

the population size N.

Math Objectives

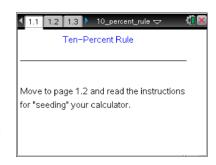
- Students will recognize that when samples are taken from a finite population without replacement the sampling distribution of sample means has a smaller standard deviation than that for samples of equal size taken with replacement.
- Students will understand that calculating the standard deviation of the sampling distribution of sample means using the formula $\sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}}$ applies reasonably well for samples taken without replacement provided the sample size n is no larger than 10% of
- (Optional/Extension) Students will understand that when samples are taken from a finite population without replacement the actual standard deviation of the sampling distribution of sample means is adjusted by multiplication by the Finite Population Correction Factor, giving $\sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}} \sqrt{1 \frac{n-1}{N-1}}$.
- Students will use appropriate tools strategically (CCSS Mathematical Practices).

Vocabulary

- Confidence interval
- Hypothesis test
- Finite Population Correction Factor
- Mean
- Normal distribution
- Population
- Sample
- Sampling distribution
- Standard deviation

About the Lesson

 This lesson involves investigating the differences between the standard deviations of sampling distributions of means for samples taken from finite populations with and without replacement.



TI-Nspire™ Technology Skills:

- Download a TI-Nspire document
- · Open a document
- Move between pages
- · Grab and drag a point

Tech Tips:

- Make sure the font size on your TI-Nspire handhelds is set to Medium.
- You can hide the function entry line by pressing ctrl
 G.

Lesson Files:

Student Activity

10_Percent_Rule_Student.pdf

10_Percent_Rule_Student.doc

TI-Nspire document

10 Percent Rule.tns

Visit <u>www.mathnspired.com</u> for lesson updates and tech tip videos.

- As a result, students will:
 - Confirm that the "textbook formula" for the standard deviation
 of the sampling distribution of sample means is exactly
 correct when samples are selected with replacement.
 - Observe that the standard deviation for sampling distributions of sample means is consistently smaller than that given by the textbook formula when samples are selected without replacement.
 - Conclude that the textbook formula for the standard deviation
 of the sampling distribution of sample means is reasonably
 correct for samples taken without replacement provided
 samples are no larger than ten percent of the population size.
 - Confirm that the factor $\sqrt{1-\frac{n-1}{N-1}}$ corrects the discrepancy between the actual standard deviation of the sampling distribution of sample means for samples without replacement and the value given by the formula for samples with replacement.

TI-Nspire™ Navigator™ System

- Transfer a File
- Use Screen Capture to compare different graphical displays.
- Use Quick Poll to check student understanding.

Prerequisite Knowledge

- Familiarity with sampling distributions.
- The ability to use probability formulas to calculate the mean and standard deviation of the sampling distribution of sample means when sampling with replacement.

Related Activities

- Sampling Distributions
- Standard Error of Sample Means

Discussion Points and Possible Answers

Teacher Tip: Teacher Tips at questions 3b, 7, 13 and 16 are marked "Vital Discussion" and should not be skipped.

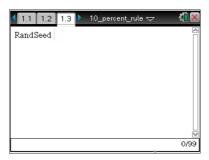
Tech Tip: Page 1.2 gives instructions on how to seed the random number generator on the handheld. Page 1.3 is a *Calculator* page for the seeding process. Ensuring that students carry out this step will prevent students from generating identical data. (Syntax: RandSeed #, where # is a number unique to each student.).

Teacher Tip: Once students have seeded their random number generators, they do not have to do it again unless they have cleared all of the memory. But it is important that this be done if the memory has been cleared or with a new device, as otherwise the "random" numbers will all be the same as those on other similarly cleared devices.

Formal statistical inference—hypothesis tests and confidence intervals—on means requires values for both the mean and the standard deviation of the sampling distribution of the sample mean. We will consider how the standard deviation of the sampling distribution of sample means change when samples are taken without replacement instead of with replacement.

Move to pages 1.2 and 1.3.

This activity involves generating a number of random samples from a population. In order to avoid having your results be identical to those for another student in the room, it is necessary to "seed" the random number generator. Read the instructions on Page 1.2 for seeding your random number generator, then carry out that seeding on Page 1.3.



Move to page 2.1.

