Math Objectives

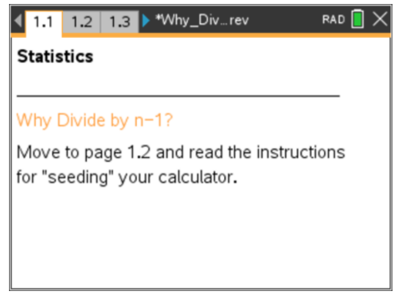
- Students will recognize that the mean of all the sample variances for samples of a given size drawn with replacement calculated using *n-1* as a divisor will give the population variance calculated using *N* as a divisor.
- Students will recognize that the mean of all the sample variances for samples of a given size drawn without replacement calculated using *n-1* as a divisor will give a value equal to the population variance if calculated using *N-1* as a divisor.
- Students will recognize that as the sample size increases the difference in calculating the variances for a sample using *n* and *n-1* as a divisor will decrease.
- Students will recognize that when a sample variance is calculated using *n*-1 as a divisor, the sample variance produces an unbiased estimator of the population variance.
- Students will construct viable arguments (CCSS Mathematical Practices).

Vocabulary

- bias
- mean
- parameter
- population
- sample size
- sampling distribution
- standard deviation
- variance

About the Lesson

- This lesson involves investigating calculating a sample variance using both *n* and *n-1* as the divisor for samples drawn with and without replacement.
- As a results, students will:
  - Investigate the sampling distribution of the variances of all possible samples of size three, drawn *with* and *without* replacement from a given population, using as divisors both *n* and *n-1*.
  - Compare the means of these distributions to the variance



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 10ˣ G.

Lesson Materials:
Student Activity
Why_Divide_by_n-1.pdf
Why_Divide_by_n-1.doc

TI-Nspire document
Why_Divide_by_n-1.tns

Visit www.mathnspired.com for lesson updates and tech tip videos.

for the population.

- Investigate the impact of sample size on the variance calculated using each of the divisors.

Teachers often focus on sampling with replacement, which would omit question 3 in this activity and its associated .tns page; however, a discussion about sampling without replacement (question 3) is necessary in order to tell the whole story.

Prerequisite Knowledge

Students should know what standard deviation is and how to calculate one and be familiar with sampling distributions. Experience with the Statistics Nspired activity Investigating Standard Deviation should precede this activity.

TI-Nspire™ Navigator™ System

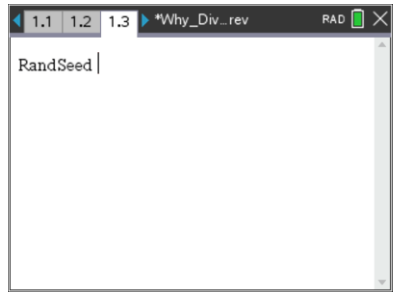
- Use *Live Presenter* to have students explain what the data points and vertical lines in the displays represent.
- Use *Screen Capture* to compare different sampling distributions.
- Use *Quick Poll* to assess student understanding at the end of the activity.

Discussion Points and Possible Answers

Tech Tip: Page 1.2 gives instructions on how to seed the random number generator of the Nspire. Page 1.3 is a calculator page for the seeding process. Carrying out this step will ensure that everyone does not produce identical data as they work through the activity. (Syntax: RandSeed #, where # should be a number unique to each student.)

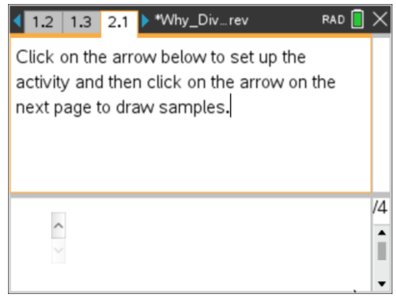
Tech Tip: When students move the cursor over a point in a dot plot, the data value associated with the dot is displayed. Clicking on a vertical line in this activity will display the value plotted on the horizontal axis. To de-select a point, click in a white space on the screen.

One of the goals of selecting a sample from an unknown population is to estimate parameters of that population. One of these parameters is variance; the square root of the variance produces the standard deviation. This activity investigates the difference in estimating the population variance calculated using a divisor of *n-1* versus *n*.



