Chapter 2: Inference and Confidence Intervals for a Single Categorical Variable

In this chapter, we will discuss methods for making decisions concerning research questions involving only a single categorical variable. Even though we'll be adding lots of terminology along the way, we will use the same logical approach to answering questions that was introduced in Chapter 1.

Formal Hypothesis Testing

In previous examples, we have used a logical process to make statistical decisions in problems involving a single categorical variable (e.g., choice – helper/hinderer; font – signet/salem; sex – woman/man). Next, we will add more structure to these statistical investigations by introducing a procedure known as *hypothesis testing*. Before we discuss this procedure, we must introduce a few more definitions.

Population Parameters vs. Sample Statistics

In each of the previous examples, we tested a claim about a **population parameter** of interest.

> **i Definitions**
>
> - A **parameter** is a numerical descriptive measure of a *population*. This value is almost always unknown, and our goal is to either estimate this parameter or test claims regarding it.
>
> - A **statistic** is a numerical descriptive measure of a *sample*. This value is calculated from the observed data.
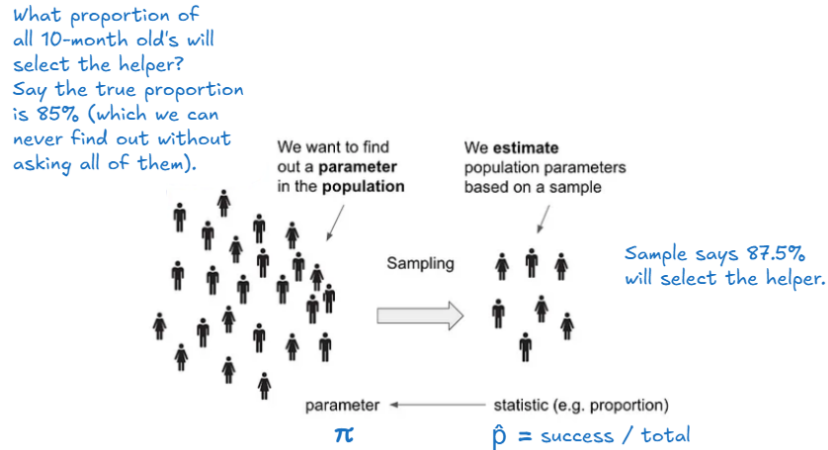
Image Source: medium.com.

Example	Statistic $\hat{p}$	Parameter π
Example 1.2: Helper vs. Hinderer	$\hat{p} = 14/16 = 0.875$	π : the population proportion of all 10-month old infants who prefer the helper.
Example 1.3: Are women passed over over for managerial training?	$\hat{p} = 9/20 = 0.45$	π: the population proportion of all employees at the large supermarket selected for the management training who are women (think long term)
Example 1.4: Font Preference	$\hat{p} = 30/40 = 0.75$	π: the population proportion of all potential U.S. consumers who prefer signet.
	$\hat{p} = 10/40 = 0.25$	π: the population proportion of all potential U.S. consumers who prefer salem.

> **i** Definitions
>
> **Hypothesis testing** is a procedure, based on sample evidence and probability, used to test a claim regarding a population parameter. The test will measure how well our observed sample statistic agrees with some assumption about this population parameter.

Before you begin a hypothesis test, you should clearly state your research question. For instance, let's reconsider the research question from three of our previous examples.

Example	Research Question
Example 1.2: Helper vs. Hinderer	Do 10-month-old infants tend to prefer the helper toy over the hinderer toy?
Example 1.3: Are women passed over for managerial training?	Is there statistical evidence for sex discrimination against women?
Example 1.4: Font Preference	Do the majority of all consumers prefer one font over the other?

Setting up the Null and Alternative Hypothesis

> **ⓘ Definitions**
>
> - The **null hypothesis**, H_0, is what we will assume to be true, and we will evaluate the observed data from our study against what we *expected* to see under the null hypothesis. This will always contain a statement saying that the population parameter is **equal** to some value.
>
> - The **alternative hypothesis**, H_A, is what we are trying to show. Therefore, the research question is restated here in the alternative hypothesis. This will always contain statements of inequality, saying that the population parameter is **less than, greater than, or different from** the value in the null hypothesis.
>
> *Note: The hypotheses are ALWAYS about the population parameter.*

For our three examples, the null and alternative hypotheses are shown below.

Research Question	Hypotheses
Do 10-month-old infants tend to prefer the helper toy over the hinderer toy?	H_0: There is no preference for one toy over the other in the population of all 10-month-olds. H_A: The majority of all 10-month-old infants prefer the helper toy.

Research Question	Hypotheses
Is there statistical evidence for sex discrimination against women?	H_0: The selection process is fair. H_A: The selection process is biased against women.
Do the majority of all consumers prefer one font over the other?	H_0: There is no preference for one font over the other in the population of all consumers. H_A: The majority of all consumers would pick one font over the other.

Note that we can also state these hypotheses in terms of the population parameter of interest:

Research Question	Hypotheses
Do 10-month-old infants tend to prefer the helper toy over the hinderer toy?	H_0: The population proportion of all 10-month olds' who prefer the helper is equal to 0.50.
	H_A: The population proportion of all 10-month olds' who prefer the helper is greater than 0.50.

Research Question	Hypotheses
Is there statistical evidence for sex discrimination against females?	H_0: the population proportion of all employees at the large supermarket selected for the management training who are women is equal to 0.60
	H_A: the population proportion of all employees at the large supermarket selected for the management training who are women is less than 0.60
Do the majority of all consumers prefer one font over the other?	H_0: the population proportion of all potential U.S. consumers who prefer signet is equal to 0.50.
	H_A: the population proportion of all potential U.S. consumers who prefer signet is different from 0.50.

Evaluating Evidence Using p-values

In in each of our three examples, we ***assumed the null hypothesis was true*** when setting up our probability for our Applet simulations. Then, we used the results simulated under this scenario to help us decide whether observing results such as our sample data would be an unusual event *if the null hypothesis were true.*

Up to this point, whether an observed result was considered unusual (or extreme) has been a rather subjective decision. Now, we will discuss the guidelines used by statisticians to determine whether an observed result is extreme enough under the null hypothesis for us to conclude that the evidence supports the research question.

First, note that in our three examples, we examined different parts of the distribution of simulated results when deciding whether the observed data was extreme. Each of these cases is an example of a specific type of hypothesis test.