1. Page 2.1 displays the calculations of the **mean** and **standard deviation** of a population containing exactly six elements, namely {3, 6, 9, 12, 15, 18}. Record these two values for use throughout this activity.

Note: The displayed mean is exact, but the standard deviation has



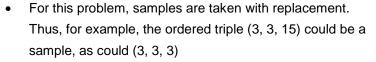
been rounded to three decimal places.

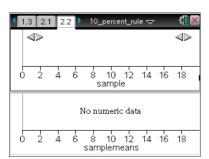
Answer: mean = 10.5, standard deviation = 5.123...

Teacher Tip: It might be necessary to remind students that the usual calculation of a population standard deviation involves division by N, the population size, instead of N-1, which is used for a sample. You might wish to have students use a Scratchpad page to verify these values.

Move to page 2.2.

The arrows on the left side of Page 2.2 will generate samples of size three from the population on Page 2.1. (The arrows on the right side of the page reset the page, erasing all samples, in case you want to start over.)





Teacher Tip: Answers to the following questions might differ because each student is generating a different set of random samples. You might want to discuss answers to 2b in order to make that clear to students. For 2c, you might have students use the Scratchpad for their calculations.

2. a. Click once on an arrow on the left side of Page 2.2 to generate a sample of size three from the given population, chosen with replacement. What do the vertical line in the upper plot and the single dot in the lower plot represent?

Sample Answers: It appears that they represent the mean of the sample.

 Click about 10 more times on the same arrow to generate more random samples of size three from the given population. Describe the resulting plots.

<u>Sample Answers:</u> The top plot shows a sample with a vertical line marking the mean. The lower plot shows the distribution of all the sample means that have been obtained so far. It seems to be centered about 10, is spread a bit less than the population, but I'm not sure of the shape after so few samples.



c. Recall that the sampling distribution of a statistic is the distribution of that statistic from all possible samples. You have only seen a few values from the sampling distribution of means of samples of size three from the given population, so your distribution in the lower work area of Page 2.2 is very incomplete. Use your answers from Question 1 and formulas from earlier in the course to calculate the exact mean and standard deviation of the complete sampling distribution of sample means of samples of size three.

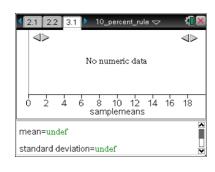
Answer: The mean of the sampling distribution is the same as the population mean, so it is 10.5. The standard deviation of the sampling distribution of means is $\frac{\sigma}{\sqrt{n}}$, so it is about $\frac{5.123}{\sqrt{3}}$, or 2.958.

Teacher Tip (Vital Discussion): Discuss student answers to Question 3b to ensure that students understand that the dotplot on Page 3.1 displays the evolving sampling distribution of sample means of samples of size three from the given population.

Teacher Tip: If students have studied counting principles you might want to ask them between Questions 3b and 3c to determine exactly how many samples of size three can be taken from the given population.

Move to page 3.1.

Page 3.1 continues building the sampling distribution of sample means from samples of size three from the given population. Again, samples are taken with replacement, so triples such as (3, 3, 15) might occur. Also, as before, the arrows on the left side of the page generate samples and the arrows on the right side of the page reset so that you can start over.



3. a. Click the arrow once, and describe the information displayed in the plot.

<u>Sample Answers:</u> When I clicked once, 12 dots appeared in the plot. Since we are using samples of size three, those dots must be the means of 12 different samples. Below the plot are values for a mean and standard deviation. They must be the mean and standard deviation for the set of 12 dots displayed in the plot.

b. Choose the highest or lowest of the displayed values (dots) in the plot on Page 3.1. State the displayed value, and identify at least one triple (sample) that could have produced it. (You might want to refer back to question 1 to recall the population from which the samples are being drawn.)

<u>Sample Answers:</u> My plot contains the value 3. The only way to obtain that value would be from the sample (3, 3, 3).

My plot's lowest value is 4. That means the sum of the three values was 12. That could have come from (3, 3, 6) or (3, 6, 3) or (6, 3, 3), but those are the only ways possible.

c. Compare the mean and standard deviation displayed below the dot plot to the values you calculated in Question 2c and explain any differences you see.

