

Describing Populations Statistically: The Mean, Variance, and Standard Deviation

BIOLOGICAL VARIATION

One aspect of biology that holds true for almost all species is that not every individual is exactly the same. In other words, measurements (values) for a given biological trait (a variable) vary from individual to individual. For example, several black cherry trees of the same age and in the same habitat may vary considerably in their heights or number of branches. We could measure all of these cherry trees and calculate a mean, or average, for height and thus describe our cherry tree population. We might even collect data on the heights of cherry trees from several different habitats, calculate the mean tree height in each habitat, and thus have a way to compare habitats.

Biological variation has both genetic and environmental causation, both of which are of great importance to biologists. It is the existence of genetic variability, for example, that allows natural selection to operate and populations of organisms to evolve.

SOME IMPORTANT STATISTICS

The Mean

A variety of techniques exist to summarize and compare sets of numerical values. Summarizing techniques permit the quick and effective communication of information. For instance, you could report your data on heights of cherry trees by listing all of the separate height measurements. At the least this would be boring and inefficient. Or you could summarize part of the information in those separate measurements by calculating their **mean** or **average**:

$$\text{mean} = \bar{x} = \frac{\sum x}{n}$$

where $\sum x$ is the sum of all the individual observations (x 's) and n is the number of observations. For five trees with heights of 2, 3, 4, 5, and 6 meters, the mean height is 4 meters; but for five trees with heights of 24, 27, 28, 29, and 32 meters, the mean height is 28 meters. These means identify a central or roughly typical value for tree heights in the two sets of data and therefore summarize the overall magnitude of values. The mean is said to be a measure of **central tendency**.

The Variance

Consider two data sets of tree heights in meters:

Data set 1: 2, 8, 12, 16, 22

Data set 2: 11, 11, 12, 13, 13

The means for both of these sets is 12 meters but the amount of variation around the mean is much greater in the first set than in the second. A person interested in

characterizing these two populations would clearly require more information than the mean alone. A statistic that measures the amount of spread in values around the mean in a data set is called the **variance**:

$$\text{variance} = s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

where Σ is the sum of all values for the parenthetical expression, x is each individual observation, \bar{x} is the mean of all observations, and n is the number of observations in the sample. TAKE A CLOSE LOOK AT THE EQUATION FOR THE VARIANCE. Notice that it is indeed a measure of variability- it literally says: “the variance is the **average** (sum divided by $n-1$) **squared** (the exponent) **deviation** of each observation from the mean ($x-\bar{x}$).” Notice also that as the deviation of the observed values from the mean ($x-\bar{x}$) increases, so does the variance. Be sure you can intuitively see what the equation says.

An easy way to keep your calculations organized when calculating the variance is to set up a table. Suppose you measured five trees with the following heights (all in meters): 4, 4.8, 6.2, 6.9, 8.1. The variance for this population is calculated as follows:

In the first column, calculate the mean:
 $\bar{x} = 30.0/5 = 6.0$ meters

In the second column, calculate the deviation of each observation from the mean

In the third column, calculate the square of each deviation and then sum those squares

x	$x - \bar{x}$	$(x - \bar{x})^2$
4.0	-2.0	4.0
4.8	-1.2	1.44
6.2	0.2	0.04
6.9	0.9	0.81
8.1	2.1	4.41
$\Sigma=30$		10.70

Now divide the sum of the squares (10.70 in this case) by $n-1$ (or 4 in this case). The variance of this tree population is $10.70/4 = 2.68\text{m}^2$. The variance is always expressed as the square of the units of measurement that have been employed. In the example above, the variance units are m^2 .

