**Inferring the Central Limit Theorem**

**Estimated Time for Lesson:** 50-75 minutes (The investigation is designed to take about one class period and can be done individually, in small groups, or as a class.)

**Compatible Instructional Formats:** Individual, small group or whole-class investigation (*Internet access needed*)

**Activity Description:** As one would expect from its title, the Central Limit Theorem is one of the foundational theorems of the realm of statistics, as it justifies and makes it feasible to use many of the fundamental analytical tools from approximating measures of typical performance and spread to hypothesis testing and construction of confidence. Unfortunately, the theorem is all too often taught from a “Here are some facts; memorize them” approach. The activity described herein is intended to provide an alternative to the traditional expository approach. That process begins by having students consider a problem set in a realistic context that the Central Limit Theorem would help the students both solve and justify their solution. Upon realizing that even if they are able to provide an informed answer to the question posed, they have no statistically sound method of justifying that answer. Out of that realization stems a necessity to learn about the nature of the sampling distribution for the mean. Construction of that knowledge is facilitated by affording students the opportunity to use an online simulation tool to quickly generate sampling distributions and identify relationships and trends that will allow them to infer both the assertions and prerequisite conditions of the Central Limit Theorem, and then apply what they have learned to determine and justify an answer to a problem set in a realistic context. In the process, students will produce sampling distributions for each of at least three sample sizes, with the sample sizes becoming progressively smaller, and students will explore how changing the parent population’s distribution impacts the inferences drawn, if at all. Upon completion of all of the preceding, students then return to the original problem, solve it and justify their solution. Thus, the lesson can be used in an introductory statistics or quantitative reasoning course, and it interrelates virtually all of the content explored in such courses, making it easily worth the time investment.

**Primary Conceptual Goals:** The proposed ARC for teaching the Central Limit Theorem is intended to support students in attaining the following conceptual goals:

* Via the use of online simulation software, students will build experientially-based, conceptual understanding of the concept of a sampling distribution as a distribution of a sample statistic that is constructed via collecting many samples of the same size, calculating the relevant statistic, and constructing a graph of the results.
* Students will understand that just as a population distribution or the distribution of a sample can be described in terms of its shape, center and spread, so too can a sampling distribution.
* For at least 2 parent populations, students will use simulation software to generate 3 sampling distributions for the mean, with each sampling distribution based on the same parent population. However, a different sample size will be used in the construction of each of the 3 sampling distributions to be produced. After synthesizing the sampling distributions, students will reflect on their results and their implications to infer relationships between the:
  + Mean of the sampling distribution and mean of the population
  + Standard deviation of the sampling distribution and the standard deviation of the population
  + Shape of the sampling distribution in relation to sample size (taking into account difference in whether or not sample size is still a relevant factor in determining the shape of the sampling distribution if the parent population is normally distributed

**Underlying structures of the concepts to be learned:**

* + 1. Via the use of online simulation software, students will build experientially-based, conceptual understanding of the concept of a sampling distribution as a distribution of a sample statistic that is constructed via collecting many samples of the same size, calculating the relevant statistic, and constructing a graph of the results.
    2. Students will understand or lay a foundation on which to understand that every variable and every sample statistic has a distribution.
    3. Students will understand that just as a population distribution or the distribution of a sample can be described in terms of its shape, center and spread, so too can a sampling distribution.
    4. Students will understand that there are typical (and more likely) values for sample statistics, just as there are typical (and more likely) values for population parameters.
    5. Students will understand that sample statistics display variability across samples, i.e., if the same statistic is calculated for different samples, different values for the statistic will likely be produced.
    6. Students will extend and refine their concepts of typical performance and variability to apply to the novel context of sampling distributions.
    7. Students will infer a relationship between the mean of a population’s distribution and the mean of the sampling distribution for the mean.
       - Specifically, students will infer that the mean of sampling distribution for the mean equals the population mean, i.e., the mean of the sample means is the population mean.

1. Students will infer a relationship between the standard deviation of a population’s distribution and the standard deviation of the sampling distribution for the mean.
   * + - Specifically, students will infer that the standard deviation of the sampling distribution for the mean equals the standard deviation for the population divided by the square root of the sample size.
2. Students will infer that the sampling distribution for the mean is normally distributed, if the sample size is sufficiently large or the parent population is normal.
   * + - Students will understand that the inferences respectively relating the mean and standard deviation of the mean’s sampling distribution to the population’s mean and standard deviation do not depend on either the sample size or the normality of the population’s distribution. However, to draw the conclusion of the Central Limit Theorem related to the normality of the sampling distribution of the mean, it is necessary that either the sample size be sufficiently large or that the parent population be normally distributed (or both).

**Reflection of MIP Components of Inquiry:** The proposed lesson reflects the principle of *active learning* in that it necessitates that students actively engage their minds in an effort to generate and interpret simulations and in the process, to identify trends and relationships, which are then used as the basis for inferences about the general traits of sampling distributions as well as the criteria that must be met before the identified traits can be assumed to hold true. In the process of doing all the above, the students will be selecting, performing, and evaluating the actions and related mathematical & statistical analytical structures and underlying structure their use reveals, which in turn reveal and equate to the structures of the concepts to be learned. All of this similarly requires that students *meaningfully apply* their knowledge of statistical concepts and structures such as measures of central tendency, measures of variability, population, sample, statistic, parameter, and symbolic and graphic representations and logical reasoning skills and processes that permeate both statistical and mathematical problem solving. Moreover, learners must make, to some extent justify, and generalize the results and requirements of the Central Limit Theorem across contexts based on identified or extracted mathematical and statistical structure. Finally, the proposed ARC reflects the principle of facilitating *academic success skills* by engaging students in activities through which they can do what mathematicians and statisticians do, and in the process, experience the joy of mathematical and statistical creation and problem solving. Through such experiences, students will also construct deep understanding and procedural fluency in applying their learning, which enhances their self-esteem as mathematical and statistical practitioners as well as their appreciation of and motivation to learn about and participate in those fields.