<u>Sample Answers:</u> The displayed mean is 10.833, and the displayed standard deviation is 2.85287. Both of these are close to what I calculated, but neither is exactly correct. However, they are based on only 12 samples, not an entire sampling distribution, so a difference is not surprising.

- 4. a. Continue clicking the arrow on Page 3.1 until the entire sampling distribution of sample means is displayed. Describe the shape of the sampling distribution, and explain how you know that it is not a normal distribution.
 - You will know the entire distribution is displayed when repeatedly clicking the arrow does not change the distribution anymore.

<u>Sample Answers:</u> The shape is very symmetric and mound-shaped. However, it cannot be normally distributed since its minimum and maximum values are 3 and 18, respectively; and only integer values are possible as sample means. The normal distribution, on the other hand, is a continuous distribution defined on all real numbers.

b. How do the mean and standard deviation of the sampling distribution compare to those you calculated in Question 2c?

<u>Sample Answers:</u> They agree exactly, at least for as many decimals as are displayed —the mean is 10.5 and the standard deviation is 2.958.

5. a. Consider the three or four most extreme sample means (dots) displayed on Page 3.1. Identify the samples from which they came.

<u>Answer:</u> The lowest value is 3. It had to have come from (3, 3, 3). Likewise, 18 had to come from (18, 18, 18). The next lowest value is 4, and the only samples that have that mean consist of two 3s and a 6, and there are three such triples, namely (3, 3, 6), (3, 6, 3), and (6, 3, 3).

b. For each sample that you identified in part a, what fraction of the entire population was included in that sample?

Answer: The most extreme sample, i.e. (3,3,3), contains only 1/6 of the population values. The next most extreme samples, each consisting of two 3s and one 6, contain 1/3 of the population.

6. a. Suppose now that the samples from the population were taken without replacement. That is, no element of the population is allowed to appear more than once in any given sample. Describe how you think that change in sampling procedure would change the extreme values in the sampling distribution of sample means.

<u>Sample Answers:</u> Since it would no longer be possible to get samples like (3, 3, 3), the lowest possible mean would be larger, coming from the three smallest members of the population—3, 6, and 9—giving a mean of 6. Similarly, the largest possible mean would now become 15.

b. What fraction of the population is included in any sample of size three taken without replacement?

<u>Sample Answers:</u> Since each sample contains exactly three values and no values might be repeated, each sample contains exactly 1/2 the entire population.

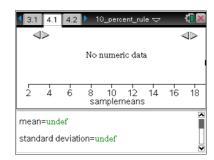
7. Standard deviation indicates something about the differences among members of the distribution it describes. Based on your work in Questions 5 and 6, discuss differences in the standard deviation of the sampling distribution of sample means that you expect between samples taken with replacement and samples taken without replacement.

<u>Sample Answers:</u> The most extreme values in the sampling distribution for sampling without replacement are closer together than those for sampling with replacement, making me think the standard deviation for sampling without replacement will be lower too. Also, since samples without replacement always include half the entire population, they can differ from each other less than those taken with replacement, which can consist of just a single repeated number. So, again I suspect that the standard deviation for 'without replacement' will be lower.

Teacher Tip: (Vital Discussion) Be sure that all students understand the reasoning behind the sample answer to question 7. They might need to see examples written out to be sure they see what is happening.

Move to page 4.1.

Page 4.1 again builds a sampling distribution of sample means from samples of size three from the same population. However, this time samples are taken without replacement, so triples such as (3, 3, 15) can never occur. Note that only the means of the samples are displayed, not the elements that make up each sample.



8. a. Click the arrow on Page 4.1 until the entire sampling distribution of sample means is displayed. Describe the shape of the sampling distribution, and explain how you know that it is not a normal distribution.

<u>Sample Answers:</u> The shape is very symmetric and mound-shaped. However, it cannot be normally distributed since its minimum and maximum values are 6 and 15, respectively; and only integer values are possible as sample means. The normal distribution, on the other hand, is a continuous distribution defined on all real numbers.

b. How do the mean and standard deviation of the sampling distribution compare to those you calculated in Question 2c?

<u>Sample Answers:</u> The mean agrees exactly, but the actual standard deviation is 2.29129, which is lower than what I calculated.

c. How do the standard deviations from the sampling distributions on Pages 3.1 and 4.1 compare? Are the results consistent with what you predicted in Question 7?