Research Question	Hypotheses	Type of Test
Do 10-month-old infants tend to prefer the helper toy over the hinderer toy?	H_0: $\pi = 0.50$ H_A: $\pi > 0.50$	 Upper-tailed Test
Is there statistical evidence for sex discrimination against women?	H_0: $\pi = 0.60$ H_A: $\pi < 0.60$	 Lower-tailed Test
Do the majority of all consumers prefer one font over the other?	H_0: $\pi = 0.50$ H_A: $\pi \neq 0.50$	 Two-tailed Test

Statisticians use the following guidelines to determine whether the observed data supports the research question:

> **i** Note
>
> **Upper-tailed test:** The observed result must fall in the upper 5% of the *null sampling distribution*.

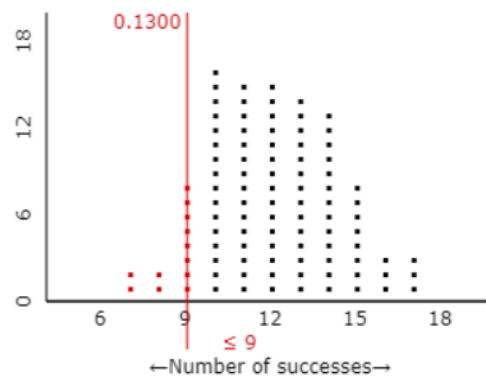Do 10-month-old infants tend to prefer the helper toy over the hinderer toy?
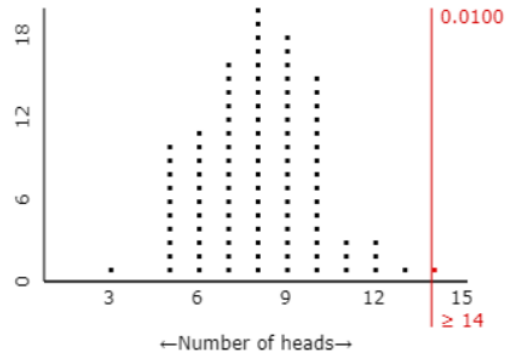
Count to the right.

> **i** Note
>
> **Lower-tailed test:** The observed result must fall in the lower 5% of the null sampling distribution.

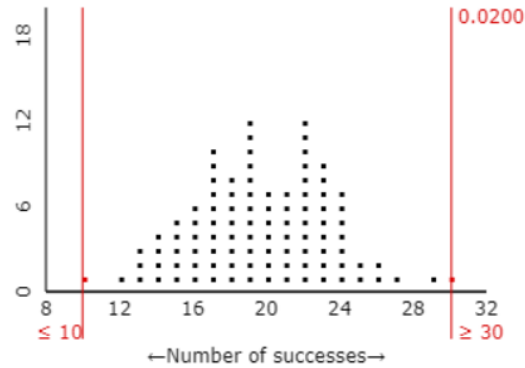Is there statistical evidence for sex discrimination against women?

Count to the left. p-value $= 1/100 = 0.01 < 0.05$ -> reject the null

> **i** Note
>
> **Two-tailed test:** The observed result must fall in either the upper 2.5% or the lower 2.5% of the null sampling distribution.

Do the majority of all consumers prefer one font over the other?

Count both tail ends (10 or fewer is just as extreme as 30 or more). p-value $= 1/100 = 0.01 < 0.05$ -> reject the null

> 💡 **Tip**
>
> In practice, two-tail tests are typically conducted over an upper- or lower-tailed test.

Statisticians use what is called a **p-value** to quantify the amount of evidence that an observed result from a set of data provides for a research question.

> ℹ️ **Definition**
>
> **p-value**: The probability of observing an outcome as extreme (or even more extreme) than the observed study result, assuming the null hypothesis is true.

Note that in each of the above examples, we obtained the simulated results ***assuming the null hypothesis was true*** (plotted on the *null sampling distribution*). Therefore, to estimate the p-value, we simply determine how often outcomes as extreme (or more extreme) than the observed study results appeared in our simulation study.

Example	Estimate of p-value
Helper vs. Hinderer?	p-value=0.01
Are Women Passed Over for Managerial Training?	p-value=0.13
Font Preference?	p-value=0.02

> **i** Making a Decision with p-values and Significance Levels (α)
>
> If the p-value is **less than** α – typically 0.05 (5%), then the data provide enough statistical evidence to *reject the null hypothesis* and support the research question.
>
> If the p-value is **greater than (or equal to)** α – typically 0.05 (5%), then the data do not provide enough statistical evidence to support the research question, and we *fail to reject the null hypothesis*.
>
> Note: We **never** make the decision to *accept the null hypothesis*. Instead, we evaluate whether the evidence supports or does not support the alternative hypothesis.

Why does this decision rule work? Consider the "Helper vs. Hinderer" example. Because the p-value falls below 5%, the observed result *must* have been in the upper 5% of the null sampling distribution. As stated earlier, this implies that the observed study result is very unlikely to happen by chance under the null hypothesis, which supports the research question.

On the other hand, consider the "Are Women Passed Over for Managerial Training" example. Because the p-value was larger than 5%, the observed result *can't* have been in the lower 5% of the reference distribution. This implies that the observed study result is not all that unusual and could have easily happened by chance under the null hypothesis. Therefore, the null hypothesis *could* be true, and we have no evidence to support the research question.

This decision rule is widely accepted for determining whether study results are statistically significant; however, some researchers do advocate using a more flexible rule similar to the following:

> ℹ **General guidelines**
>
> - If the p-value falls below 0.05, we have *strong evidence* to support the alternative hypothesis (i.e., the research question).
>
> - If the p-value falls below 0.10 but above 0.05, we have *marginal evidence* to support the alternative hypothesis (i.e., the research question).
>
> - If the p-value is above 0.10, we *do have not have enough evidence* to support the alternative hypothesis (i.e., the research question).

Next, we will review the steps involved in a formal hypothesis test for each of our three examples. Note that our conclusions are written in the context of the problem. Moreover, even a person with no statistical background should be able to understand these conclusions (i.e., a conclusion should NOT say something like "We reject the null hypothesis." – this is your *decision.*)

Helper vs. Hinderer

Research Question	Do 10-month-old infants tend to prefer the helper toy over the hinderer toy?
Hypotheses	H_0: There is no preference for one toy over the other in the population of all 10-month-olds; that is, the probability a 10-month-old selects the helper toy is 50%. H_A: The majority of all 10-month-old infants prefer the helper toy; that is, the probability a 10-month-old selects the helper toy is greater than 50%.
p-value estimated from simulation	p-value=0.01 < 0.05 -> Reject the null
Conclusion	At an $\alpha = 0.05$, we have strong evidence to conclude the population proportion of all 10-month old infants who prefer the helper is greater than 0.50 (p-value = 0.01).