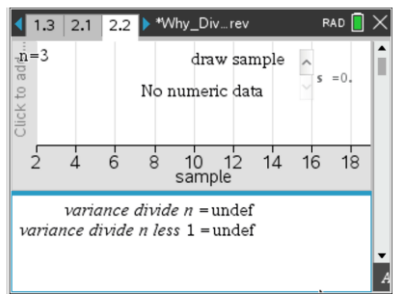
Move to page 2.1.

Click on the arrow to begin the activity.



Tech Tip. Students will generate samples during the activity. The arrow on Page 2.1 will set up the activity to ensure that not all the samples students produce will be the same.

Move to page 2.2.



1. The arrow generates a sample of size three, with replacement, from an unknown but fixed population. Generate a sample and note the sample variances calculated by dividing by *n* and by *n-1*.
   a. How do the two values differ and why?

   <u>Sample Answers:</u> The variance calculated using *n-1* as the divisor will be larger than the one calculated using *n* because *n-1* is smaller than *n,* and when you divide by a smaller number you get a larger result.

Teacher Tip. Students do not have enough information about the population to answer the following question yet. The purpose is to motivate their thinking about how to make an inference from a sample.

b. Record the values from question a in the table below. Then generate three more samples, and record the sample values and variances in the table. Which method do you think will give a better estimate for the population variance and why? (Remember that variance is population squared.)

Sample Answers:

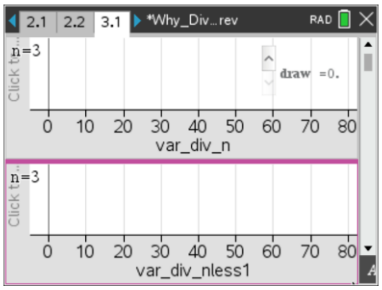
Sample values	Variance (*n* divisor)	Variance (*n-1* divisor)
3, 15, 15	**24**	**36**
9, 12, 15	**6**	**9**
15, 18, 18	**2**	**3**
6, 6, 12	**8**	**12**

Some students might choose dividing by *n-1* because the value is always larger, and it would be safer to find a large value for the variance. Others might note that it is difficult to make a good estimate from just a few samples.

Teacher Tip. The activity is designed to study methods of calculating estimators in an environment in which we know the "answer," so we can judge how well the different methods perform. The knowledge we learn from this investigation allows us to use methods that consistently seem to perform "best" in real situations where we do not know what the answer is.

Move to page 3.1.

To determine whether one estimator is better than another requires knowing the population and investigating the behavior of the estimators across all possible samples. The population from which the samples in question 1 were drawn is {3,6,9,12,15,18}. The population variance, which is calculated with *N*=6 as a divisor, is 26.25.



2. The arrow will generate eight possible samples of size three, drawn with replacement from the population and calculate the variance of each sample. Those variances, calculated using both divisors, are plotted in the dot plots on page 3.1. Use the arrow to generate a set of eight samples.

> **Tech Tip:** Note that when the cursor is over a data point, the value associated with the point is shown and the corresponding point, calculated using the other divisor, is highlighted in the second plot. To see the value associated with that point, it is necessary to move to that panel and hover over that point.

a. What does a data value in each plot represent? Give an example to illustrate your answer.

Sample Answers: A data value in the upper plot represents the variance calculated using *n* as a divisor for one sample of size three drawn with replacement. A data value in the bottom plot represents the variance calculated using *n-1* as a divisor for the same sample. For example, in one set of eight samples of size three, one of the samples has a variance calculated using n as a divisor of about 50, while the corresponding variance calculated using *n*-1 as a divisor is about 75.

b. Make a conjecture about what the vertical line in each plot represents. Then check your conjecture by inspecting the data values in each plot.

Sample Answers: The vertical line in the each plot represents the mean of the eight sample variances, calculated by using the indicated divisor. For example, in one set of eight samples using *n* as a divisor, the variance were 0, 2, 2, 2, 8, 18, 32 and 50; the vertical line represents the mean, 14.25, of those sample variances. Using *n-1* as a divisor, the corresponding sample variances were 0, 3, 3, 3, 12, 27, 48 and 75 for a mean of 21.375.