The Standard Deviation

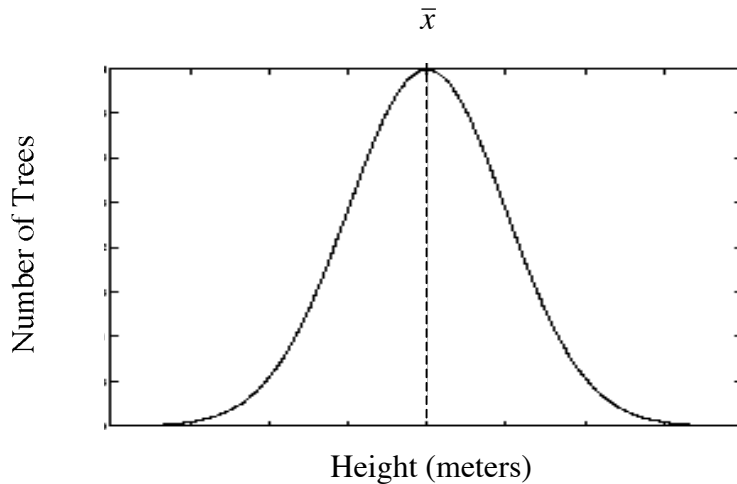
In scientific papers, the measure of variability for a population is usually expressed in an unsquared state so that it is in the same units as the observations that were made. The square root of the variance is called the **standard deviation**:

$$\text{standard deviation} = s = \sqrt{s^2} = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}; s = \sqrt{2.68} = 1.64\text{m}$$

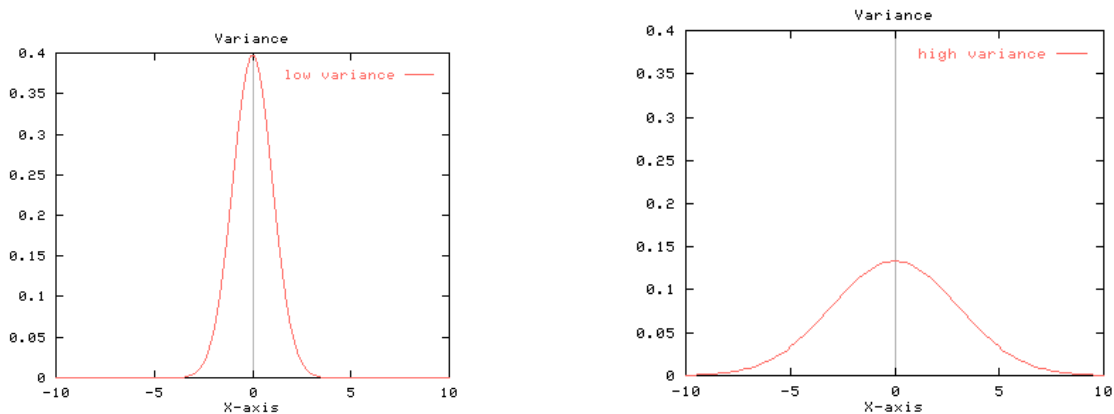
A given set of data is commonly summarized and expressed as the mean plus and minus the standard deviation ($\bar{x} \pm \text{SD}$). For the above data set, this would be 6.0 ± 1.6 meters

THE NORMAL DISTRIBUTION

If individual values deviate from the mean in a random manner, and we usually assume that they do, then a typical set of data will look like this in a frequency distribution plot:



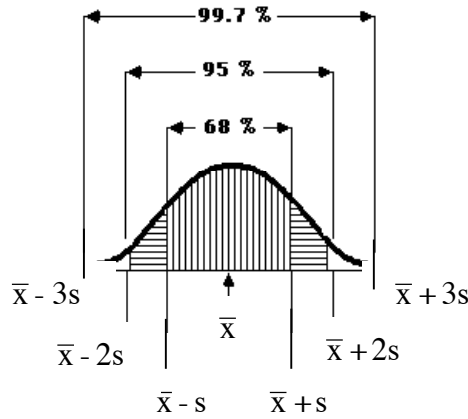
This is what is called a bell-shaped curve or normal distribution. Notice in this type of distribution, the majority of individuals have values near the mean, and fewer are at the extremes. Roughly the same number of individuals have values below the mean as have values above the mean. Normal curves can differ in shape due to differences in their variances. In the graphs below, the population depicted on the left has a lower variance (less spread) than the population on the right (more spread):



