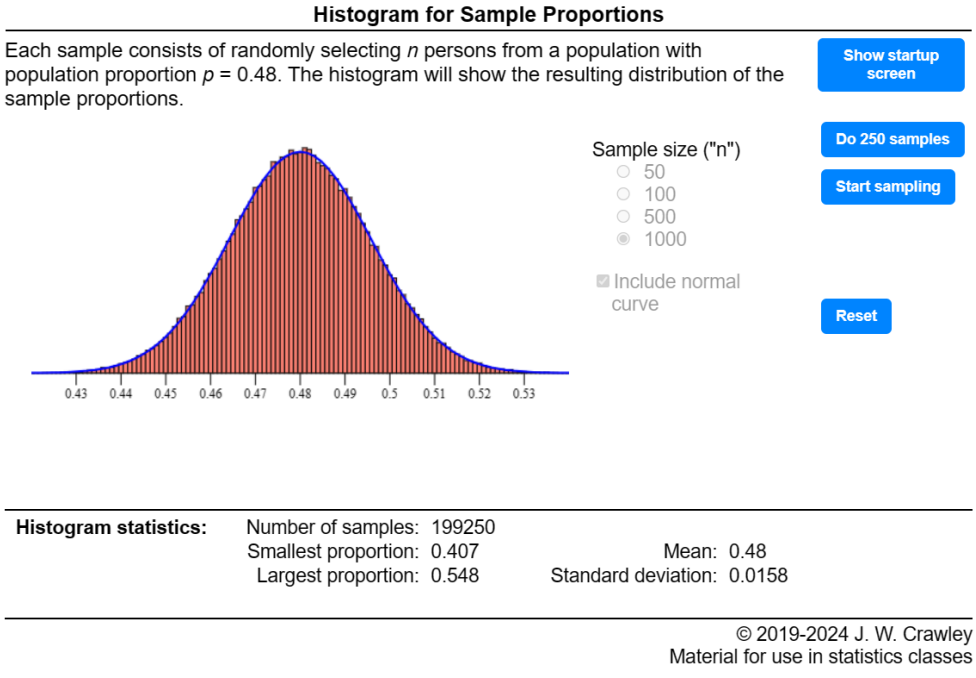
b. Reload the applet and repeat part (a) for a different population proportion.

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.** If it is clear that we are talking about sample proportions, we will simply refer to the **sampling distribution**.

Mathematicians have established three fundamental facts about this theoretical sampling distribution (the sampling distribution of the sample proportions):

- The mean of the distribution is equal to the population proportion – what we have labeled with the variable p.
- The standard deviation of the distribution can be calculated using the formula $\sqrt{\frac{p(1-p)}{n}}$
- The distribution is indeed mound shaped. (In fact, provided n is large enough, the theoretical sampling distribution is not just mound-shaped, it is in fact approximately normal; we will say more about this in the next lesson.)

The author decided to explore this a bit further. Because the sampling distribution refers to *all* possible samples, he decided not to stop at 2500 samples, but rather to use the *Start sampling* and *Stop sampling* buttons to generate approximately 200,000 samples. Here are the results.



Notes:

1. The resulting histogram is definitely mound-shaped, and matches the overlaid normal curve quite closely. The mean of the histograms (which has been rounded to four places) exactly matches the population proportion (0.48). As noted above, the sampling distribution has standard deviation $\sqrt{\frac{p(1-p)}{n}}$, which for $p = 0.48$ is $\sqrt{\frac{0.48(1-0.48)}{1000}} = 0.0157987341$; when rounded to four places this exactly matches the standard deviation for the histogram. That is, what we have observed by generating 199,250 samples is in close agreement with what mathematicians have proved to be true for the theoretical sampling distribution of the sample proportions.

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***. To help you get used to this terminology, we will follow that custom in what follows, and we will frequently use the abbreviation *se* or sometimes *SE* to stand for standard error. Whenever you see the term "standard error" or its abbreviation *se* or *SE*, you should remind yourself that it is nothing more than a standard deviation.

Exercise 12. Use the applet to experiment, as follows. Load the applet and record the population proportion (p) it will be using. Then use the start sampling / stop sampling buttons to generate about 200,000 samples with $n = 1000$ and with overlaid normal curve, and record your answers to these questions:

- **a.** Is the histogram mound-shaped? Does it closely match the overlaid normal curve?
- **b.** What is the mean for the histogram? Is this close to the population proportion p?
- **c.** Use the $\sqrt{\frac{p(1-p)}{n}}$ formula to calculate the standard error *se* (that is, the standard deviation for the sampling distribution of the sample proportions).
- **d.** What is the standard deviation for the histogram? Is this close to what you calculated in step (c)?

We now address our original question in this section: Why does it work? Why does using 1000 as the sample size and 3% for the margin of error result in a confidence level of approximately 95%? We will consider one example that illustrates the general pattern. In the next lesson we will examine the situation more thoroughly and more precisely; in particular, in that lesson we will zero in on what exact margin of error should be used to achieve a 95% confidence level in general.