Are Women Passed Over for Managerial Training?

Research Question	Is there statistical evidence of sex discrimination against women?
Hypotheses	H_0: The selection process is fair; that is, the probability a woman is selected is 60%. H_A: The selection process is biased against women; that is, the probability a woman is selected is less than 60%.
p-value estimated from simulation	p-value=0.13 > 0.05 -> Fail to Reject the Null
Conclusion	At an $\alpha = 0.05$, we do not have enough evidence to conclude the population proportion of all employees at the large supermarket selected for management training who are women is less than 0.60.

Font Preference

Research Question	Do the majority of all consumers prefer one font over the other?
Hypotheses	H_0: There is no preference for one font over the other in the population of all consumers; that is, the probability of a consumer selecting the Signet font is 50%. H_A: The majority of all consumers would pick one font over the other; that is, the probability of a consumer selecting the Signet font is different from 50%.
p-value estimated from simulation	p-value=0.02 < 0.05 -> Reject the Null
Conclusion	At an $\alpha = 0.05$, we have strong evidence to conclude the population proportion of all U.S. consumers who prefer signet font is different than 0.50.

Next, let's carry out a formal hypothesis test for a few new examples.

Example 2.1 Claims of Numbness After Automobile Accident

A 28-year-old developed pain involving the spine and the left side of her body after an auto-mobile collision. They were actively involved in a personal litigation against the company that owned the other vehicle, and they reported constant pain and numbness in the left arm. To test their claims, researchers touched their left arm with either 1 finger or 2 fingers simultaneously while their eyes were closed. The word "touch" was said simultaneously with the presentation of the tactile stimulus so that the subject knew when to respond. She then had to indicate whether she felt 1 single touch or 2 simultaneous touches (with the double-touch stimulus, the fingertips were always spaced 2 inches apart). The subject received 100 stimuli overall; they were correct on 30 of them. Is there statistical evidence that they are intentionally answering incorrectly?

1. Identify both the population and sample of interest.

Population: all possible stimuli trials that could have been given to the 28-year old

Sample: n = 100 stimuli the 28 year-old received

2. Identify the single categorical variable of interest.

Outcome (Correctly identified the touch/Incorrectly identified the touch)

3. Identify both the parameter and statistic of interest.

Parameter: the proportion of all possible stimuli trials the 28 year old would answer correctly.

Statistic: $\hat{p} = 30/100 = 0.3$

4. Carry out the formal hypothesis test to address the research question.

- **Research Question** Is there statistical evidence that she is intentionally answering incorrectly?

- Hypotheses:

Wording based on research question -

Null: The 28 year-old is just guessing for all possible stimuli.

Alternative: The 28 year-old is intentionally answering incorrectly for for all possible stimuli.

Wording based on parameter -

Null: the proportion of all possible stimuli trials the 28 year old would answer correctly is equal to 0.5.

Alternative: the proportion of all possible stimuli trials the 28 year old would answer correctly is less than 0.5.

Using symbols -

$H_0 : \pi = 0.5$

$H_A : \pi < 0.5$

- Estimate p-value (Carry out the simulation study to investigate this p-value, sketch or paste your simulation results here)

Use the simulation results to estimate the p-value: <0.001 < 0.05 -> Reject the Null

(note that when we never have a simulation set as extreme or more extreme than what we observed, the p-value is "0", very very small!)

- Conclusion:

At an $\alpha = 0.05$, we have evidence to conclude the 28 year-old is intentionally answering incorrectly for for all possible stimuli (p < 0 .0001).

Example 2.2: Effectiveness of an Experimental Drug

Suppose a commonly prescribed drug for relieving nervous tension is believed to be only 70% effective. Experimental results with a *new* drug administered to a random sample of 20 adults who were suffering from nervous tension show that 18 received relief. Is there statistical evidence that the new experimental drug is more than 70% effective?

1. Identify both the population and sample of interest.

Population: all adults suffering from nervous tension

Sample: n = 20 adults who received the new drug for nervous tension

2. Identify the single categorical variable of interest.

Relief (yes/no)

3. Identify both the parameter and statistic of interest.

Parameter: the proportion of all adults who receive relief from nervous tension

Statistic: $\hat{p} = 18/20 = 0.9$n

4. Carry out the formal hypothesis test to address the research question.

- **Research Question** Is there statistical evidence that the new drug is more than 70% effective?

- Hypotheses:

In words:

Null: the proportion of all adults who receive relief from nervous tension is equal to 0.7.

Alternative: the proportion of all adults who receive relief from nervous tension is greater than 0.7.

In symbols:

$H_0 : \pi = 0.7$

$H_A : \pi > 0.7$

- Estimate p-value (Carry out the simulation study to investigate this p-value, sketch or paste your simulation results here)

Describe process:

Probability of success (π): 0.7

Sample size (*n*): 20

Number of samples: 100

☐ Show animation

Draw Samples

Total Samples = 100

Choose statistic:

○ Number of successes

◉ Proportion of successes

Count samples

As extreme as [=] [0.9] [Count]

Proportion of samples:
3 / 100 = 0.0300

Most recent results

Number of Successes = 11

Number of Failures = 9

☐ Summary Statistics

0.0300

←Proportion of successes→

= 0.9000
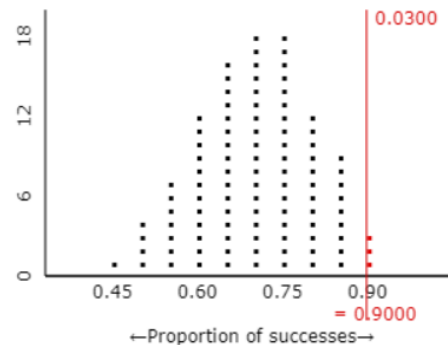
☐ Show previous results

☐ Show sliders

Use the simulation results to estimate the p-value: $0.03 < 0.05$ -> Reject the Null

- Conclusion:

At an $\alpha = 0.05$, we have evidence to conclude the proportion of all adults who receive relief from nervous tension is greater than 0.70 (p = 0.03).

Theoretical Approach to p-values

There is one caveat regarding our current approach to obtaining a p-value. Certainly, different simulations will produce slightly different null sampling distributions. The general pattern will be the same, but variations do exist. For example, consider the Helper vs. Hinderer study.