In all normal curves, the mean \pm one standard deviation will include 68.3% of the data points (i.e., 68.3% of the total area under the curve), or about $2/3$ of the data; the mean \pm 2 standard deviations includes 95.4% of the data, and the mean \pm 3 standard deviations includes 99.7% of the data. Expressing the data as the mean \pm one standard deviation (e.g., 6.0 ± 1.6 meters) enables one to mentally construct a normal distribution of the

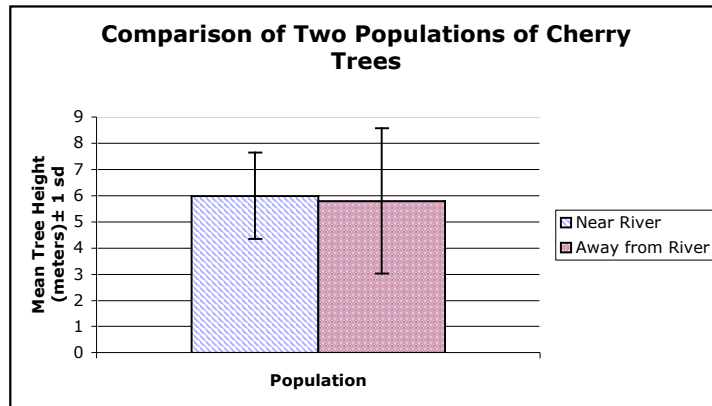
data. In the example on tree heights, we know that 68% of the trees in our data set have a height that is between 4.4 and 7.6 meters.

Normal Distribution



Bar graphs

Often in biological studies we wish to compare values of a certain trait in two different populations. For example say we want to compare height of cherry trees in a population near a river (population 1) and a population away from a river (population 2). We collect data from each population and calculate the mean and standard deviation for the trees in each population. To display this data graphically, we could produce a separate frequency distribution for each population as above. However, if we want to place the data side by side for comparison, we can create a bar graph displaying the mean and standard deviation:



From this graph, we can see that the population near the river has a slightly higher mean tree height, but the population away from the river has a higher variance so it must have a wider range of tree heights (short and tall).

THE T-TEST

Looking at the graph above, there appears to be a slight difference between the two means of the populations. However, do you think this is a significant difference? If you look at the error bars you can see that they overlap. That means that there is variance around each mean, so some individuals of similar size are found in each population.

To determine if two populations really have significantly different means, we can perform a statistical test called the Student's t-test. The t-statistic was invented by William Sealy Gosset for cheaply monitoring the quality of beer brews. "Student" was his pen name. Gosset was statistician for Guinness brewery in Dublin, Ireland, hired due to Claude Guinness's innovative policy of recruiting the best graduates from Oxford and Cambridge for applying biochemistry and statistics to Guinness's industrial processes. In simple terms, the t-test compares the actual difference between two means in relation to the variation in the data (expressed as the standard deviation of the difference between the means).

The t-test will calculate what is called the t-statistic using the formula:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2 + s_2^2)/n}}$$

From this t-statistic, we can determine whether the two populations are statistically different. To test the significance, you need to set a risk level (called the alpha level). In most social research, the "rule of thumb" is to set the alpha level at .05. This means that five times out of a hundred you would find a statistically significant difference between the means even if there was none (i.e., by "chance"). You also need to determine the degrees of freedom (df) for the test. In the t-test, the degrees of freedom is the sum of the persons in both groups minus 2. Given the alpha level, the df, and the t-value, you can look the t-value up in a standard table of significance (available as an appendix in the back of most statistics texts) to determine whether the t-value is large enough to be significant. If it is, you can conclude that the difference between the means for the two groups is different (even given the variability). Fortunately, statistical computer programs routinely print the significance test results and save you the trouble of looking them up in a table.

If the P value that is calculated is less than the threshold chosen for statistical significance (usually the 0.05 level), then the null hypothesis that the two groups do not differ is rejected in favor of the alternative hypothesis, which typically states that the groups do differ.