**Actions students will select, perform, and evaluate:**

1. Given a population mean and standard deviation as well as a sample size and possibly whether or not the population is normally distributed, students will be able to:
   * + - Determine or estimate the value of the mean and standard deviation of the sampling distribution and use the Central Limit Theorem to explain why their stated values will be accurate.
       - State that the sampling distribution for the mean will be normally distributed and use the Central Limit Theorem to justify that conclusion or state that not enough information is known to definitively conclude that the sampling distribution for the mean will be normally distributed and use the Central Limit Theorem to provide a justification for that conclusion.
2. Given the mean and standard error (standard deviation) of a sampling distribution for the mean along with a sample size, students will be able to state the parent population’s mean and standard deviation and to use the Central Limit Theorem to explain why their stated values will be accurate.
3. In doing all of the above, students will apply and refine their understandings of:
   * + - The normal distribution and its characteristic shape
       - The mean as a measure of typical performance or central tendency,
       - The standard deviation as a measure of spread
4. Students will infer that the mean of sampling distribution for the mean equals the population mean, i.e., the mean of the sample means is the population mean.
5. Students will infer that the standard deviation of the sampling distribution for the mean equals the standard deviation for the population divided by the square root of the sample size.
6. Students will infer that the sampling distribution for the mean is normally distributed, if the sample size is sufficiently large or the parent population is normal.
7. Students will apply their learning form i-vi to solve a problem in which it is necessary to determine whether or not a sample mean is significantly different enough from the mean of the sampling distribution, i.e., is unusual or extreme enough to warrant the shutting down of a production line.
8. Students will apply their knowledge of the Central Limit Theorem in conjunction with their knowledge of probabilities of occurrence, and the Range Rule of Thumb for Identifying Significant Values and/or the Empirical Rule.

**Meaningful applications:**

1. The idea that a sampling distribution can be described in terms of shape/center/spread is a mathematically meaningful application (and generalization) of the understanding that population distribution can be described in terms of those three characteristics, because both are distributions.
2. In using graphic representations of populations in conjunction with the relevant simulation software, students will apply their understanding of graphic representations to create graphic representations for a new class of distributions, sampling distributions. Accordingly, students will refine and extend their understanding of and capacity to apply those understandings for the purposes of investigating statistical questions and using logical reasoning to make inferences based on analysis of relationships between graphic representations of populations and sampling distributions.
3. In using means and standard deviations (as well as graphic representations) of populations in conjunction with the relevant simulation software, students will apply their understanding of those parameters to compute means and standard deviation of sampling distributions. Accordingly, students will refine and extend their understanding of and capacity to apply those understandings for the purposes of investigating statistical questions and using logical reasoning to make inferences based on analysis of relationships between the pertinent means and standard deviations (and graphic representations) of populations and sampling distributions.
4. In doing parts i and ii immediately preceding this entry, students will extend application of the concept of mean and standard deviation to new contexts and thereby, refine their understandings of those terms and related concepts. Relatedly, by connecting population and sampling distribution means and standard deviations, and in the process applying their understandings of the terms parameter and statistic, it is believed that students will build a deeper and clearer understanding of the relationship and the difference between the two.
5. Students will use simulations to carry out the process of constructing a sampling distribution, and in doing so, will build a deeper, more tangible understanding of that process and the related concept of a sampling distribution.
6. See also i-viii under the preceding section devoted to actions students will perform as those also constitute meaningful applications of logical reasoning as well as the concepts, representations, and processes relevant to the activity.
7. Through applications vii and viii of the preceding section on student actions, resolution of the introductory problem coupled with simulation-based explorations and critical thinking require pupils to infer mathematical or statistical structure via pattern recognition as well as to interrelate the Central Limit Theorem, concepts of probability and statistical significance, the Range Rule of Thumb for Identifying Significant Values and potentially the Empirical Rule.
8. It is also noteworthy that in subsequent lessons, students will use conclusions of the Central Limit Theorem to compute the probability that a sample mean will lie in a given interval, which also constitute the foundation on which the reasoning and processes for constructing confidence intervals and hypothesis tests are based. This coupled with the preceding point (vii) accounts connects the tasks and explorations noted herein to virtually all of the content explored in introductory statistics courses or statistics components of quantitative reasoning courses, thus making the time invested in the activity a very worthwhile endeavor.

**Inferring the Central Limit Theorem: Exploring the Sampling Distribution of the Sample Mean**

1. A production line is set to fill juice boxes with 6 ounces of juice, and when it is functioning properly, the line produces a population of juice boxes with a mean juice content of 6 ounces and a standard deviation of .12 ounces. As part of the quality control procedure, each hour a sample of 36 randomly selected juice boxes is pulled from the end of the production line, and the mean juice content for the boxes is computed. If the sample mean is deemed to be significantly high or significantly low, the production line is shut down for maintenance and recalibration. One such sample has a mean of 6.05 ounces. Is that sample mean unusual or extreme enough, i.e., is the mean significantly high or low enough for the Quality Control Manager to order the line to be shut down for maintenance and recalibration? Explain why or why not. (If you get stuck, proceed to #2, and you will have an opportunity to revisit this problem at the end of the activity).
2. Go to the following website: <http://onlinestatbook.com/stat_sim/sampling_dist/>
3. Click on “Begin” in the upper left section of the screen.