Helper vs. Hinderer

- **?**Do 10-month-olds tend to prefer the helper toy over the hinderer toy?
- Hypotheses

- H_0: There is no preference for one toy over the other in the population of all 10-month-olds; that is, the proportion of all 10-month-old infants who select the helper toy is equal to 0.50.
- H_A: The majority of all 10-month-old infants prefer the helper toy; that is, the proportion of all 10-month-old infants who select the helper toy is greater than 0.50.
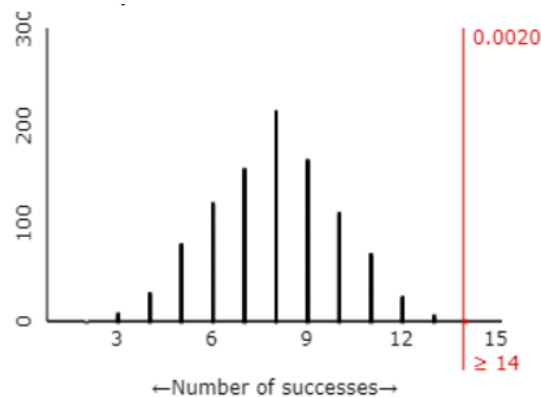
The *observed* result was as follows: 14 out of 16 infants chose the helper toy. What if two different researchers each carried out their own simulation study with 1,000 repetitions to estimate the p-value?

- p-value

Simulation #1: p-value: $0.004 < 0.05$ -> Reject the Null
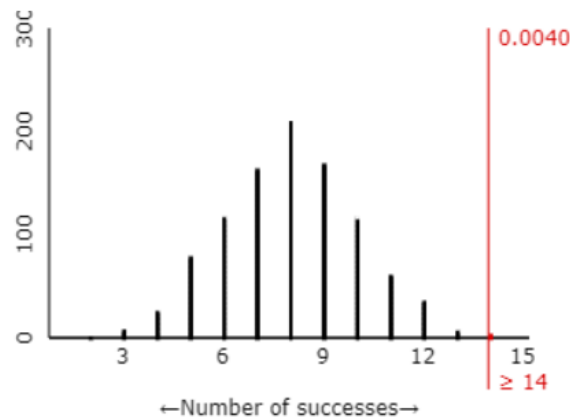
Simulation #2: p-value: $0.002 < 0.05$ -> Reject the Null

- Conclusion: At an $\alpha = 0.05$, we have evidence to conclude the proportion of all 10-month-old infants that prefer the helper toy is greater than 0.5 (p = 0.002).

Fortunately, regardless of which simulation study we use in the previous example, the final conclusion is the same and the discrepancy between the two estimated p-values is minimal. Note that as the number of repetitions in our simulation study increases, we would expect less discrepancy between these two estimates. So, instead of using a **simulation study** with only 100 (or 1000) repetitions to *estimate* the p-value, we would ideally like to simulate this experiment over and over again, say an *infinite* number of times. This would provide us with the **theoretical probabilities** of interest so that can get exact p-values instead of an estimate of the p-value.

The following graphic shows what the null sampling distribution would look like if we kept repeating the simulation study over and over again, each time counting and plotting the number of infants that chose the helper toy (assuming there was no real preference in the population of all infants). This is known as the **binomial distribution**.
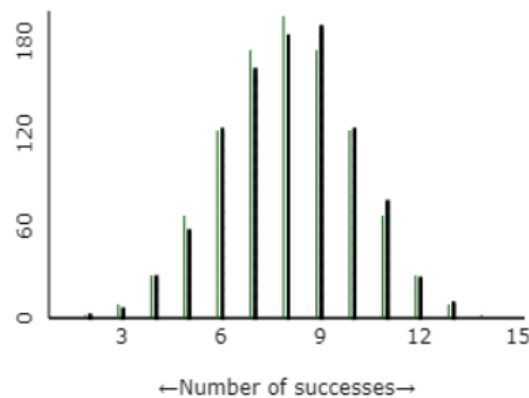

Figure 1: The black dots/bars show results from 1000 simulations; the green bars show the results from the binomial distribution. Notice they follow a similar pattern.

Statisticians often use the binomial distribution to calculate p-values when testing claims about a population proportion. However, before using the binomial distribution, we should check to make sure the following conditions are met.

> **i** The Binomial Distribution - When can we use it?
>
> This distribution can be used whenever the following assumptions are met:
>
> - There exist a fixed number of trials, n.
> - There are only two possible outcomes on each trial ("success" or "failure").
> - The probability of "success" π remains constant from trial to trial.

- The n trials are independent.

1. Check whether these assumptions seem reasonable for the Helper vs. Hinderer study.

- There exist a fixed number of trials, n. n = 16

- There are only two possible outcomes on each trial ("success" or "failure"). helper/hinderer

- The probability of success (π) remains constant from trial to trial. $\pi = 0.50$

- The n trials are independent. The choice of one infant does not give us information about the choice of another infant

Once the conditions for the binomial distribution have been met, we can calculate binomial proportions of interest. While there is a mathematical formula for calculating these probabilities, we will not go into that in this class. Instead, I have provided you the calculated probabilities for the Helper vs. Hinder example in the graph below:
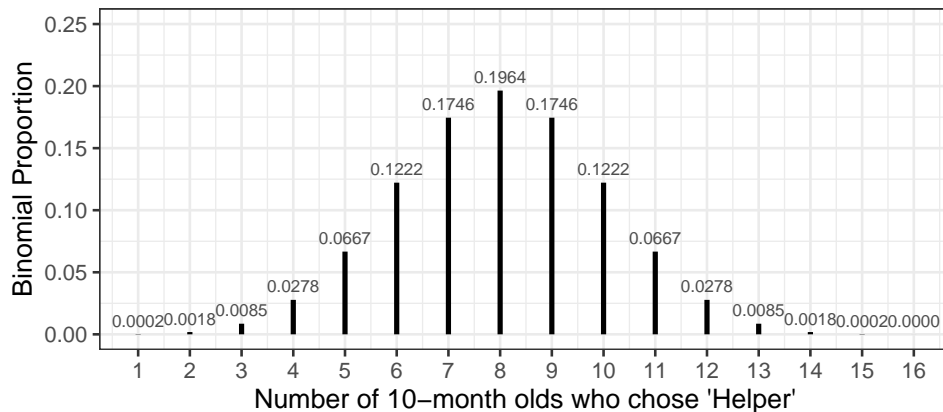
1. Find the probability of observing *exactly* 14 of the 16 10-month-old infants choosing the helper toy, assuming that the population of all 10-month-olds has no preference.

0.0018 -> 0.0018*100 = 0.18%

2. What is the exact **p-value** found using the binomial proportions?

0.0018 + 0.0002 + 0.0000 = 0.002

Note you add up the observed + the extremes (in this case, upper-tailed test).