Example. Suppose we take a sample of size $n = 1000$ from a population whose population proportion is $p = 0.40$. Explain why the strategy of using 3% as the margin of error results in a confidence level of approximately 95%.

Solution. Remember that saying the confidence level is about 95% means that this strategy yields a correct answer about 95% of the time. So we can rephrase the question as, "Explain why the strategy of using 3% as the margin of error yields a correct answer about 95% of the time." The key to understanding why this is so lies in the sampling distribution of the sample proportions for this population.

- First of all, the samples that yield a correct answer are those for which $\hat{p}$ is between 0.37 and 0.43. For example, if $\hat{p}$ is 0.38 the corresponding interval will be (0.38 – 0.03, 0.38 + 0.03), or (0.35, 0.41), and this interval does contain $p = 0.40$. On the other hand, for example, if $\hat{p}$ is 0.44, the interval will be (0.41, 0.47) which does not contain p.

- Since the sampling distribution is mound shaped, the Empirical Rule tells us that about 95% of the data lies within 2 standard deviations (that is, 2 standard errors) on either side of the mean.
- Also, we know that the mean is $p = 0.40$, and (rounded to the nearest hundredth) two times the standard error is $2 \cdot se = 2\sqrt{\frac{0.4(1-0.4)}{1000}} = 0.03$. So, "within 2 standard deviations of the mean" translates as between 0.37 and 0.43.

Putting all these observation together, we see that:

For n = 1000 and p = 0.40, approximately 95% of all samples have a sample proportion that is within about 3% of the population proportion. So choosing n = 1000 and m.e. = 3% gives a correct answer for approximately 95% of all samples; that is, give a 95% confidence level.

COMMENT: The analysis in this example boils down to this: the reason 3% is a reasonable margin of error for this situation is that for the sampling distribution, $2 \cdot se$ (that is, 2 standard deviations) turns out to be approximately 3%. It turns out that for any population proportion, $2 \cdot se$ is approximately 3% or perhaps even lower. (If it is lower, we could have achieved 95% confidence with an even smaller margin of error.) Again, these ideas will be explored in more detail in the next lesson.

Exercise 13. Illustrate the statement in this comment by calculating $2 \cdot se = 2\sqrt{\frac{p(1-p)}{n}}$ for $n = 1000$ and these values of p. Record your answers as a decimal rounded to 4 places, then as a percent rounded to the nearest percent. The first is already done, using the example above.

p	**Decimal**	**Percent**
0.40	0.0310	3%
0.30		
0.85		
0.57		
0.68		

Summary: *In general, using a sample of size 1000 and a margin of error of 3% yields correct answers about 95% of the time or more. This is true because the "sampling distribution of the sample proportions" is mound-shaped, with two standard errors (two standard deviations) equal to about 3% or less.*

Solutions to Exercises

Most of the exercises have no specific solutions, since the results from running the applet will vary. In some cases we show the author's results; yours will be different but similar.

11. Use the applet to reproduce what the author did – that is, generate 2500 samples each with n = 1000. Because the app chooses a different population proportion p each time it is run, your results will not be identical, but they should be similar to those above. Is the resulting histogram mound shaped? Is the mean for the histogram close to the population proportion for your run of the applet?

b. Reload the applet and repeat part (a) for a different population proportion.

For the author's first run, the population proportion was p = 0.44. The histogram was mound shaped with mean 0.4396, very close to the population proportion. For the second run, the population proportion was p = 0.61. The histogram was again mound shaped, with mean 0.6098, very close to the population proportion.

12. Use the applet to experiment, as follows. Load the applet and record the population proportion (p) it will be using. Then use the start sampling / stop sampling buttons to generate about 200,000 samples, and record your answers to these questions:

a. Is the histogram mound-shaped? Does it closely match the overlaid normal curve?

b. What is the mean for the histogram? Is this close to the population proportion p?

c. Use the $\sqrt{\frac{p(1-p)}{n}}$ formula to calculate the standard error *se* (that is, the standard deviation for the sampling distribution of the sample proportions).

d. What is the standard deviation for the histogram? Is this close to what you calculated in step (c)?

Here are the author's results, yours will be different but the yes/no answers should be the same: Population proportion for the run was 0.58. (a) Yes, yes (b) 0.58, yes (c) 0.0156076904 (d) 0.0156, yes

13. Illustrate the statement in this comment by calculating $2 \cdot se = 2\sqrt{\frac{p(1-p)}{n}}$ for n = 1000 and these values of p. Record your answers as a decimal rounded to 4 places, then as a percent rounded to the nearest percent. The first is already done, using the example above.

p	Decimal	Percent
0.40	0.0310	3%
0.30	0.0290	3%
0.85	0.0226	2%
0.57	0.0313	3%
0.68	0.0295	3%