<u>Sample Answers:</u> The standard deviation for samples without replacement is 2.29129, which is a bit lower than the standard deviation for samples with replacement, namely 2.95804. That is what I predicted in Question 7.

The idea of **residual** runs through all of statistics. Remember that a residual is the difference "Observed – Expected." Here we can think of the actual standard deviation of the sampling distribution

as the "Observed" value and the value predicted by the formula $\sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}}$ as the "Expected" value. Since a difference of five units can be "large" or "small" in different contexts, one useful way to judge

the size of a residual is to express it as a "relative difference" or fraction of the Expected value. That is, **Observed - Expected**

compute Relative Difference = Expected

9. Compute the relative difference for the standard deviation of the sampling distribution you generated on Page 4.1, and interpret what it means.

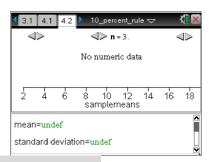
<u>Answer:</u> The relative difference is approximately -0.225. That means that the actual standard deviation was over 22% lower than what the formula said it should be.

10. Based on your answers in Questions 8 and 9, name at least one advantage and one disadvantage to sampling without replacement.

<u>Sample Answers:</u> One advantage is that variability in the sampling distribution of sample means is reduced, so sample means are generally closer to the population mean. A disadvantage is that the formula for calculating the standard deviation does not give the right answer when sampling without replacement.

Move to page 4.2.

Page 4.2 behaves exactly as Page 4.1 did, except that now the center set of arrows will change the sample size to values between 1 and 6 inclusive. Samples are again taken without replacement.



Teacher Tip: If students have studied combinations, you might want to have them calculate how large each sampling distribution will be before they begin work with Page 4.2.

11. a. Record the standard deviation information from Page 4.1 into the table below.

Answer:

Sample Size	SD by Formula	Actual SD	Relative Difference
1	5.12348	5.12348	0.0
2	3.62284	3.24037	-0.106
3	2.95804	2.29129	-0.225
4	2.56174	1.62019	-0.368
5	2.29129	1.0247	-0.553
6	2.09165	0.0	-1.000

b. Set the sample size to a value other than 3, then generate the corresponding sampling distribution of sample means. Record the standard deviation information for that distribution in the table above.

Answer: See table entries above.

c. Complete the summary table for sample size and standard deviation of sampling distributions
by choosing other sample sizes and generating their corresponding sampling distributions.

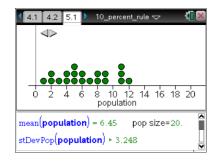
Describe how sample size is related to the "agreement" of the actual standard deviation of

sampling distributions and the standard deviation predicted by the formula

<u>Answer:</u> See table entries. In general, as sample size goes up, the agreement between formula and reality gets worse. The actual standard deviation decreases more and more than predicted as sample size increases.

Move to page 5.1.

12. Page 5.1 displays a new, larger population. The arrow allows you to change to a new population. Using the table below, record descriptions of each possible population.

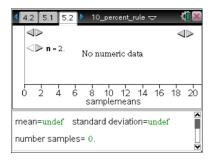


Answer:

Population Size	Mean	Standard Deviation	Shape
20	6.45	3.248	approx uniform
40	4.575	1.909	skew right
60	16.15	1.833	skew left
80	9.6875	5.387	bimodal
100	9.48	3.324	slight skew left

Move to page 5.2.

Page 5.2 behaves exactly as Page 4.2 did. The population from which samples will be taken is the one you selected on Page 5.1.



Teacher Tip: (Vital Discussion) Discuss student answers to Question 13. Ensure that students understand that the sampling distributions they have seen in previous pages are complete and exact, but those for the remainder of the activity will be only approximate due to the very large number of samples possible from each population.

13. a. Page 5.2 generates samples without replacement and plots the means of those samples. Thus, as before, you will build a simulated sampling distribution. However, now the population is larger, so it will no longer be possible to generate all possible samples. That is, you will have only an approximate sampling distribution. How do you think this will affect your standard deviation calculations?

<u>Sample Answers:</u> Since I will be missing some (maybe many) samples, my calculated values will probably differ a little from those predicted by the formulas.

b. In carrying out this simulation, what might we do to improve our results?