Alternatively, there is a handy function, `binom.test()` that helps us calculate our p-value from binomial proportions and takes the following arguments:

- `x`: the number of successes in our study
- `n`: the sample size for our study
- `p`: the proportion of successes we would expect to see under the null

- `alternative`: "two-sided", "greater", "less"

Below are the results from the Helper vs. Hinder example.

```
binom.test(x = 14, n = 16, p = 0.5, alternative = "greater")
```

```
    Exact binomial test

data:  14 and 16
number of successes = 14, number of trials = 16, p-value = 0.00209
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.656175 1.000000
sample estimates:
probability of success
                 0.875
```

Confidence Intervals for a Single Categorical Variable

When carrying out hypothesis tests, we are testing claims about population parameters. Sometimes, however, our goal is simply to **estimate** a parameter of interest. Statisticians typically do this with a **confidence interval**

> **i** Definition
>
> **Confidence interval:** a range of plausible values for the parameter of interest.

The big difference between *hypothesis testing* and *confidence intervals* is as follows: the construction of a confidence interval does NOT require any hypotheses concerning our population parameter of interest. The goal of a *hypothesis test* is to draw a conclusion about a claim whereas the goal of a *confidence interval* is to make an *estimate* for the parameter.

For example, consider the following scenario.

Example 2.3: Effectivenss of Acupuncture

In 2008, researchers from the University of Pennsylvania used data from the 2002 National Health Interview Survey (NHIS) to determine the proportion of survey respondents who reported effectiveness of acupuncture treatment for specific conditions. The NHIS is a nationally representative household interview survey of the U.S. civilian non-institutionalized population aged 18 years or older.

In one part of this study, the researchers identified 89 survey respondents who reported seeing a practitioner of acupuncture to treat back pain. Of these 89 respondents, 77 reported that the acupuncture helped improve their symptoms.

Source: 2008. Patrick J. LaRiccia et al. "Perceived Effectiveness of Acupuncture: Findings from the National Health Interview Survey." *Medical Acupuncture,* Volume 20, Number 4.

1. Identify the population of interest.

All U.S. civilian non-institutionalized individuals above 18 who reported seeing a practitioner of acupuncture to treat back pain.

2. Identify the sample

n = 89 survey respondents

3. Identify the parameter of interest.

π the population proportion of all U.S. adults seeing an acupuncture specialist who saw improvement in their back pain.

4. Identify the observed sample statistic.

$\hat{p} = 77/90 = 0.865$ -> 86.5% saw improvement in back pain

Note that the goal of this study was clearly identified in the abstract for this manuscript:

ABSTRACT

Background: Knowledge of perceived benefit from acupuncture treatment is important to predict who is using acupuncture, to inform physicians of the possible benefits of acupuncture, to determine where rigorous research should be focused, and to help policy makers predict future demand.
Objectives: To determine the proportions of survey respondents who reported perceived effectiveness of acupuncture treatment for specific conditions;

Recall that for the specific condition of back pain, 77 respondents reported that the acupuncture helped improve their symptoms.

The objectives don't mention anything about a specific research hypothesis (for example, they're not conducting this study in order to show that "the majority of patients with back pain get relief from acupuncture"). So, we have no null or alternative hypothesis of interest. Instead, the goal is simply to estimate π, the population parameter of interest. What is our best guess for π based on the observed data from the sample?

$\hat{\pi} = 77/89 = 0.865$

Note! The point estimate here (symbol $\hat{pi}$) is the same value as our observed statistic ($\hat{p}$). We just use a different notation mainly to talk about this being our "best guess" for π, but you can think of these as interchangeable.

This is called a **point estimate.** However, we know that if we obtained a different random sample of 89 survey respondents, this point estimate would most likely change. The purpose of using an **interval estimate** (i.e., a confidence interval) instead of this point estimate is to account for the inherent sample-to-sample variation that exists in a sample statistic. Once we determine how much variation is expected in the statistic, we can obtain a range of likely values for the population parameter (called the confidence interval). So, our next goal is to determine how much variation we should expect under repeated sampling.

Approach 1: Simulation

Describe process:

Probability of success (π): 0.865
Sample size (n): 89
Number of samples: 100

    5. Why is the Sample size (n) set to 89?

The number of individuals in our original observed sample.

    6. Where did the 0.865 value come from?

This is our point estimate of π, the population proportion of all US adults seeing a specialist who saw improvement in their back pain. Note this comes from our one observed sample. Our "best guess".

    7. Why did we use 0.865 for the probability of success?
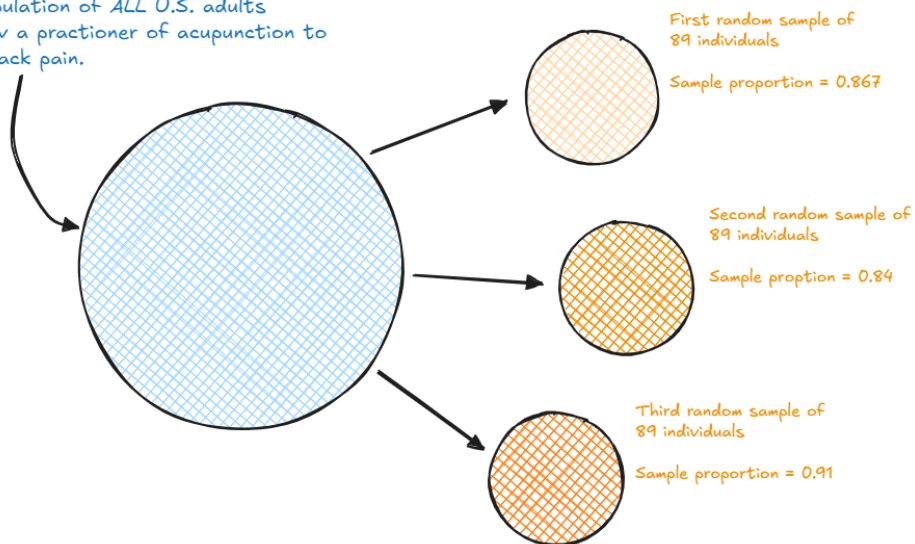
We want to know what the sampling distribution looks like *assuming* the population proportion really is the observed point estimate.

Note: this is the process we have been doing with our null distributions – *under the assumption we are just guessing etc.*. We just change our assumptions to be under the assumption the population parameter really is what we saw in our sample. Then we want to know the sample to sample variability (see below).