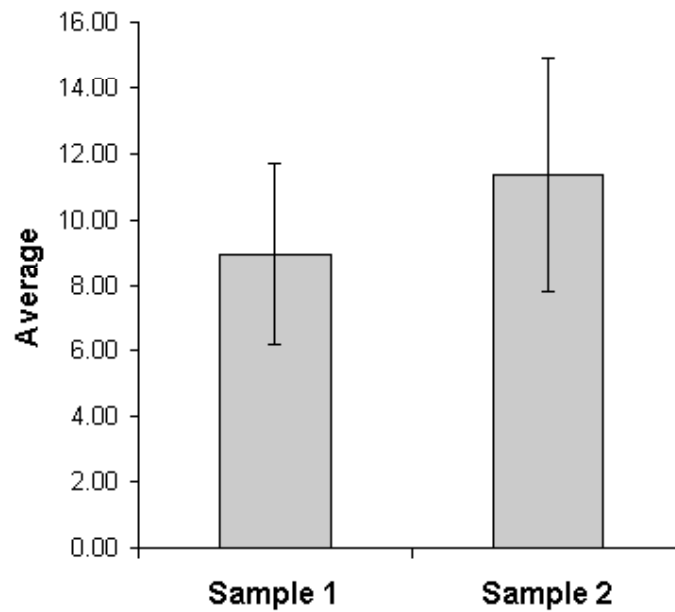
Therefore, if the P-value calculated is say, 0.01, then since, this value is below 0.05, so we reject the hypothesis that the 2 groups are the same, and conclude the 2 populations are statistically different from one another.

Example: Suppose we have the following data from two populations

Sample 1	Sample 2
7.85	12.50
8.51	12.94
13.66	6.26
11.03	6.10
6.59	13.19
8.04	10.74
14.16	6.06
8.13	12.53
6.79	15.45
11.06	15.64
5.83	15.19
10.73	14.93
6.68	7.94
5.02	8.28
10.37	12.65

It is always useful to look at a summary graph of your data. In this case, we can plot the mean value of the variable for each sample in a bar graph with error bars showing the standard deviation of each group. This will produce a graph like Figure 1 below.

Figure 2



By looking at this graph, it is evident that the mean for sample 1 is lower than the mean for sample 2. Is this difference statistically significant? Well, the answer to that depends not only on the difference between the means of the two samples, but also on the difference between their variability. This is exactly what a t-test takes into account.

The computer program Excel will calculate the t-statistic along with the mean, standard deviation, etc. and will give you a table of results like the one below:

	Sample 1	Sample 2	
Mean	8.96284	12.04958	(1)
Variance	7.627848	4.055883	(2)
Observations	15	15	(3)
Pooled Variance	5.841865		(4)
Hypothesized Mean Difference	0		(5)
df	28		(6)
t-stat	-3.49748		(7)
P(T<=t) one-tail	0.000793		(8)
t Critical one-tail	1.70113		(9)
P(T<=t) two-tail	0.001587		(10)
t Critical two-tail	2.048409		(11)

Rows (1), (2), and (3) give you the mean, variance and number of observations for each variable. Row (4) gives you the “pooled” variance (i.e., for both samples together), used to calculate the t statistic. Row (5) gives the hypothesized mean difference (usually zero if you want to know if there is no significant difference between the groups). Row (6) gives the “degrees of freedom.” Row (7) presents the t-statistic (the higher the absolute value, the less similar the means of the two samples are). Row (8) gives you the one-tailed probability that the t-statistic calculated for your data is lower than or equal to the critical t-value [given in row (9)]. Rows (10) and (11) give the probability and critical t-value for two tails. (You should use a one-tailed test if your hypothesis is that the mean of sample 1 is either higher or lower than the mean of sample 2; you should use a two-tailed test if your hypothesis is that the means of the two samples differ, no matter which one is higher and which is lower.)

So, in this example, the mean of sample 2 is higher than the mean of sample 1. Suppose our hypothesis was that the means are different, no matter which one is higher; we would then use the two-tailed test. This difference is statistically significant, since the two-tailed probability is 0.001587, which is much lower than alpha (i.e., 0.05).

Exercises

1. As an aspiring earthworm enthusiast, you decide to compare earthworms found in your vegetable garden vs. those found in your yard. You collect and measure the length of 10 worms in each location, the values of which are listed below:

Vegetable Garden	Yard
6.0cm	5.5
5.2	4.6
8.2	4.9
4.8	6.8
6.5	7.5
6.6	7.7
3.9	6.2
5.7	6.6
7.4	5.7
6.9	5.6

Calculate the mean, variance and standard deviation for each population.

