Why $n - 1$?
An Exploration through Simulation

One of the most often asked—and most difficult to answer—questions in AP Statistics is, "Why do we divide by $n - 1$ when we calculate the standard deviation and variance of a sample?" This activity will explore that question through a simulation.

Let us quickly review the manner in which variance and standard deviation are calculated.

Variance of a population: $$s^2 = \frac{\sum (x - m)^2}{n}$$

Standard Deviation of a population: $$s = \sqrt{\frac{\sum (x - m)^2}{n}}$$

Variance of a sample: $$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

Standard Deviation of a sample: $$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

It is very rare, if ever, that one knows the parameters of a population, mean m, variance s^2, and standard deviation s. (If we did, we would not need the statistic!) Therefore one uses the sample statistics $\bar{x}$, s^2, and s to estimate them. One wants the best estimate possible, meaning one that, on average, equals the population parameter. In this activity, you will explore how the behavior of several estimates compares with their real population values.

Exploration 1: Sampling from a Known Population

Procedure

A. Write each of the digits 0 through 9 on slips of paper 5 times. This set of 50 slips of paper is your population.

B. Place the population in a box, bag, or some other container, mixing the slips thoroughly.

C. Calculate the mean, variance, and standard deviation of the population.

D. Draw one of the slips from the box, record its value, and replace it in the box, mixing thoroughly.

E. Repeat Step D until you have recorded 5 values ($n = 5$). This group of 5 values is your sample from the population.

F. Calculate the mean of the sample.

G. Calculate the variance of the sample using both formulas, s^2 and s^2.

H. Calculate the respective standard deviations.

I. Repeat Steps D–H until you have done six samples.

Sample Number	Sample Data ($n = 5$)	Mean	Variance s^2	Variance s^2	Std. Dev s	Std. Dev s
1						
2						
3						
4						
5						
6						

J. Combine your data with the rest of your class. Make a histogram or dotplot for each sample statistic. Locate the centers of the distributions.

- Where is the center of the distribution of sample means? How does it compare with the mean of the population?
- Where are the centers of the distributions of sample variances, calculated with n and $n - 1$? How do they compare with the variance of the population?
- Where are the centers of the distributions of sample standard deviations, calculated with n and $n - 1$? How do they compare with the standard deviation of the population?
- What conclusions can you draw?

Why $n - 1$?
Exploration 2: Speeding Up the Sampling Process

In Exploration 1, you gathered a fair number of samples and analyzed them. The process was a bit cumbersome, however. Writing the 0's through 9's on the slips of paper, physically drawing them, recording them, replacing them, then repeating the process was as tedious as reading this extraordinarily long and involved sentence. One can speed up the process through use of a random digit table, which you will use such in this activity.

Step back for a moment to Exploration 1 and consider a different approach. If you had written 0 through 9 only once, thus creating only ten slips of paper instead of fifty, how would this have affected your samples? Would you have had any more or less of a chance of drawing, say, a 6? Does this population of ten slips have a different mean, variance, and standard deviation from the population of fifty?

Now, suppose someone had a hat containing 10 slips of paper, the digits 0 through 9 recorded on the slips once each. The person sat for hours drawing a slip, recording the digit, replacing the slip, mixing, and repeating the process. This was done until the slips of paper were a pile of lint. The result is known as a random digit table. The first line of such a table is shown below.

63157 00354 57651 14793 76682 89281 84037 46926 92345 47280

Because the digits in the table were created in the same manner as randomly drawing from the population of ten paper slips, the digits' distribution is the same as sampling from the population of slips. Therefore, selecting a single digit from the random digit table is equivalent to drawing a single slip from our population.

Use the random digit table to speed up the sampling process. Each digit in a random number table represents a selected member of the population.

Procedure

A. Select a starting digit in your random digit table and record the value. Your teacher may tell you a specific place to start.

B. Move right to the next digit and record it. If you reach the end of a line, move down to the next line.

C. Repeat Step B until you have recorded 5 digits.

D. Calculate the mean of the sample.

E. Calculate the variance of the sample using both formulas, s^2 and s^2.

F. Calculate the respective standard deviations.

G. Repeat Steps B–F until you have done siz samples.

Sample Number	Sample Data ($n = 5$)	Mean	Variance s^2	Variance s^2	Std. Dev s	Std. Dev s
1						
2						
3						
4						
5						
6						

H. Combine your data with the rest of your class and the data from Exploration 1.

I. Make a histogram or dotplot for each sample statistic. Locate the centers of the distributions.

- Where is the center of the distribution of sample means? How does it compare with the mean of the population?
- Where are the centers of the distributions of sample variances, calculated with n and $n - 1$? How do they compare with the variance of the population?
- Where are the centers of the distributions of sample standard deviations, calculated with n and $n - 1$? How do they compare with the standard deviation of the population?
- What conclusions can you draw?