> ℹ **Definition**
>
> **Sampling Distribution:** is the distribution of a statistic (like the sample proportion) calculated from many random samples of the same size from a population. It shows how the statistic varies across different samples.
>
>
>
> *Note: We have been creating null sampling distributions – "under the assumption the null is true". This is the same concept, except we now assume the population parameter really is the point estimate.*

8. Once we obtain the results of several runs of the simulation and plot the outcomes, on what value do you expect this distribution to be centered? Why?

←Proportion of successes→
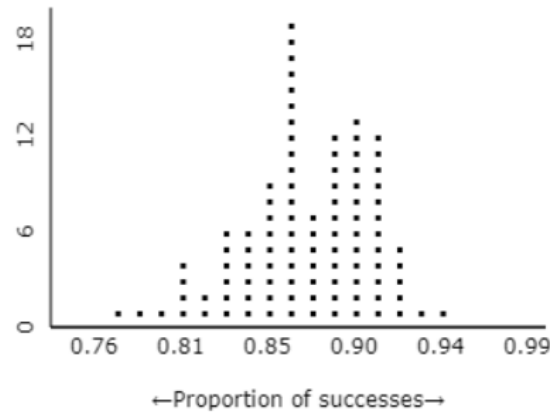
Note there are 100 dots and the sampling distribution is centered at 0.865 (our point estimate!).

A formal approach to calculating a confidence interval involves identifying the middle 95% of the sampling distribution. That is, we need to find a lower endpoint that separates the bottom 2.5% of the distribution and an upper endpoint that separates the top 2.5% of the distribution.

9. Identify the lower and upper endpoints that separate close to the bottom 2.5% and the top 2.5% of the outcomes.

(0.79, 0.93)

Note that there are ~95 "dots" between these two end points.

10. What information does this range of values really provide us?

A *plausible interval* for the population proportion of all U.S. adults seeing an acupuncture specalist who see improvement in back pain.

Approach 2: Binomial

Next, recall that we previously discussed the fact that running a simulation study to enables us to estimate proportions for our sampling distribution that we actually could obtain using the binomial distribution.

The distribution appears as follows:
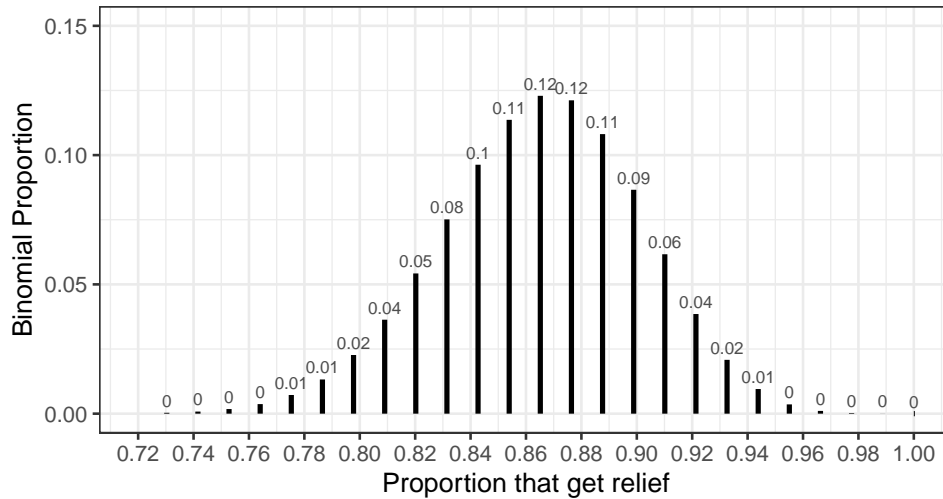
11. Identify the lower and upper endpoints that separate close to the bottom 2.5% and the top 2.5% of the outcomes.

(0.80, 0.93)

Note this leaves ~95% in the "middle". I know this actually is more like 94%, but close enough! We will talk about why this feels subjective and what is done in practice instead (aka below!).

Approach 3: Normal Approximation (aka Theoretical)

With the previous two approaches, we were simply trying to understand the basic idea behind a confidence interval. A statistician would probably not use Simulation or the binomial distribution in the same way that we have just illustrated to calculate the endpoints for this confidence interval (think about some reasons why not).

Instead, to calculate a confidence interval for a proportion, a statistician could make use of the normal curve (i.e., the bell-shaped curve). Note that the normal curve approximates the binomial distribution with a sample size of 89 and $\pi = 0.865$ fairly well:
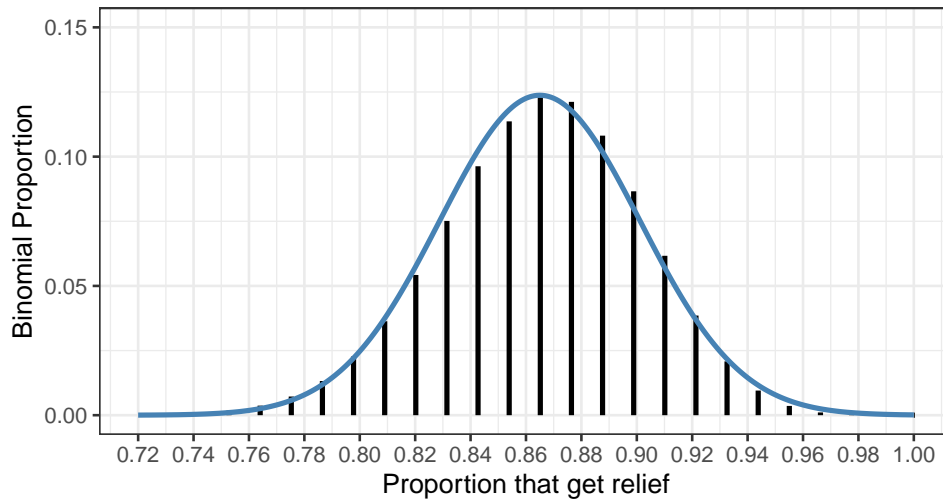
Figure 2: The normal curve superimposed on binomial proportions.

We think of the area under that blue line as 100%! It is now continuous and we want to shade in and find the middle 95%. What end points capture that? we will use a math formula!

General formula –> point estimate ± critical value * standard error

To calculate the endpoints for a 95% confidence interval constructed using normal theory methods, we find the values on the x-axis of the above graph that separate the middle 95% from the rest. This is done using the following steps:

12. Start with the **point estimate**, $\hat{\pi}$: $77/99 = 0.865$

13. Calculate the **standard error** associated with this point estimate:

$$\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} =$$

$$\sqrt{\frac{0.865(1-0.865)}{89}} = 0.0362$$

Careful about order of operations! Take it step by step if you need.