2. Create a bar graph comparing these two populations by following the instructions below:
- Open the computer program “Excel” on either a Windows or Macintosh computer. When it asks what type of window you wish to open, chose “Excel workbook”. A blank spreadsheet should appear.
 - Under “column A”, type “Vegetable Garden” in the first cell. Under that, type in the values starting with 6.0, and typing in one value per cell until all 10 values have been entered. Repeat in “column B” with the data in the “Yard” column.
 - In the cell below the last entry in column A (cell A12), type “=average(“ (without the quotation marks). Then use the mouse to click on the first number in the column (6.0) and drag to the last (6.9). This should select all the numbers in the first column. Finally, type “)” and then press the return button. The average for the numbers you selected should appear in the cell. Make sure it matches the number that you calculated in #1. Repeat for column B.
 - In the cell below the one in which you just calculated the average in column A (A12), type “=stdev(“ (without the quotation marks). Then use the mouse to click on the first number in the column (6.0) and drag to the last (6.9). This should select all the numbers in the first column. Finally, type “)” and then press the return button. The standard deviation for the numbers you selected should appear in the cell. Make sure it matches the number that you calculated in #1. Repeat for column B.

- e. In the toolbar near the top of the window, click on the “Chart Wizard” button. This is the button that is 5th from the right and is a picture with a bar graph with a magic wand over it. If you are not sure, you can pass the mouse over the buttons and it will identify each button.
- f. In the dialogue box that appears, under chart type click on “Column” (1st one). Under chart sub-type, click on the first picture in the first row. Click on the button that says “Next>”.
- g. In the next dialogue box that appears click on the button that says “Series” that is next to “Data Range”.
- h. In the series dialogue box that appears, click on the button that says “Add”. In the text box next to “Name:” type in “Vegetable Garden”. Click on the box next to “Values:”, and delete anything that is in the box. Then click on the cell in your spreadsheet in which the average of your vegetable garden appears (should be A12).
- i. Click on the “Add” button again. In the name box, now enter “Yard”. Next to values, select the cell in your spreadsheet that has the average for your yard data. Then click on “Next>”.
- j. If it is not already selected, click on the button that says “Titles”. In the dialogue box, Under “Chart title”, enter a name for you graph. It should be descriptive enough that someone looking at just the graph could understand what is being shown. “Worm length” is **NOT** an adequate title.
- k. Under Category (X) axis, type in “Population”. This names your horizontal axis, or the independent variable in our experiment. Under Value (Y) axis, type “Mean Earthworm Length (centimeters)”. This names your vertical axis, the dependent variable in our experiment and tells us the units in which it was measured. Click on “Next>”.
- l. In the new dialogue box click under Place chart on the button next to “As new sheet”. Click on “Finish”.
- m. Now we need to add our standard deviation. Double click on the first bar in your graph. In the dialogue box that appears, click on the button that says “Y error bars”. Under display, chose the error bars that go both up and down (\pm from the mean). This is the first option. Under “Error amount”, chose “custom” and then type in your calculated value for standard deviation for your first population in both the “+” and “-“ box. Click “OK”.
- n. When your graph reappears, double click on the second bar and repeat step m, but add the standard deviation you calculated for your second set of data. Your graph is done!
- o. Look over your graph. Are the axes labeled correctly? Could a reader easily interpret what the graph is showing? Is the legend clear? Is the title sufficiently descriptive? Make sure the reader knows what the error bars indicate.
- p. To perform a t-test, go to the “Tools” menu, and select the “Data analysis” option; this will open the Analysis ToolPak. (If there is no “Data analysis” option in your “Tools” menu, then you have to install it by

- choosing “Add-ins” under the Tools menu.) Select the two-sample t-test or the paired t-test option, as appropriate.
- q. In the t-test window, select the ranges of each of your two variables. Select the significance level ($\alpha = 0.05$ is the conventional value). In the “Output options” section select “New Worksheet Ply;” this will create a new page with your results. Click “OK.” Are the 2 populations statistically different from each other? Explain.
 - r. When you’re satisfied, print out your graph and attach it with your statistical results.