<u>Sample Answers:</u> If I can collect lots of samples, maybe the results will be pretty close to what would occur with a full sampling distribution.

Tech Tip: : Each complete simulated sampling distribution on Page 5.2 will take approximately a minute to run on the handheld device. If you have convenient access to a computer lab, using the TI-Nspire software will speed up the investigation.

Teacher Tip: Since each simulated sampling distribution will take some time for students to generate, the activity is set up to have the investigation shared across groups. An ideal set-up would have each group responsible for exactly one population size, requiring five (5) groups.

At this point, your teacher will split your class up into groups and assign each group a specific population to investigate. Within your group, divide up the work however you think will be most efficient.

Tech Tip: Sample means will be calculated and displayed in groups of 100 at a time. Because of the amount of calculation going on in generating those samples, there will be several seconds of delay before your display is updated. Be sure to watch for the plot to update before your next click. Be patient.

- 14. Return to Page 5.1, and use the arrow to set the population to the one assigned to your group.
 - Then move back to Page 5.2, and use the lower arrow on the left side of the page to select a sample size for which to generate the simulated sampling distribution. (Each person in your group can use a different sample size.)
 - Click the upper arrow on the left side of the page to generate sample means.
 - Continue until you have 1500 samples. (Remember, wait for the plot to update between clicks.)
 - Record the information for your group's simulations in the table below. Describe any patterns you see.

<u>Sample Answers:</u> See tables below (each table represents a possible investigation of one particular population size, corresponding to one group of students). In general, the relative difference becomes more pronounced as the sample size increases, with actual SD becoming smaller and smaller with respect to what the formula predicts.

Pop Size	Sample Size	SD by Formula	Simulated SD	Real Difference
20	2	2.2966	2.21	-0.03774
20	3	1.8752	1.694	-0.09664
20	5	1.4525	1.301	-0.10433
20	7	1.2276	1.024	-0.16587
20	10	1.0271	0.7339	-0.28546
20	15	0.8386	.04216	-0.94972

Pop Size	Sample Size	SD by Formula	Simulated SD	Real Difference
40	2	1.3498	1.364	0.01047
40	3	1.1021	1.07	-0.02918
40	4	0.9545	0.9125	-0.04400
40	5	0.8537	0.8098	-0.05145
40	10	0.6036	0.5346	-0.11442
40	20	0.4268	0.315	-0.26206

Pop Size	Sample Size	SD by Formula	Simulated SD	Real Difference
60	2	1.2961	1.297	0.00067
60	5	0.8197	0.7863	-0.04079
60	6	0.7483	0.7119	-0.04866
60	7	0.6928	0.6602	-0.04706
60	10	0.5796	0.5349	-0.07719
60	20	0.4098	0.3467	-0.15412

Pop Size	Sample Size	SD by Formula	Simulated SD	Real Difference
80	2	3.8091	3.834	0.00651
80	7	2.0360	1.964	-0.03540
80	8	1.9045	1.796	-0.05701
80	9	1.7956	1.671	-0.06942
80	15	1.3909	1.272	-0.08549
80	25	1.0774	0.9162	-0.14961

Pop Size	Sample Size	SD by Formula	Simulated SD	Real Difference
100	2	1.108	2.408	0.02449
100	9	1.0511	1.089	-0.01714
100	10	1.0022	0.9996	-0.04903
100	11	0.8582	0.9293	-0.07276
100	15	0.6648	0.8051	-0.06193
100	25	1.108	0.5742	-0.13628

TI-Nspire Navigator Opportunity: *Screen Capture* See Note 1 at the end of this lesson.

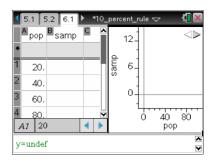
15. Using the results of your group's work in Question 14 (investigating additional sample sizes if needed), determine the largest sample size for which the simulated standard deviation and the "formula" standard deviation differ by no more than 5%.

<u>Sample Answers:</u> (Refer to sample answer tables in question 14.) Our group was working with a population of 60 numbers. The largest sample size we were able to use and keep the SD close enough to the formula value was n = 7. (Results will depend on population size and will vary due to simulation variability, but in general sample size should be about 1/10 the population size.)