Why $n - 1$?
An Exploration through Simulation
Teacher Notes

Since the variance of a population is unknown, we must estimate it in order to do any inference. The question becomes, "What is the best estimator for the population variance?" This is answered in detail at http://phywww1.ncssm.edu/green/Math/Stat_Inst/Notes.htm, under Theory of Inference, Student's t-Test, beginning on page 7. In a nutshell, dividing by $n - 1$ **provides** a sample variance that is an unbiased estimator of s^2 and dividing by n does not. A formal definition of bias is given:

> A statistic with mean value equal to the value of the population characteristic being estimated is said to be an **unbiased statistic.** A statistic that is not unbiased is said to be **biased.** (Peck, Olsen, Devore, 2001)

Students sometimes feel the need for validating that dividing by $n - 1$ is the correct approach. The proof is beyond the scope of AP Statistics and students do not always accept this explanation on faith. In fact, dividing by $n - 1$ seems counter-intuitive to many statistics students—and their teachers!

Since the topic of dividing by $n - 1$ vs. n comes up early in the course, students may not have done many simulations. This activity is designed to expose students to various tools of simulation—concrete objects, random digit tables, and technology—gradually progressing from less abstract to more abstract methods.

In the commentary that follows, the various simulation techniques are discussed. Projected results and possible trouble spots are also addressed.

Exploration 1: Sampling from a Known Population

In this exploration, students sample using concrete objects, i.e. drawing slips of paper from a box. The simulated population consists of 50 pieces of paper, 5 each having the digits 0 through 9. Students will sample from the population, with replacement, until they get a sample of size 5. Each will repeat this process six times.

While a population of ten pieces having the digits 0–9 occurring once would be equivalent to the "five in fifty" method, the larger population size seems more logical to students at first. They often are puzzled why one would sample from a population of only 10 members. Also, we sample with replacement here so that successive observations are independent. Sampling without replacement can produce satisfactory results if the population size is very large, but such a population is cumbersome to create using hands-on methods.

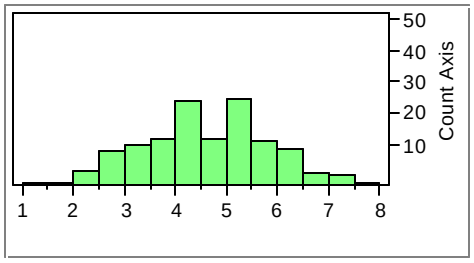
The mean, variance, and standard deviation of the population are $m = 4.5$, $s^2 = 8.25$, $s \approx 2.872$, respectively.

A typical student's result may look something like this:

Sample Number	Sample Data ($n=5$)	Mean	Variance denom = n	Variance denom = $n-1$	Std. Dev denom = n	Std. Dev denom = $n-1$
1	10674	3.60	7.44	9.30	2.73	3.05
2	87656	6.40	1.04	1.30	1.02	1.14
3	87396	6.60	4.24	5.30	2.06	2.30
4	82455	4.80	3.76	4.70	1.94	2.17
5	01703	2.20	6.96	8.70	2.64	2.95
6	52629	4.80	6.96	8.70	2.64	2.95

Below are the histograms and selected summary statistics of the samples collected by a simulated class of 30 students. JMP-INTRO was used to analyze the data.