Assignment 2

Your assignment is to use the tools described so far to compare two populations. We realize that this is a very open assignment -- that is intentional. You could compare two ant nests, the same species of tree in two woodlots, pillbugs, fish, plants, etc. All we require is that each population have at least 20 individuals to insure a significant sample. While you may work on this assignment with others, each student must utilize their own, unique data set for the analysis. If you have any questions or problems or just need suggestions, please check with Dr. Davis.

Please provide the following in your response:

- 1.) You need to first clearly define what you are calling a population. You may want to look up the biological definition of a population!
- 2.) Then compare two (or more!) such populations by measuring a variable on at least 20 individuals in each population. For example, you could measure tree diameter at breast height, the number of flowers on each plant, the length of insects, earthworms, etc. You then need to calculate the mean, variance and standard deviation for each population.
- 3.) Construct at least one good graph demonstrating some factor that you think is more important for demonstrating that these populations are or are not different from one another. Perform a t-test to determine if the populations are statistically different from one another. Write a short synopsis of your results.
- 4.) AS AN EXAMPLE (please do not use this exact approach): Student Bob and Sally worked together and sampled two areas of lawn on campus for the small plant, *Oxalis stricta* (yellow wood sorrel). In each location, they examined 30 plants. Bob counted the number of leaves on each plant, and Sally counted the number of flowers. In this case, each population consists of all the plants of *Oxalis* on one lawn on Tuesday, October 7, 2003. By taking a small area only, they are obtaining a subsample of the larger population and assuming that it represents the entire population.

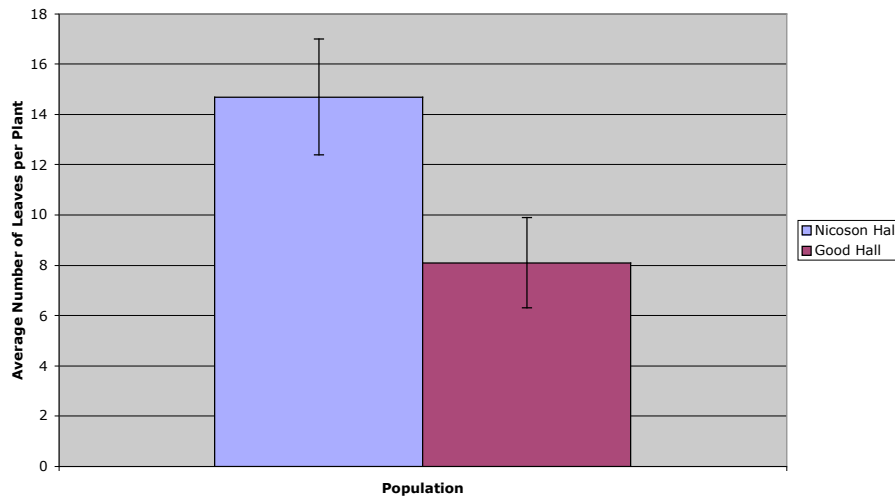
They found:

Shaded lot between Nicoson and Lilly Hall: $n = 30$ plants, mean size $\pm 1SD = 14.7 \pm 2.3$ leaves per plant, mean number of flowers $\pm 1SD = 0.27 \pm 0.16$ flowers per plant

Lawn next to Good Hall: $n = 30$ plants, mean size $\pm 1SD = 8.1 \pm 1.8$ leaves per plant, mean number of flowers $\pm 1SD = 0.10 \pm 0.09$ per plant

Clearly, the plants were smaller and produced fewer flowers in the shaded lot than in the lawn Good Hall. This comparison is shown in the following graphs:

Size of *Oxalis* Plants in a shaded lot (Nicoson Hall) and an unshaded lawn (Good Hall) as measured by the number of leaves per plant $\pm 1SD$



Flower Production on *Oxalis* Plants in a shaded lot (Nicoson Hall) and an unshaded lawn (Good Hall)