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

c. Continue to generate groups of samples until you have 104 samples, observing what happens to the means of the sample variances calculated with both methods. Describe how the two means compare.

Sample Answers: The mean for the variances calculated by dividing by *n* is consistently lower than the mean calculated by dividing by *n-1*. The mean of the variances calculated by dividing by *n* seems to vary around 20, while the mean calculated by dividing by *n-1* seems to vary around about 28.

d. Continue to generate samples until you have the sampling distribution of all 216 possible sample variances, calculated using both divisors, for samples of size three drawn with replacement from the given population. How is it possible that some samples have variances of 0 for both methods?

<u>**Answer:**</u> When all values in the sample are identical, for example {3,3,3}, the variances will be found by dividing 0 by either *n* or *n*-1. In both cases, the result will be 0.

> **Teacher Tip:** Some students might wonder why there are 216 possible samples of three elements from the given population. It is 6^3 because there are six choices for each of the three elements in the sample, but the focus of the question is not on the number of possible samples.

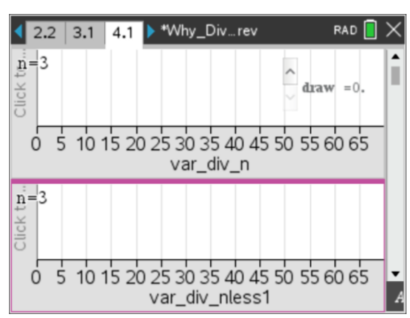
e. Use shape, center, and spread to compare the sampling distributions for the sample variances found by the two methods. (Remember that the population variance is 26.25)

<u>**Sample Answers:**</u> The sampling distributions have the same shape, both are skewed right, but the sampling distribution using *n-1* as a divisor is spread out (dilated) by a factor of 1.5 (the ratio of the divisors *n*=3 to *n-1* =2) to produce the sampling distribution for the divisor *n-1,* which thus has a longer tail to the right. The distribution using *n* as the divisor has a range of 50, from 0 to 50, while the sampling distribution using *n-1* as a divisor has a larger range, 75 (3/2 of 50), from 0 to 75. The mean of the sampling distribution using *n-1* as a divisor is 26.25, which is the population variance. The mean of the sampling distribution of sample variances calculated using a divisor of *n* is 19.5 and, as samples were generated, was consistently smaller than the population variance.

> **Teacher Tip:** Note that the samples involved in question 2 were all of the possible samples of size three, chosen *with* replacement, for the given population. The samples generated in the next set of questions beginning on Page 4.1 are generated *without* replacement, which decreases the number of possible samples of size three to 20.

Move to page 4.1.

In many cases, a sample is drawn without replacement because, for a variety of reasons, it is impossible to replace the sample elements. The following questions investigate how to estimate the population variance when a sample is drawn without replacement.



Teacher Tip: Ask students to identify situations where it is impossible to sample with replacement and those where it is possible. For example, testing the life of light bulbs would not be a good setting for sampling with replacement because the sample is destroyed. Calling the same telephone number twice in a telephone survey would be sampling with replacement.

Teacher Tip: Note this activity is about sampling distributions of sample variances. It asks the question, "Under what conditions does the sample variance, on average, produce the population variance?" This question is not related to the Finite Population Correction factor (FPC) or the 10% rule for sampling from a population, which deal with the sample mean within the sampling distribution of sample means. The FPC corrects the calculation of the variance of the sampling distribution of sample means when sampling without replacement. Why a correction factor is needed is not the same question as examining the conditions in which the sample variance, on average, produces the population variance.

The arrow will generate all possible samples of size three, drawn without replacement from the population in Question 1. The sample variances calculated using both divisors are plotted in the dot plots on Page 4.1. Remember if the divisor were *N-1*, or 5, the result would be 31.5. The population variance, which is calculated with *N*=6 as a divisor, is 26.25.

3. a. Make a conjecture about which method of calculating the variance for a sample will give the best estimate of the population variance. Explain your reasoning.