Mean



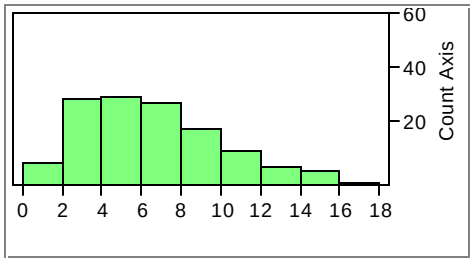
Quantiles

50.0%	median	4.6000

Moments

Mean	4.513333

Var n



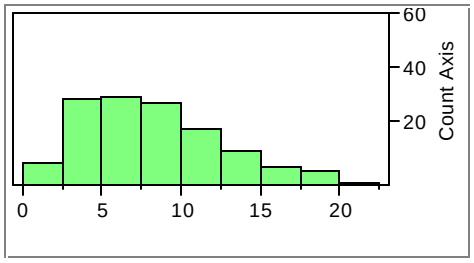
Quantiles

50.0%	median	6.160

Moments

Mean	6.610133

Var n-1



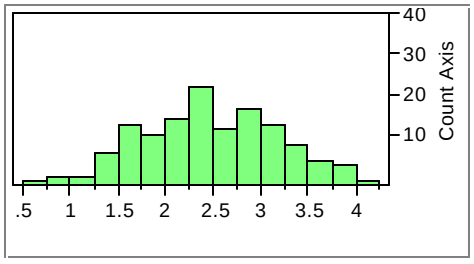
Quantiles

50.0%	median	7.700

Moments

Mean	8.262667

SD n



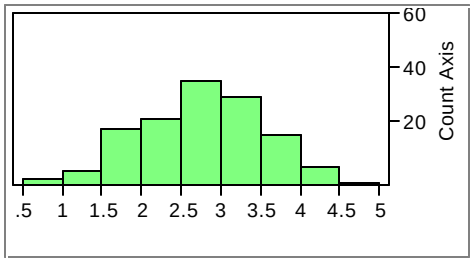
Quantiles

50.0%	median	2.4819

Moments

Mean	2.473532

SD n-1



Quantiles

50.0%	median	2.7749

Moments

Mean	2.765493

Students should notice that the center of the distribution of means is very close to the true mean of 4.5.

Looking at the variances when we divide by n, the center is somewhere between 6.2 and 6.6. This range is much farther from the population variance of 8.25 than the center of the distribution when we divide by $n - 1$, which is somewhere between 7.7 and 8.3.

A similar observation can be made for the standard deviation. The concern is that both are too low when compared with the true standard deviation of approximately 2.872. In fact, neither dividing by n nor $n - 1$ produces an unbiased estimate of standard deviation. Though s^2 is an unbiased estimator of s^2, the square root of s^2 is *not* an unbiased estimator of the square root of s^2.

When evaluating the data, the "centers" we are interested in the means of the sample statistics, not the medians. In the simulated class above, the mean of the sample means is 4.513, very close to the population mean of 4.5. The means of the sample variances using the formulas for s^2 and s^2 are 8.263 and 6.610, respectively. The formula for s^2 produces results much closer to the population variance of 8.25 than does the formula for s^2.

Both the mean and median have been provided here to give you a sense of what to expect from students when they describe the "centers" of their distributions. If class data can be combined in an efficient manner—perhaps into a common graphing calculator—the means of the sample statistics can be quickly calculated and the point driven home more quickly.

Also, it is sometimes difficult to see the differences in the means from a hastily made histogram or dotplot. If number of runs of the simulation is small, then distinction is less clear. The more runs that students can do, the better. Later parts of this activity address this concern.

Exploration 2: Speeding up the Sampling Process.

This exploration is designed to introduce students to random digit tables as a means to generate data faster and without concrete materials. If students have not had much exposure to probability, you may have to guide them through the narrative on their worksheet. It is important that students understand three points before beginning the simulation process.

1. That a population of fifty slips with the digits 0 through 9 represented five times each has the same parameters as a population of ten slips with the digits 0 through 9 represented once each.

2. That a random digit table is essentially created by sampling digits, with replacement, from a population of ten slips with the digits 0 through 9 represented once each. (It is more likely created by generating random digits in a computer. This can be discussed at the conclusion of the activity.)

3. That selecting digits from the random digit table is the same as sampling digits, with replacement, from a population of ten slips with the digits 0 through 9 represented

once each. Also, the digit table *is* random, and further some further randomization scheme does not matter (like skipping digits or lines or starting in "random" places on the table).

If students are using the same random digit table from their text, or provided by you, it is important that they all begin in a different place. One method would be to have them begin at the row number that corresponds to their birth month and the digit on that row that corresponds to their birth date. For example, a student born on July 23 would begin at the 23rd digit of the 7th row. (Beware of classes with twins, triplets, etc.)