A screenshot of a computer

Description automatically generated

1. You should now see the following on your screen.

A screenshot of a computer

Description automatically generated

What you are seeing is a set of **normally distributed population** data with a **population mean of μ = 16** and **population standard deviation of σ = 5.00**. The symbol μ is the lower-case, Greek letter mu (an equivalent for m), and it is commonly used to represent a population mean. The symbol σ is the lower-case, Greek letter sigma (an equivalent for s) and it is commonly used to represent a population standard deviation. For the purposes of this exercise, you will use μ to represent a population mean and M to represent a sample mean. Similarly, you will use σ to represent a population standard deviation and s to represent a sample standard deviation.

The above diagram has 4 sets of axes for graphs; you will be restricting our attention to the first 3 graphs.

The Clear lower 3 button at the top right of the first graph will clear the lower 3 graphs. The box below it, which currently reads normal will allow you to change the shape of the population distribution, if you click on the down arrow on the right side of the box. Later, you will change the shape of the parent distribution, but for now leave it set to normal.

For the time being, skip over the second graph and direct your attention to the third set of axes bearing the heading Distribution of Means, N = 5. The box on the upper right of the graph should be set to “Mean,” which means that the app is going to generate a histogram of sample means. You will leave that set to “Mean” for the duration of this investigation. If you collect a large number of means of samples that are randomly selected from the parent distribution, you will generate a histogram which is a very good model of the sampling distribution for the sample means. Remember, a sampling distribution is simply a data set for a sample statistic, and you can describe that distribution in terms of its center, shape and spread. **Your first goal for this investigation is to generate a model of the sampling distribution for the sample mean M, when the samples are randomly selected from a parent distribution with a known shape, mean (center) and standard deviation (spread).** You will then alter either the sample size, the shape of the parent population or both, and along with it the population mean and standard deviation may also change. **This will lead you to the principal goal of your investigation, which is to see what trends do you notice in the (mean), shape and spread (standard deviation) of the sampling distribution**.

A screenshot of a computer

Description automatically generated

Now direct your attention to the box below the box reading “Mean” on the upper right side of the third graph. The relevant box is circled in red in the above screen shot. That box should read N = 5, which means that the app is set to generate randomly selected samples of size 5 from the Parent Distribution at the top of the screen. If you left-click on the downward pointing arrow beside the five, you will see a drop-down menu of sample sizes from which you may choose. Choose 25 for the sample size.

Your app screen should now look like the following:

A graph of a normal distribution

Description automatically generated

Now direct your attention to the second set axes, entitled Sample Data. Above and to the right of the graph area is a set of buttons that are command buttons. They will cause the app to generate a given number of randomly selected samples of the size noted in the Distribution of Means (Graph 3) settings, which should currently be set to 25.

* The “Animated” button will generate 1 randomly selected sample of size N, which is currently set to 25.
* The button labeled 5 will generate 5 randomly selected samples of size N
* The button labeled 10,000 will generate 10,000 randomly selected samples of size N
* The button labeled 100,000 will generate 100,000 randomly selected samples of size N

1. Click the “Animated” button. You should see 25 black rectangles fall from the Parent population graph into the Sample Data graph, and then 1 blue rectangle should fall from the Sample Data down to the Distribution of Means graph. The 25 black rectangles which fell from the Parent population are a sample of 25 data values that were randomly selected from the population. The blue rectangle, which fell from the Sample Data graph into the Distribution of Means, is the mean of the sample data you just generated. Your top three graphs should now look something like the following:

A screenshot of a graph

Description automatically generated

*Question: Of what is the Sample Data graph an example?*

1. Click the “Animated” button 4 more times. Then click the 5 button. Your top 3 graphs should now look something like the following:

A graph of a normal distribution

Description automatically generated with medium confidence

1. Now click the 100,000 button. Your graph should now look something like the following.

A graph of a normal distribution

Description automatically generated

7A) *Of what is the blue Distribution of Means graph an example?*

7B) Now examine the shape of the blue Distribution of Means graph. *What shape does that graph appear to have, i.e., what distribution does the distribution of means at least roughly approximate?*

7C) *How does the center of the blue Distribution of Means graph compare to the center of the black Parent population graph? Are they similar or different?*

7D) Now read the value of the mean of the Distribution of Means, which is reported to the left of the graph of that graph. *How does the size of the mean of the Distribution of Means compare to the size of the mean of the parent population? That is, is the mean of the Distribution of the Means close to the mean of the parent population?*

7E) *Is the standard deviation of the Distribution of Means data 1.00?*

1. Now set the sample size N = 16 by clicking on the down arrow in the N = 25 box above and to the right of the graph and selecting 16 from the drop-down menu that appears. Your first three graph axes should now look as follows.

A graph of a graph

Description automatically generated

1. Click on the 100,000 button 10 times. Your graphs should now look something like:

A graph of a normal distribution

Description automatically generated

9A) *Does the shape of the blue Distribution of Means graph appear to be roughly normal*?

9B) *Is the mean of the Distribution of Means equal to the mean of the parent population?*

9C) *Is the standard deviation of the Distribution of Means data 1.25 or very close to it?*

1. Now set the sample size N = 2 by clicking on the down arrow in the N = 16 box above and to the right of the graph and selecting 2 from the drop-down menu that appears.