   <u>Sample Answers:</u> Some students might suggest that using *n-1* as a divisor is more likely to give a good estimate for the population variance because dividing by *n-1* will result in larger values when the sum of the squared deviations from the mean does not equal zero.

   b. Use the arrow to generate the sampling distribution of the sample variances, and observe what is happening to the means of the sample variances in each of the distributions. Describe what you observed.

   <u>Sample Answers:</u> The mean of the sample variances using *n* as a divisor is always less than the mean calculated using *n-1* as a divisor and goes a bit above 25 but ends at 21. The mean using *n-1* as a divisor goes above 40 for some samples but ends at 31.5.

   c. Use shape, center, and spread to compare the sampling distributions for the variances found by the two methods and to check your conjecture in 3a. (Remember that the population variance, which is calculated with *N*=6 as a divisor, is 26.25. If the divisor were *N-1*, or 5, the result would be 31.5.)

Sample Answers: The sampling distributions of the sample variances have the same shape, both skewed right, but the sampling distribution using *n* as a divisor is dilated by a factor of 1.5 (the ratio of the divisors *n*=3 to *n-1* =2) to produce the sampling distribution for the divisor *n-1*. The distribution using *n-1* as the divisor has a range of 54, from 9 to 63, while the sampling distribution using *n* as a divisor has a smaller range, 36, from 6 to 42. The mean of the sampling distribution using *n-1* as a divisor is 31.5, which is the value that would result if the population variance were calculated using *N*-1 as a divisor. The mean using *n* as a divisor is 21, which is lower than the population variance and seems to have no relation to it.

4. An *unbiased estimator* for a population parameter is one for which the average of the values of the sample statistic from all possible sample values equals the value of the population parameter. A biased estimator for a parameter is one whose values on average over- or under- estimate the true value of the parameter.

   a. When sampling with replacement, which estimator, the variance calculated by dividing by *n* or the variance calculated by dividing by *n-1*, will be an unbiased estimator for the population variance? Explain your reasoning.

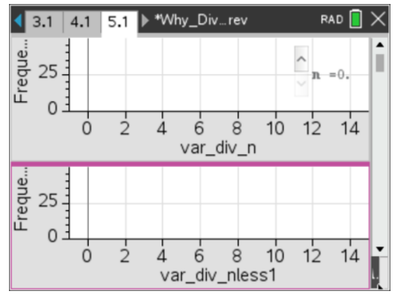
   Sample Answers: The variance calculated using *n-1* as a divisor is unbiased when sampling with replacement because on average, it will give you the population variance. Dividing by *n* to calculate a sample variance will consistently underestimate the population variance.

   b. Lin claims that the *n-1* sample variance is better than the *n* sample variance. Do you agree with her? Why or why not?

   Sample Answers: Lin is probably correct because calculating sample variance using *n-1* as a divisor will produce an unbiased estimate of the population variance.

> **Teacher Tip:** Note that the samples involved in questions 2-5 were all of the possible samples, chosen without replacement, of the given size for the given population. The samples generated in the next set of questions beginning on page 6.1 are generated *with* replacement, which increases the number of possible samples of size three to 216.

Move to page 5.1.



Using the arrow will select a sample size, *n,* and the display will show 100 random samples generated from a normal population with mean 10 and standard deviation 2 for the sample size you selected. The screens display the simulated sampling distributions of the sample variances calculated for each sample size using divisors of *n* and of *n-1*.

5. a. If the standard deviation is 2, what is the variance?

<u>Answer:</u> The variance is 4, the square of the standard deviation.

> **Teacher Tip:** The samples generated on Page 7.1 are randomly selected from a normal population and represent only some of the possible samples. The displayed simulated distribution of the 100 sample variances provides information about the distribution of all possible variances from that normal population for the given sample size. Thus, the distributions of the sample variances will vary from student to student. The move from the finite to an infinite population makes the issue of sampling with or without replacement unimportant.

TI-Nspire Navigator Opportunity: *Live Presenter*
See Note 2 at the end of this lesson.

b. Make a conjecture about how the means of the two simulated sampling distributions of sample variances will compare as the sample size increases. Explain why your conjecture seems reasonable.