Teacher Tip: Coordinate the collection of all the class results from Question 15, ensuring that each group reports only the sample size that "just barely" stayed within 5% of the formula value for standard deviation for entry on Page 6.1.

Move to page 6.1.

16. a. The left work area of Page 6.1 shows two columns of a spreadsheet. Collect the class results from Question 15 into this page, entering the appropriate sample size in the cell next to the population from which it was obtained. Exactly what does the first pair of values (20, ___) tell you about sampling distributions?



<u>Sample Answers:</u> The pair is (20, 2). That means that for samples larger than n = 2, sampling without replacement will result in a standard deviation for the sampling distribution of sample means that will be smaller than $\sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}}$ by more than 5%.

b. Describe the plot in the right work area.

Sample Answers: The scatterplot appears pretty linear with a positive slope.

c. Click the arrow in the scatterplot to display the graph and equation of the least squares regression line for your class data. Interpret that equation and relate your interpretation to the title of this activity.

<u>Sample Answers:</u> The equation has a slope of 0.095. That's about 0.1. So we could say that the sampling distribution of sample means for samples taken without replacement but smaller than about 10% of the population size will have actual standard deviations within about 5% of that calculated from the formula $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$.

Teacher Tip: (Vital Discussion) Question 16 is the "punch line" for this activity. Ensure that students understand exactly what the entries of their class table mean. Summarize their answers to 16c as the usual textbook "10% Rule."

Teacher Tip: Note that samples larger than 10% of the population size are better than expected (less variable), not worse. So violations of the 10% Rule lead, for example, to confidence intervals that capture the true mean more often than demanded by the stated confidence level because they are longer than necessary (computed margin of error is too large).

Extension

Theory says that the standard deviation of the sampling distribution of sample means for samples

taken without replacement is actually $\sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}} \sqrt{1 - \frac{n-1}{N-1}}$

This is the usual formula but with an additional factor: the quantity $\sqrt{1-\frac{n-1}{N-1}}$ is called the **Finite Population Correction (FPC)** factor.

17. Return to your data table in Question 14, and verify the accuracy of the new formula by carrying out the indicated calculations and comparing the results to your simulation data.

<u>Sample Answers:</u> The new formula works pretty well. It still misses in some cases, but not in a systematic way. I think that might be because the simulation does not include all possible samples and is only approximate.

18. Use algebra on the FPC factor to find an exact rule for when the relative difference in standard

deviations will be more than 5% of $\sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}}$

Answer: Setting the FPC ≥ 0.95 and solving gives the relationship:

 $n \le 0.0975*N+0.9025$

for the critical sample sizes.

TI-Nspire Navigator Opportunity: Quick Poll

See Note 2 at the end of this lesson.

Wrap Up

Upon completion of the lesson, the teacher should ensure that students are able to understand:

- When samples are taken from a finite population without replacement the sampling distribution of sample means has a smaller standard deviation than that for samples of equal size taken with replacement.
- Calculating the standard deviation of the sampling distribution of sample means using the formula $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$ applies reasonably well for samples taken without replacement provided the sample

size n is no larger than 10% of the population size N.

Assessment

Suppose the population from which you are selecting consists of 100 rectangles and the statistic of interest is the mean area of the rectangles. How would you calculate the standard deviation of the sampling distribution of sample means if you wanted to investigate samples of

1. size 20 with replacement? Is it exact?

Sample Answers: Use
$$\sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}}$$
. It is exactly correct.

2. size 10 without replacement? Is it exact?

Sample Answers: Use
$$\sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}}$$
. It is not correct but is a good approximation.

3. size 30 without replacement? Is it exact?

Sample Answers: Use
$$\sigma_{\overline{x}} = \frac{\sigma_x}{\sqrt{n}} \sqrt{1 - \frac{n-1}{N-1}}$$
. The shorter formula is not correct and is a poor approximation.

TI-Nspire Navigator

Note 1

Name of Feature: Screen Capture

A Screen Capture can be used to compare sampling distributions and calculations of standard deviations for a variety of populations and sample sizes.

Note 2

Name of Feature: Quick Poll

Quick Poll can be used to check for student understanding throughout the activity. The suggested assessment questions could be used to check for understanding at the end of the activity.