Before going further, after you click on the 100,000 button several times,

* *What do you think will be the shape of the blue Distribution of Means graph?*
* *What do you think will be the mean of the Distribution of Means?*
* Consider the following table:

|  |  |  |  |
| --- | --- | --- | --- |
| Sample Size N | 25 | 16 | 2 |
| Standard Deviation of the Distribution of Means | 1.00 | 1.25 |  |

* *Do you expect the standard deviation (the spread) of the Distribution of Means to get smaller, stay the same or get larger? Explain why or why not.*

1. Click on the 100,000 button until the Distribution of Means has a mean = 16.00 and a standard deviation of 3.54.

A graph of a normal distribution

Description automatically generated

11A) *Did your predictions hold true? What is a mathematical or statistical reason / explanation / justification for why your predictions held true or did not hold true?*

11B) *Now summarize the conclusions you believe to be true about the sampling distribution for the mean, and state the conditions under which you believe those conclusions will hold true, i.e., state the requirements that have to be met in order for your conclusions to hold true.*

Look for a pattern relating the sample size N and the standard deviation. Because the mean of the sampling distribution is related to the mean of the population, it may be that the standard deviation of the sample means is related to the standard deviation of the population; so, add a row to your table for the population standard deviation.

|  |  |  |  |
| --- | --- | --- | --- |
| Population Std. Dev. | 5.00 | 5.00 | 5.00 |
| Sample Size N | 25 | 16 | 2 |
| Std. Dev. of Sampling Distribution for Mean | 1.00 | 1.25 | 3.54 |

Now, you are searching for a pattern, so whatever ideas you produce, will need to work for all of the cases you have encountered so far and any you encounter in the future.

Think about what you could do to 5.00 to get to 1, and if possible, to relate whatever you do to the sample size N. When you have found an idea that seems to work for all 3 cases, fill in the table with computations to verify that the proposed relationship between the standard deviation of the sample means, the standard deviation of the population, and the sample size does in fact yield the values for the standard deviation of the sample means in the above table.

Ideas:

|  |  |  |  |
| --- | --- | --- | --- |
| Population Std. Dev. | 5.00 | 5.00 | 5.00 |
| Sample Size N | 25 | 16 | 2 |
| Std. Dev. of Sampling Distribution for Mean |  |  |  |

1. *Now if needed, revise the conclusions you believe to be true about the sampling distribution for the mean, and state the conditions under which you believe those conclusions will hold true, i.e., state the requirements that have to be met in order for your conclusions to hold true*.

You have made great progress, but what if the parent population is not normally distributed?

1. Click on the Clear Lower 3 button.
2. Click on the down arrow in the box reading Normal, which is above and to the right of the top or first graph, and select Skewed.
3. Check to be sure the sample size N = 25. Your top 3 graphs should now look as follows.

A graph of a parent population

Description automatically generated

Notice that the population mean (μ) is now 8.08 and the population standard deviation (σ) = 6.22

Take a moment to consider what you expect to happen based on your prior investigations. *Write down what you think will be true about the shape, mean and standard deviation of the sampling distribution for the mean*.

1. Generate a sampling distribution based on 1,000,000 samples (click 100,000 ten times). Your graphs should look something like:

A graph of a number of data

Description automatically generated with medium confidence

*Did your expectations prove to be the case? What is a mathematical or statistical reason for why your expectations about center and spread held true or did not hold true?*

1. You are going to run this simulation one more time, with a sample size of 2; so, *state what you expect to happen*:
2. Set the sample size to N = 2.
3. Generate a sampling distribution based on 5,000,000 to 8,000,000 or so samples. Your graphs should look something like the following:

A graph of a parent population

Description automatically generated with medium confidence

19A) *Did your prediction about standard deviation hold*? *What is a mathematical or statistical reason for why your prediction did or did not hold true?*

19B) *Did the mean of the sampling distribution equal the population mean of 8.08; did μM = μ = 8.08? What is a mathematical or statistical reason for why those means were equal or why they were not equal?*

19C) *Was the sampling distribution normal in shape?*

19D) Consider how this last result changes things. The sampling distribution was normally distributed or normal in shape, in all of our previous cases, including the one other case in which the parent population was not normally distributed.*So, what was different about this case from the other example involving a skewed parent population?*

*Use the trends you identified in your investigation to fill in the blanks to make true statements:*

**CENTRAL LIMIT THEOREM:**

If the parent population is \_\_\_\_\_\_\_\_\_\_, **OR** if the population is not \_\_\_\_\_\_\_\_\_\_, but the sample \_\_\_\_\_ is

\_\_\_\_\_\_\_\_\_\_ enough, then:

1. The sampling distribution for the \_\_\_\_\_\_\_\_ is \_\_\_\_\_\_\_\_\_\_\_ distributed,

or the distribution of sample \_\_\_\_\_\_\_ is \_\_\_\_\_\_\_\_\_\_\_\_.

1. µsample means = \_\_\_ or µM = \_\_\_

(The \_\_\_\_\_ of the \_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_ for the sample \_\_\_\_\_\_ = \_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_.)

1. σM = \_\_\_\_\_\_\_\_

(The \_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_ of the sample \_\_\_\_\_\_ = population \_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_ by the \_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_ of the \_\_\_\_\_\_\_\_\_ size.)

1. Now that you know the Central Limit Theorem, what it implies, and the requirements for applying it, return to and answer the original problem, which was:

A production line is set to fill juice boxes with 6 ounces of juice, and when it is functioning properly, the line produces a population of juice boxes with a mean juice content of 6 ounces and a standard deviation of .12 ounces. As part of the quality control procedure, each hour a sample of 36 randomly selected juice boxes is pulled from the end of the production line, and the mean juice content for the boxes is computed. If the sample mean is deemed to be significantly high or significantly low, the production line is shut down for maintenance and recalibration. One such sample has a mean of 6.05 ounces. Is that sample mean unusual or extreme enough, i.e., is the mean significantly high or low enough for the Quality Control Manager to order the line to be shut down for maintenance and recalibration? Explain why or why not.