<u>Sample Answers:</u> The difference between the means of the two distributions will probably get smaller because as *n* gets larger, dividing by *n* or *n-1* will not make as big a difference in the results as it does when *n* is small. For example, dividing 200 by 4 gives 50; dividing by 3 gives 66 and 2/3, but dividing 200 by 50 is 4 and dividing by 49 is about 4.08.

c. Use the arrow to choose a sample size of 10. Click on the vertical line through the mean in each display and note the values of the means of the two simulated sampling distributions. Repeat this process for at least three other sample sizes greater than 40. Do the results support your conjecture in part a? Why or why not?

Sample Answers: The results support my conjecture; in each case, the mean of the distribution of the variances calculated by dividing by *n-1* was close to 4, the population variance. But as the sample size increased, the mean of the distribution of the variances calculated by dividing by n also became close to 4. In one set of simulations, when *n* was 10, the means differed by approximately 0.5; when *n*=40, the means of the two distributions for the variances differed by 0.36. When *n*=100, the means of the two distributions differ only by 0.04. For *n*=100, the distributions look nearly identical.

Teacher Tip: Note that hovering over the bins in a histogram will display the number of values in each bin.

TI-Nspire Navigator Opportunity: ***Screen Capture***
See Note 3 at the end of this lesson.

6. Describe how *sample size* seems to affect the difference in approximating the population variance using a divisor of *n* or a divisor of *n-1*.

Sample Answers: As the sample size increases, the difference in approximating the population variance using a divisor of *n* and using a divisor of *n-1* decreases. For large enough *n*, the difference is very small. Since calculating sample variance using *n-1* always matches either the population calculation based on n or the one based on *n-1*, and these are almost identical for reasonable sized populations, the *n-1* sample calculation is "always" appropriate.

Wrap Up

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

- The mean of all the sample variances for samples of a given size drawn with replacement and calculated using a divisor of *n-1* will give the population variance
- The mean of all of the sample variances for samples of a given size drawn without replacement calculated using *n-1* as a divisor will give a value that would result if the population variance were calculated using *N-1* as a divisor
- Using a divisor of *n-1* to calculate the variance for a sample of size *n* produces an unbiased estimate of the population variance
- Calculating sample variance using *n-1* always matches either the population calculation based on *N* or the one based on *N-1*, and because these are almost identical for reasonable sized populations, the *n-1* sample calculation is "always" appropriate
- As the sample size increases the difference in calculating the variances for a sample using *n* and *n-1* as a divisor will decrease

Assessment

1. When estimating the population variance from sample data, do you divide by *n* or by *n-1* in calculating the sample variance? Justify your answer.

   Answer: When calculating the sample variance, dividing by *n-1* will always give you an unbiased estimator of the population variance if you sample with replacement..

2. Describe how sample size seems to affect the bias in approximating the population variance when using a divisor of *n* instead of a divisor of *n-1*.

   Answer: The bias decreases as the sample size increases.

3. Does sampling with or without replacement make a difference in obtaining an unbiased estimate of the population variance? Why or why not?

   Answer: Yes. Sampling with replacement gives an unbiased result, and the bias in sampling without replacement becomes very small when samples are large.

TI-Nspire Navigator

Note 1

Question 1b, *Live Presenter*

Use Live Presenter to have several students generate samples and explain what they are generating; each data point represents a sample variance and the vertical line represents the mean of the eight sample variances.

Note 2

Question 5a, *Live Presenter*

You might want to use Live Presenter again to have someone generate samples and to explain what the elements in the display represent; the histogram displays the sample variances for 100 samples of size 10. The vertical line is the mean of those sample variances. Hovering over a bin will display the endpoints of the bin and the number of sample variances in that bin.

Note 3

Question 5c, *Screen Capture*

Use Screen Capture to compare different sampling distributions for some subset of n = 10 to 100. Group those with the same sample size and discuss how these are alike or different; then compare them across sample size. Call attention to the fact that the sampling distributions across the class become more and more similar as the sample size increases.