14. Calculate the **margin of error**: This is defined as 1.96 standard errors for a 95% confidence interval. *(This is discussed more on the next page)*.

26

$$1.96 \times \sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}} =$$
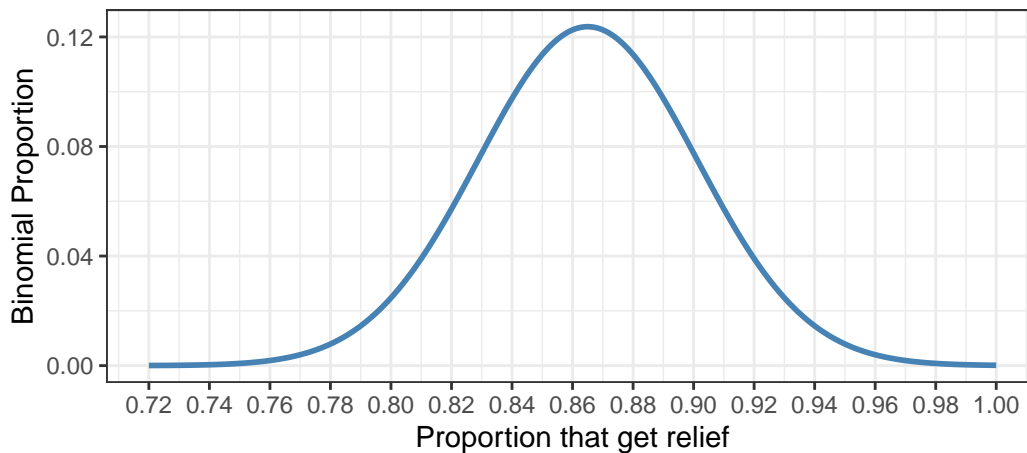
$1.96 * 0.0362 = 0.071$

15. Find the **endpoints** of the confidence interval:

Lower endpoint $= \hat{\pi} - 1.96 \times \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = 0.865 - 0.071 = 0.794$

Upper endpoint $= \hat{\pi} + 1.96 \times \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = 0.865 + 0.071 = 0.936$

Sketch the following values for our example on the normal curve shown below:

- the point estimate
- the lower and upper endpoints
- the margin of error.

16. What is the 95% confidence interval based on the normal theory approach for this problem?
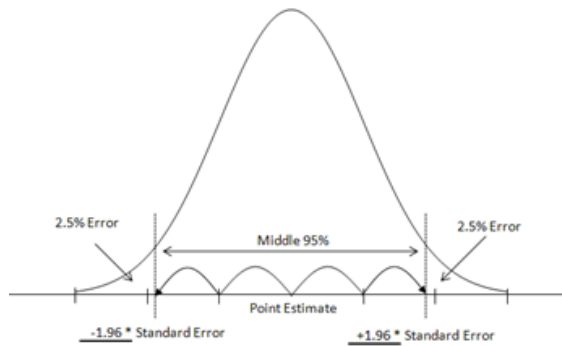
(0.794, 0.936)

17. What is the interpretation of this confidence interval?

We are 95% confident the population proportion of all U.S. adults seeing an acupuncture specialist who see imporvement in back pain is betwen 0.794 and 0.936.

> ### 💡 More on the Margin of Error
>
> **Question:** Why is 1.96 used in the formula for a 95% confidence interval?
> **Answer:** Because with the normal distribution, using a margin of error of 1.96 × standard error gives us an error rate of 2.5% in each tail, resulting in a total error rate of 5%.
>
>
>
> - If you are constructing a 95% confidence interval, then margin of error = 1.96 * standard error
> - Similarly, if you are constructing a 90% confidence interval, it can be shown that margin of error = 1.645 * standard error
> - If you are constructing a 99% confidence interval, it can be shown that margin of error = 2.575 * standard error

> ### ⚠️ Warning
>
> Several methods exist for constructing a confidence interval for a binomial proportion. We have considered a method known as the "Wald" interval (normal approximation). **Note that this method does not typically perform well for very small sample sizes or when the point estimate is very close to either zero or one.**

This is because the normal curve does not approximate the binomial distribution very well in these situations. For example, consider the binomial distribution with $\pi = 0.10$ and $n = 20$:
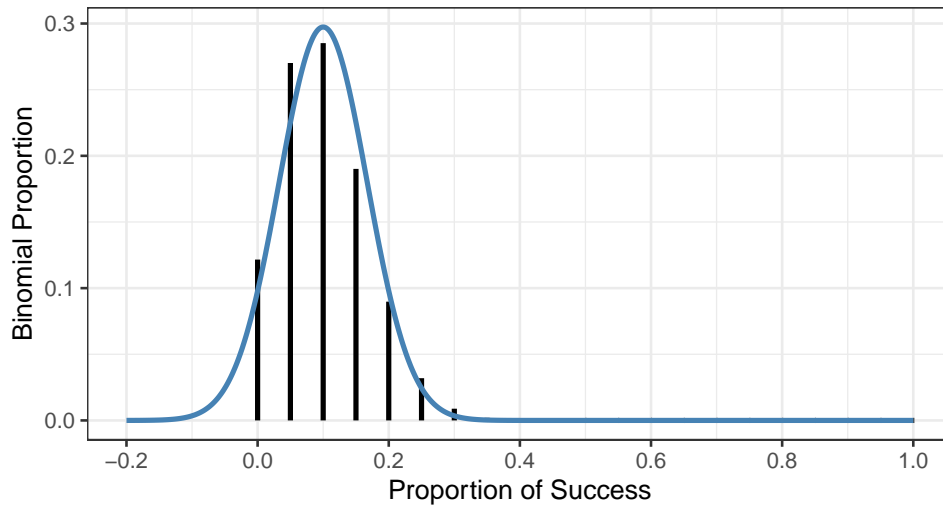
This distribution is not approximated well by a bell-shaped curve. If we encountered a study with $n = 20$ and $\pi = 0.10$, then the normal approxmiation method would not provide us with a very reliable interval estimate of the true population proportion.

> **ℹ Conditions for One Proportion Normal Approximation**
>
> One rule of thumb is that as long as there are 10 "successes" and 10 "failures", the normal approximation should work reasonably well.

18. Check this condition for the Acupuncture example. Does the Wald interval seem to be appropriate here?

77 reported relief (success) > 10 – GOOD! 12 did not reported relief (failure) > 10 – GOOD! Thus, our conditions for using the normal approximation are met!