**Instructor Version of Activity**

1. A production line is set to fill juice boxes with 6 ounces of juice, and when it is functioning properly, the line produces a population of juice boxes with a mean juice content of 6 ounces and a standard deviation of .12 ounces. As part of the quality control procedure, each hour a sample of 36 randomly selected juice boxes is pulled from the end of the production line, and the mean juice content for the boxes is computed. If the sample mean is deemed to be significantly high or significantly low, the production line is shut down for maintenance and recalibration. One such sample has a mean of 6.05 ounces. Is that sample mean unusual or extreme enough, i.e., is the mean significantly high or low enough for the Quality Control Manager to order the line to be shut down for maintenance and recalibration? Explain why or why not.

It is imperative that instructors require students to justify their answer, because it is via the effort to justify a proposed answer that the need to describe the center, spread and shape of the sampling distribution for the mean arises, which is a task students will probably not be able to do without knowledge of the Central Limit Theorem, which readies them to learn.

1. Go to the following website: <http://onlinestatbook.com/stat_sim/sampling_dist/>
2. Click on “Begin” in the upper left section of the screen.

A screenshot of a computer

Description automatically generated

1. You should now see the following on your screen.

A screenshot of a computer

Description automatically generated

What you are seeing is a set of **normally distributed population** data with a **population mean of μ = 16** and **population standard deviation of σ = 5.00**. The symbol μ is the lower-case, Greek letter mu (an equivalent for m), and it is commonly used to represent a population mean. The symbol σ is the lower-case, Greek letter sigma (an equivalent for s) and it is commonly used to represent a population standard deviation. For the purposes of this exercise, you will use μ to represent a population mean and M to represent a sample mean. Similarly, you will use σ to represent a population standard deviation and s to represent a sample standard deviation.

The above diagram has 4 sets of axes for graphs; you will be restricting our attention to the first 3 graphs.

The Clear lower 3 button at the top right of the first graph will clear the lower 3 graphs. The box below it, which currently reads normal will allow you to change the shape of the population distribution, if you click on the down arrow on the right side of the box. Later, you will change the shape of the parent distribution, but for now leave it set to normal.

For the time being, skip over the second graph and direct your attention to the third set of axes bearing the heading Distribution of Means, N = 5. The box on the upper right of the graph should be set to Mean, which means that the app is going to generate a histogram of sample means. You will leave that set to mean for the duration of this investigation. If you collect a large number of means of samples that are randomly selected from the parent distribution, you will generate a histogram which is a very good model of the sampling distribution for the sample means. Remember, a sampling distribution is simply a data set for a sample statistic, and you can describe that distribution in terms of its center, shape and spread. **Your first goal for this investigation is to generate a model of the sampling distribution for the sample mean M, when the samples are randomly selected from a parent distribution with a known shape, mean (center) and standard deviation (spread).** You will then alter either the sample size, the shape of the parent population or both, and along with it the population mean and standard deviation may also change. **This will lead you to the principal goal of your investigation, which is to see what trends do you notice in the (mean), shape and spread (standard deviation) of the sampling distribution**.

A screenshot of a computer

Description automatically generated

Now direct your attention to the box below the box reading Mean on the upper right side of the third graph. The relevant box is circled in red in the above screen shot. That box should read N = 5, which means that the app is set to generate randomly selected samples of size 5 from the Parent Distribution at the top of the screen. If you left-click on the downward pointing arrow beside the five, you will see a drop-down menu of sample sizes from which you may choose. Choose 25 for the sample size.

Your app screen should now look like the following:

A graph of a normal distribution

Description automatically generated

Now direct your attention to the second set axes, entitled Sample Data. Above and to the right of the graph area is a set of buttons that are command buttons. They will cause the app to generate a given number of randomly selected samples of the size noted in the Distribution of Means (Graph 3) settings, which should currently be set to 25.

* The “Animated” button will generate 1 randomly selected sample of size N, which is currently set to 25.
* The button labeled 5 will generate 5 randomly selected samples of size N
* The button labeled 10,000 will generate 10,000 randomly selected samples of size N
* The button labeled 100,000 will generate 100,000 randomly selected samples of size N

1. Click the “Animated” button. You should see 25 black rectangles fall from the Parent population graph into the Sample Data graph, and then 1 blue rectangle should fall from the Sample Data down to the Distribution of Means graph. The 25 black rectangles which fell from the Parent population are a sample of 25 data values that were randomly selected from the population. The blue rectangle, which fell from the Sample Data graph into the Distribution of Means, is the mean of the sample data you just generated. Your top three graphs should now look something like the following:

A screenshot of a graph

Description automatically generated

*Question: Of what is the Sample Data graph an example?*

Answer: A sample distribution.

1. Click the “Animated” button 4 more times. Then click the 5 button. Your top 3 graphs should now look something like the following:

A graph of a normal distribution

Description automatically generated with medium confidence

1. Now click the 100,000 button. Your graph should now look something like the following.

A graph of a normal distribution

Description automatically generated

7A) *Of what is the blue Distribution of Means graph an example?*

Answer: A sampling distribution (or at least a very good approximation or model of one)

That is, the blue Distribution of Means graph is a distribution formed from the values of a statistic that was calculated for each of a large number of samples. It is both natural and sample variance that make this happen, i.e., sampling distributions arise from the fact that a sample statistic varies from sample to sample because the samples usually do not contain exactly the same elements or elements that represent exactly the same performance with respect to or possession of the trait or variable of interest. Further, the elements themselves naturally vary with respect to their performances with respect to or possession of the trait or variable of interest. So, combining these facts, we know that if we have a sample statistic from each of a large number of samples, we can generate a very good model of the sampling distribution for that sample statistic.

7B) Now examine the shape of the blue Distribution of Means graph. *What shape does that graph appear to have, i.e., what distribution does the Distribution of Means at least roughly approximate?*

The answer of yes to the above question may not be particularly surprising since the parent population was similar. Do you think the shape of the sampling distribution would still be normal? Do not worry, we will investigate that question later.