Optional Exploration: Automating the Process

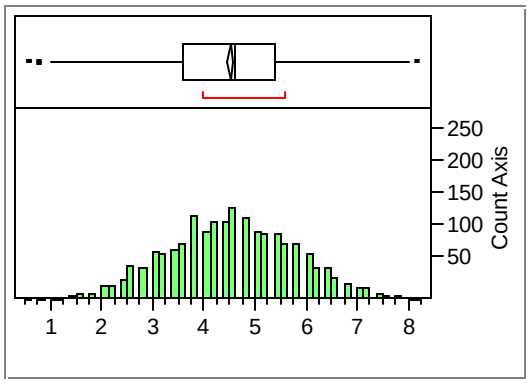
The next step is to automate the sampling process. Graphing calculators and statistics software packages allow for the creation of random data (actually pseudo-random) very quickly and in large numbers. Data can be stored in lists on graphing calculators, or columns in software data tables, for quick analysis.

For increasingly large sample sizes, the difference between dividing by $n - 1$ and n converges to zero. This can be seen in the results of a single sample of size 1000 shown below. Recall, $m = 4.5$, $s^2 = 8.25$, and $s \approx 2.872$.

Mean = 4.541
Var (n) = 8.2343
Var ($n - 1$) = 8.2426
SD (n) = 2.8696
SD ($n - 1$) = 2.8710

For small sample sizes, differences are more apparent after larger numbers of simulations. Shown below is a JMP-INTRO analysis of 2,000 samples of size 5. A boxplot where the mean is indicated by a diamond is added to each histogram to assist in seeing the means of the distributions. Since we are only interested in the mean of the distribution, the other characteristics of the boxplot—median, quartiles, etc.—are not relevant.

Distributions
Mean



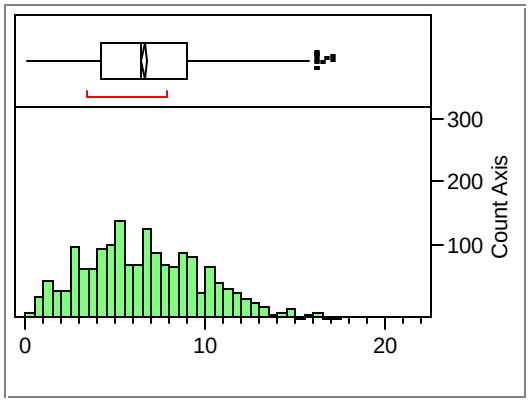
Moments

Mean	4.517800
N	2000

The center of the distribution of means is almost exactly the population mean of 4.5. When examining the variance, it is very clear that the sample variance calculated with n is biased—it is systematically too low, while the variance calculated with $n - 1$ has a mean very close to $s^2 = 8.25$.

Distributions

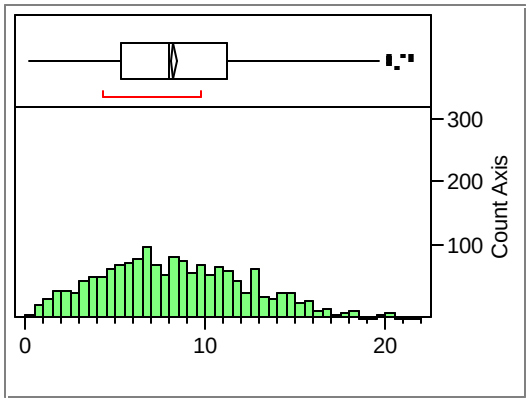
Var n



Moments

Mean	6.618920
N	2000

Var n-1



Moments

Mean	8.273650
N	2000

Final Thoughts

Students may wonder why we discuss variance. While statistical analysis utilizes the standard deviation as the statistic to describe spread, the theoretical framework uses the variance.

The conclusions from this Exploration are: (1) dividing by $n - 1$ rather than n makes the sample variance an unbiased estimator for population variance, and (2) s^2 is an unbiased estimator of s^2, but the square root of s is *not* an unbiased estimator of s.

And if you are still worried about this fact, consider, in closing, the MINITAB analysis of 2000 runs of samples of size 1000. Recall, $m = 4.5$, $s^2 = 8.25$, and $s \approx 2.8723$.

```
Variable        N     Mean
Mean         2000   4.4985
Var n        2000   8.2419
Var n-1      2000   8.2502
SD n         2000   2.8706
SD n-1       2000   2.8720
```

For further investigation of the behavior of standard deviation and variance, refer to the activity *The Algebra of Random Variables: An Exploration through Simulation.*