> **ℹ Big Picture: Confidence Intervals**
>
> Confidence intervals provide a range of plausible values for a population parameter (like the population proportion) based on a sample. Rather than giving a single estimate, they reflect the uncertainty in our estimate due to random sampling from our population. Confidence intervals allow us to account for this variability and make more informed statements about what the population parameter could be.

Extension (on own if desired): What Does "95% Confident" Really Mean?

Once again, consider Example 2.3 regarding acupuncture treatment for back pain. Note that if we were to obtain another random sample of 89 individuals, the sample proportion from

this new study will probably not be equal to $77/89 = 86.5\%$. Thus, the confidence interval we calculate based on this statistic will also change.

Also, note that the population parameter ($\pi =$ the proportion of the population that get relief) is equal to one true value, which happens to be unknown. In a single study conducted in the real world, we have no way of knowing for sure whether or not the confidence interval obtained from this study has captured this one true value.

We can, however, conduct a simulation study to see how well our methods for constructing a confidence interval work. In this simulation study, we will set the population proportion to a value that is known, construct confidence intervals from several random samples, and see how often the resulting interval actually captures the true value of the population proportion.

This process is easy to do using an applet available on the web. `Online Simulation Applets > Simulating Confidence Intervals for Population Parameter`

Link: http://www.rossmanchance.com/applets/2021/confsim/ConfSim.html

Carry out the following steps.

1. Let's assume that the true population proportion for our acupuncture example is $\pi = 0.80$. In the applet, set the value of π to be 0.80 and n to be 89. Also, make sure that the method is set to "Wald" for "Proportions."

2. Click on "Sample." Then, click on the interval to see the sample proportion that was obtained. You should see something similar to the following.
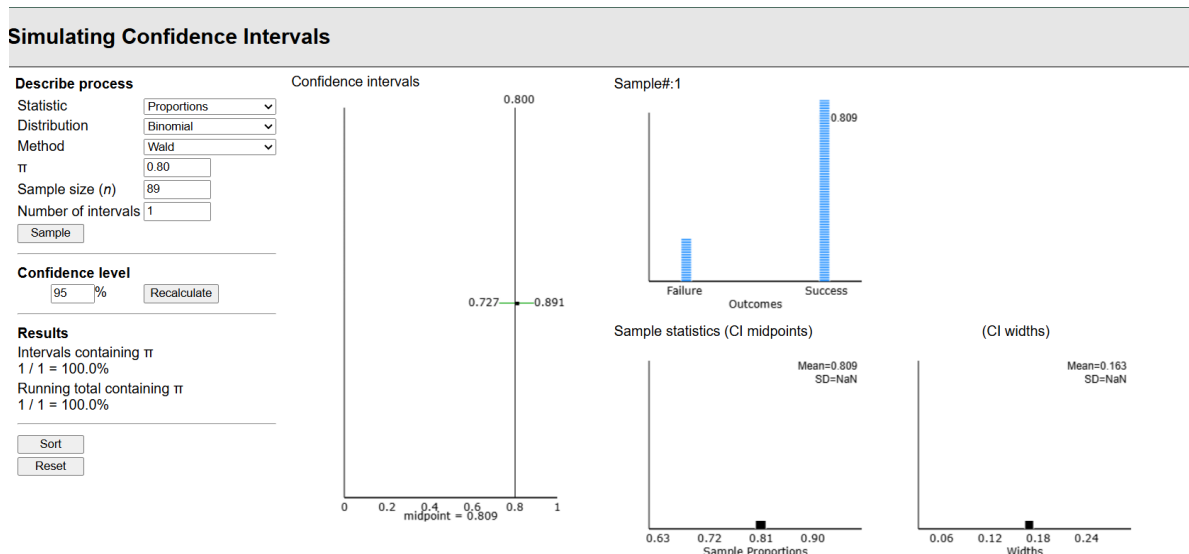
In the example shown above, the first random sample resulted in $\hat{\pi} = 0.809$. The confidence interval constructed using the Wald method was $0.727 \leq \pi \leq 0.891$. Note that this interval did in fact capture the true population proportion we specified, $\pi = 0.80$.

Identify the following from your first random sample in the simulation study.

- Sample proportion :

For this example, $\hat{\pi} = 0.809$. Note yours is a different sample so it might be different!

- 95% confidence interval:

(0.727, 0.891)

- Did your confidence interval capture the true value of $\pi = 0.80$?

Yes! 0.80 falls beween 0.727 and 0.891.

3. Click "Sample" again. Did you get the same interval? Does this new interval capture the true value of $\pi = 0.80$?

My next sample resulted in a 95% confidence interval of (0.740, 0.900). Yes, this still captured 0.80. Let's do this again and again!

4. To investigate what happens in the long run, we can use the applet to take many more random samples and construct a confidence interval for each. Change the number of "Intervals" from 1 to 198 and click "Sample."

   What percentage of the 200 intervals captures the true value of the population proportion, $\pi = 0.80$? Note that this information is given in the "Running Total."

Mine had 187 of the 200 repeated samples capture the population proportion of 0.80 (this is 93.5%). Those capturing the population proportion are shown in green and the intervals not are shown in red (terrible colors, I know).

5. Change the number of "Intervals" from 198 to 200 and click "Sample" over and over again until you've created 1,000 confidence intervals. What percentage of these 1000 intervals captures the true value of the population proportion, $\pi = .80$? This is often referred to as the *coverage rate*.

My running total sasy $937/1000 = 93.7\%$ of the repeated samples have confidence intervals that capture the population proportion.

6. Predict how the coverage rate will change if you were to construct 90% confidence intervals instead of 95% confidence intervals. Then, in the applet, change the confidence level to 90% and click "Recalculate." Continue to press "Sample" until you have constructed 1,000 intervals. What percentage of these 1000 intervals captures the true value of the population proportion, $\pi = 0.80$?

The coverage rate will be closer to 90% of the repeated samples having confidence intervals that capture 0.80 instead of 95% capturing it.

7. Now, recall that in the actual study, you get only one sample and thus one confidence interval. Do you know for sure whether this one interval captures the truth?

Nope. Hence why we interpret with, "We are 95% confident…"

> **ℹ Big Picture: Meaning of 95% Confident**
>
> These simulations show us that if take repeated samples from a population and construct 95% confidence intervals each time, in the long run approximately 95% of our intervals will succeed in capturing the true value of the population proportion, π. So, when we collect data and calculate a single 95% confidence interval, we say we are 95% confident (or certain) that this interval captures the truth.