7C) *How does the center of the blue Distribution of Means graph compare to the center of the black Parent population graph? Are they similar or different?*

They are similar.

7D) Now read the value of the mean of the Distribution of Means, which is reported to the left of the graph of that graph. *How does the size of the mean of the Distribution of Means compare to the size of the mean of the parent population? Is the mean of the Distribution of the Means close to the mean of the parent population?*

If the mean of the Distribution of Means data does not equal the mean of the parent population hit the 100,000 button 9 more times. You should now have generated somewhere in the vicinity of about 1,000,000 samples, each having a sample size of 25. *Are the means of the parent population and the distribution of means now equal?*

7E) *Is the standard deviation of the Distribution of Means data 1.00?*

If the answer to the preceding question is no, press the Clear Lower 3 button, and then press the 100,000 until the mean of the Distribution of Means is 16 and the standard deviation is 1. You probably will not have to clear your data and start over, but later in our investigation, such a step may become necessary to save time.

1. Now set the sample size N = 16 by clicking on the down arrow in the N = 25 box above and to the right of the graph and selecting 16 from the drop-down menu that appears. Your first three graph axes should now look as follows.

A graph of a graph

Description automatically generated

1. Click on the 100,000 button 10 times. Your graphs should now look something like:

A graph of a normal distribution

Description automatically generated

9A) *Does the shape of the blue Distribution of Means graph appear to be roughly normal?*

Yes.

9B) *Is the mean of the Distribution of Means equal to the mean of the parent population?*

Yes.

9C) *Is the standard deviation of the Distribution of Means data 1.25 or very close to it?*

If the answer to the preceding question is no, press the Clear Lower 3 button, and then press the 100,000 until the mean of the distribution of means is 16 and the standard deviation is 1.25.

1. Now set the sample size N = 2 by clicking on the down arrow in the N = 16 box above and to the right of the graph and selecting 2 from the drop-down menu that appears.

Before going further, after we click on the 100,000 button several times,

* *What do you think will be the shape of the blue Distribution of Means graph?*
* *What do you think will be the mean of the Distribution of Means?*
* Consider the following table:

|  |  |  |  |
| --- | --- | --- | --- |
| Sample Size N | 25 | 16 | 2 |
| Standard Deviation of the Distribution of Means | 1.00 | 1.25 |  |

* *Do you expect the standard deviation (the spread) of the Distribution of Means to get smaller, stay the same or get larger? Explain why or why not.*

Sample Explanation for Expectation Concerning Standard Deviation: The standard deviation for the Distribution of Means or the sampling distribution for sample mean should get larger because the sample size decreased, and variability tends to increase as sample size decreases.

1. Click on the 100,000 button until the Distribution of Means has a mean = 16.00 and a standard deviation of 3.54.

A graph of a normal distribution

Description automatically generated

11A) *Did your predictions hold true? What is a mathematical or statistical reason / explanation / justification for why your predictions held true or did not hold true?*

Yes.

Sample Justification for Prediction Concerning Shape: Since the parent population is normal, randomly selected samples should model the parent population reasonably well and therefore, be normal or roughly normal, especially if a large number of samples are taken.

Sample Justification for Prediction Concerning Center: Since a population mean is a measure of the center or typical performance for a population, the center of a distribution of sample means, which are all estimates of the center of the population, should be the of the center of the population, i.e., the population mean

Sample Justification for Prediction Concerning Spread: The standard deviation for the sampling distribution equal or σM = σ / √N = 5.00 / √2 = 3.54 We would expect the standard deviation of the sample means to be less than the standard deviation of the population since samples are typically less spread out than the population and the standard deviation is a measure of spread. Also, variability tends to increase as sample size decreases, so it is reasonable for the standard deviation of the means to be related to the sample size. The fact that the formula for calculating a sample’s standard deviation involves or is related to sample size may also play a role.

11B) *Now summarize the conclusions you believe to be true about the sampling distribution for the mean, and state the conditions under which you believe those conclusions will hold true, i.e., state the requirements that have to be met in order for your conclusions to hold true.*

Summaries should be similar to the following:

When the parent population was normal:

* The Distribution of Means, a.k.a., the sampling distribution of the mean was normally distributed, i.e., had the symmetric, bell-like shape of a normal curve.
* The mean of the Distribution of Means or the mean of the sampling distribution for the mean equaled the population mean.
* The standard deviation got larger as the sample size decreased, i.e., the sampling distribution for the mean got more spread out as the sample size decreased.

Look for a pattern relating the sample size N and the standard deviation. Because the mean of the sampling distribution is related to the mean of the population, it may be that the standard deviation of the sample means is related to the standard deviation of the population; so, add a row to your table for the population standard deviation.

|  |  |  |  |
| --- | --- | --- | --- |
| Population Std. Dev. | 5.00 | 5.00 | 5.00 |
| Sample Size N | 25 | 16 | 2 |
| Std. Dev. of Sampling Distribution for Mean | 1.00 | 1.25 | 3.54 |

Now, you are searching for a pattern, so whatever ideas you produce, will need to work for all of the cases you have encountered so far and any you encounter in the future.

Think about what you could do to 5.00 to get to 1, and if possible, to relate whatever you do to the sample size N. When you have found an idea that seems to work for all 3 cases, fill in the table with computations to verify that the proposed relationship between the standard deviation of the sample means, the standard deviation of the population, and the sample size does in fact yield the values for the standard deviation of the sample means in the above table.

Commonly Suggested Ideas: Subtract 4; divide by 5.

The subtract 4 idea works for the case when N=25, but not when N=16; 5-4 ≠ 1.25, and 4 does not seem to be related to the sample size. We might think to try 5-3, but 5-3 ≠ 1.25 and no other ideas related to the 4 are coming to mind, so we drop the subtraction idea.

Dividing the population standard deviation of 5.00 by 5 works when N = 25 since 5.00 / 5 = 1.00, but the dividing by 5 option fails when N = 16: 5/5 ≠ 1.25. However, 5.00/4 = 1.25.

Is 25 related to 5 and 16 related to 4 in any special way? Aha! √(25) = 5 and √(16) = 4, so now we have an idea that has worked in two cases and does in fact relate the population standard deviation, the sample size, and the standard deviation of the sampling distribution for the mean.

It appears that (Std. Dev. Of Sampling Distrib.) = (Pop. Std. Dev. / (Square Root of the Sample Size)

If this idea works, then if we apply it when the sample size or N = 2, we should get 3.54, let us see what happens.

N = 2 Pop. Std. Dev. / √N = 5.00 / √2 = 3.54 Yeah!!!! We are brilliant!

|  |  |  |  |
| --- | --- | --- | --- |
| Population Std. Dev. | 5.00 | 5.00 | 5.00 |
| Sample Size N | 25 | 16 | 2 |
| Std. Dev. of Sampling Distribution for Mean | 5.00/√25 = 1.00 | 5.00/√16 = 1.25 | 3.54/√2 |

1. *Now if needed, revise the conclusions you believe to be true about the sampling distribution for the mean, and state the conditions under which you believe those conclusions will hold true, i.e., state the requirements that have to be met in order for your conclusions to hold true*.

At this point, students should have something to the effect of:

When the parent population is normal:

* The sampling distribution is normally distributed, i.e., has a symmetric, bell-like shape.
* The mean of the sampling distribution for the mean equals the population mean.
* Standard deviation of the sampling distribution = Pop. Std. Dev. / square root of the sample size.

You have made great progress, but what if the parent population is not normally distributed?

1. Click on the Clear Lower 3 button.
2. Click on the down arrow in the box reading Normal, which is above and to the right of the top or first graph, and select Skewed.
3. Check to be sure the sample size N = 25. Your top 3 graphs should now look as follows.

A graph of a parent population

Description automatically generated

Notice that the population mean (μ) is now 8.08 and the population standard deviation (σ) = 6.22

Take a moment to consider what you expect to happen based on your prior investigations. *Write down what you think will be true about the shape, mean and standard deviation of the sampling distribution for the mean*.

Expectations should be similar to the following:

* Sampling distribution should be normal (or possibly it will match the shape of the parent population).
* The mean of the sampling distribution should equal the population mean of 8.08.
  + Or, μM = μ = 8.08
* The standard deviation of the sample means = population standard deviation / √(sample size)
  + Or, σM = σ / √N = 6.22 / √25 = 1.24

1. Generate a sampling distribution based on 1,000,000 samples (click 100,000 ten times). Your graphs should look something like:

A graph of a number of data

Description automatically generated with medium confidence

*Did your expectation prove to be the case? What is a mathematical or statistical reason for why your expectations about center and spread held true or did not hold true?*

Answers concerning center and spread should be similar to the sample explanations given for 10A. We recommend omitting an explanation for why the shape of the sampling distribution is normal because to us, that seems to much to expect introductory level students to be able to independently generate, and because the time necessary to help students construct any semblance of understanding of why the sampling distribution for the mean is normally distributed is likely to take more time than such an explanation merits in an introductory level statistics or quantitative reasoning course.

1. You are going to run this simulation one more time, with a sample size of 2; so, *state what you expect to happen*:

* Sampling distribution should be normal (or possibly it will match the shape of the parent population).
* The mean of the sampling distribution should equal the population mean of 8.08.
  + Or, μM = μ = 8.08
* Std. Dev. Of Sampling Distrib. = Pop. Std. Dev. / √(sample size)
  + Or, σM = σ / √N = 6.22 / √2 = 4.40

1. Set the sample size to N = 2.
2. Generate a sampling distribution based on 5,000,000 to 8,000,000 or so samples. Your graphs should look something like the following:

A graph of a parent population

Description automatically generated with medium confidence

19A) *Did your prediction about standard deviation hold*? *What is a mathematical or statistical reason for why your prediction did or did not hold true?*

Yes! If the answer for your data is no, you should be close to 4.40, if you add on a few more million trials, you should get there, but if not, clear the lower 3 and try again. The standard deviation for the sampling distribution equal to = 4.40 should hold, i.e., it should be the case that σM = σ/√N = 6.22/√2 = 4.40. We would expect the standard deviation of the sample means to be less than the standard deviation of the population since samples are typically less spread out than the population and the standard deviation is a measure of spread. Also, variability tends to increase as sample size decreases, so it is reasonable for the standard deviation of the means to be related to the sample size. The fact that the formula for calculating a sample’s standard deviation involves or is related to sample size may also play a role.

19B) *Did the mean of the sampling distribution equal the population mean of 8.08; did μM = μ = 8.08? What is a mathematical or statistical reason for why those means were equal or why they were not equal?*

Yes, since a population mean is a measure of the center or typical performance for a population, the center of a distribution of sample means, which are all estimates of the center of the population, should be the of the center of the population, i.e., the population mean

19C) *Was the sampling distribution normal in shape?*

No, or at least most likely not.

19D) Consider how this last result changes things. The sampling distribution was normally distributed or normal in shape, in all of our previous cases, including the one other case in which the parent population was not normally distributed. So, *what was different about this case from the other example involving a skewed parent population?*

*Answer*: The sample size was decreased.

Be sure the students do not overlook the fact that the predictions about the mean of the sampling distribution and the standard deviation of the sampling distribution held in every case, regardless of what we did to the sample size.

So, it appears that only the prediction about the shape of the sampling distribution is dependent on or influenced by the sample size. If that is the case, then we can adjust our initial predictions by including a statement about having a sample size that is large enough as a requirement for getting a sampling distribution that is normally distributed or normal in shape. Doing so yields the following, which is known as the Central Limit Theorem. It and the Empirical Rule undergird the reasoning on which all that we have done or will do in statistics in this class.

***Use the trends you identified in your investigation to fill in the blanks to make true statements:***

**CENTRAL LIMIT THEOREM:**

If the parent population is normal, **OR** if the population is not normal but the sample size but the sample size is large enough (N ≥ 30), then:

1. The sampling distribution for the mean is normally distributed,

or the distribution of sample means is normal

1. µsample means= µ or µM = µ

(The mean of the sampling distribution for the sample means = population mean.)

1. σM = σ/√N

(The standard deviation of the sample means = population standard deviation divided by the square root of the sample size.)

NOTE 1: The symbol σM is called the standard error. So, the lines of conclusion 3 can be read as:

standard error of the sample means = pop. std. dev. / square root of sample size, or to put it another way the std. dev. of the sampling distribution = pop. std. dev. / square root of the sample size.

NOTE 2: The results concerning the mean of the sample means and standard error do not depend on the sample size, but the claim concerning the normality of the sampling distribution does.

NOTE 3: If the standard deviation of the population is not known, which is usually the case, the population standard deviation can be approximated by the standard deviation of a sample. So, in practice, the 3rd conclusion of the Central Limit Theorem is modified, and σM ≈ s / √N, where Nn is the sample size and s is the standard deviation of the sample.

Note 4: Some textbooks and instructors include “If the parent population is normal” in their statements of the Central Limit Theorem, and some do not. Herein, it is included in the theorem’s statement in order to highlight all of the circumstances in which the shape component of the theorem can be concluded and remove what students may see as an “exception” to the Central Limit Theorem, but it is certainly not the only way to word the theorem.

1. Now that you know the Central Limit Theorem, what it implies, and the requirements for applying it, return to and answer the original problem, which was:

A production line is set to fill juice boxes with 6 ounces of juice, and when it is functioning properly, the line produces a population of juice boxes with a mean juice content of 6 ounces and a standard deviation of .12 ounces. As part of the quality control procedure, each hour a sample of 36 randomly selected juice boxes is pulled from the end of the production line, and the mean juice content for the boxes is computed. If the sample mean is deemed to be significantly high or significantly low, the production line is shut down for maintenance and recalibration. One such sample has a mean of 6.05 ounces. Is that sample mean unusual or extreme enough, i.e., is the mean significantly high or low enough for the Quality Control Manager to order the line to be shut down for maintenance and recalibration? Explain why or why not.

Answer: Yes, the mean is unusual or extreme enough for the Quality Control Manager to order the production line to be shut down for maintenance and recalibration. This is because the *Range Rule of Thumb for Identifying Significant Values* indicates that if a distribution of a quantitative variable X has a mean of μx and a standard deviation of σx, then a data value is considered significantly high if it is more than 2 standard deviations above the mean of the distribution and is considered significantly low, if the data value is more than 2 standard deviations below the distribution mean. The sample mean of 6.05 oz can be considered a data value from the sampling distribution or the mean, i.e., a data value for the distribution of sample means (for samples of 36 boxes). Since the sample size is greater than or equal to 30, we can apply the Central Limit Theorem. From the Central Limit Theorem, we know the mean of the sampling distribution equals the population mean (µM = µ = 6 oz), or to put it another way, from the Central Limit Theorem, we know the mean of the distribution of sample means equals the population mean of 6 oz. By the Central Limit Theorem, we also know that the standard error or standard deviation of our distribution of sample means equals population standard deviation divided by the square root of the sample size, or σM = σ/√N = .12/√36 = .02 oz.

So, applying the *Range Rule of Thumb for Identifying Significant Values* yields that:

A sample mean will be significantly high if: > µM + 2 σM = 6 oz + 2 (.02 oz) = 6.04 oz.

A sample mean will be significantly low if: < µM - 2 σM = 6 oz - 2 (.02 oz) = 5.96 oz.

Due to all of the above, the sample mean of 6.05 oz is greater than 6.04 oz, so that sample mean is significantly high enough or unusually high or extreme enough for the Quality Control Manager to shut down the production line for maintenance and recalibration.

Alternatively, we could also conclude that by the Central Limit Theorem, the sampling distribution of the mean is normally distributed, so we can apply the Empirical Rule to that distribution (to the distribution of sample means, for samples of 36 boxes). Now, by the Empirical Rule we know that 95% percent of all sample means will lie within 2 standard deviations of the distribution mean of µM (95% of sample means lie in the interval determined by µM ± 2 σM = 6 oz ± 2 (.02 oz) or the interval (5.96 oz, 6.04 oz). Since, 6.05 falls outside the interval we know that if the production line were operating properly, the probability p of getting a sample mean of 6.05 is less than or equal to p =.05 or 5% (p ≤ 5%), which is small enough to allow us to conclude that the sample mean of 6.05 is significantly higher than 6 oz, so again, the sample mean is significantly high enough for the Quality Control Manager to shut down the production line for maintenance and recalibration. Moreover, since the sampling distribution for the mean is normal, we can use the fact that normal distributions are symmetric, which would allow us to conclude that the probability of getting a sample mean above the µM ± 2 σM or (5.96 oz, 6.04 oz) interval equals the probability of lying below it, so the probability of getting a sample mean above the (5.96 oz, 6.04 oz) interval equals .05/2 = .025 = 2.5%. In light of that, we could even say that if the production line was operating properly, the probability p of getting a sample mean of 6.05 oz is less than or equal to .025 or 2.5% (p ≤ .025 = 2.5%).