7

Descriptive Statistics

Introduction

Congratulations! You have completed your experimental data collection and are now ready to organize and analyze your data to determine whether or not your hypothesis is supported. Chapters 7, 8, and 9 of this *STEM Student Research Handbook* contain guidelines on how to graphically represent data and provide tips on using descriptive and inferential tests in STEM research.

Recording Calculations in Your Laboratory Notebook

I discussed in Chapter 6, "Organizing a Laboratory Notebook," how important it is to write down everything you do for your experiment in one place—in either a paper or an online laboratory notebook. By "everything," I mean the organization of your data and the application of statistical tests to your experimental research. In other words, in your laboratory notebook, don't put only the final polished tests you plan on using in your research paper or presentation but rather all the calculations you make to help you determine what is significant and what is not. Therefore, your laboratory notebook will be full of charts, graphs, tables, calculations, and computer printouts that may never make it into your paper or presentation. Remember, one of the main purposes of the laboratory notebook is to provide a place for you to record your observations, inferences, and analyses throughout the process. And recall the laboratory notebook guidelines from Chapter 6: never rip out pages (or delete text), and never use correction fluid to hide changes or errors. Simply cross out numbers with a single line.

When beginning statistical analysis in your laboratory notebook, you should clearly label which statistics you are using on what data. Write out—directly on the notebook page—preliminary tables to help you in your calculations. If you use a calculator to record the results, round only at the end

Key Terms

Arithmetic mean: A measure of central tendency that is the centermost point when there is a symmetric distribution of values in the data set; also referred to as an *average*.

Bimodal: A data set in which there are two clear data points that are represented more often than the other data.

Central tendency: A group of descriptive statistics that measure the middle or center of a quantitative data set, whose value, or number, best represents the entire data set.

Data set: The collection of similar data recorded within a researcher's laboratory notebook.

Descriptive statistics: Statistics that describe the most typical values and the variations that exist within a data set.

Interquartile range (IQR): A type of statistical variation that presents a data set as a graphical representation that indicates the range of data organized into four quartiles.

Median: A measure of central tendency that represents the number that appears in the middle of an ordered, finite list of data.

Mode: A measure of central tendency that represents the value that appears most often in a data set.

Outliers (outlier data): Data points that seem to lie outside the data set and do not appear to belong with that data set.

Range: A type of statistical variation that indicates the difference between the largest and smallest values in a data set. It is a measure of how spread out the data are, and therefore it is sometimes called the *spread*.

Standard deviation (σ, SD, s): A commonly used type of statistical variation that measures how close data are from the mean.

Statistical variation: A measure of how scores differ from one another; also known as *variation*, *statistical dispersion*, *spread*, or *dispersion*.

Unimodal: A data set in which there is one clear data point that is represented more often than the other data.

Variance (s^2): A type of statistical variation that is the standard deviation squared.

of the calculations by using the memory feature of the calculator. It is important to correctly apply the order of operations and put every number with the right expression in the technology you use. If you are using Microsoft Excel or another statistical program, use printouts as a record that you performed the statistical test. Always write down your interpretation of the statistics and thoughts regarding the usefulness of the statistical test in relationship to what it tells you about the data within your experiment. Use formal, scientific language. An *unacceptable* laboratory notebook entry would be:

> *Cool, finally after messing with those gross bean sprouts, the t-test shows that the number of plants makes a difference to how many of these weird things actually come up.*

An *acceptable* laboratory notebook entry should be an interpretation of the statistics including explanations about how results might introduce new questions. For example:

> *The t-test results support the hypothesis, and therefore the number of bean sprouts per square cm of soil does have an impact on speed of germination. I wonder if, based on this, I may be able to determine the optimal number of seeds to plant in order to shorten germination time of an entire crop.*

If you are using a paper laboratory notebook, either write directly within the notebook or attach printouts to the page with your personal interpretation nearby. If you are using an online notebook, the same principles apply except that you will have links to Excel (or other statistical software) files showing the same processes. When performing statistical tests using computer technology, refrain from running statistics and simply deleting the results when they do not appear to highlight important findings. The laboratory notebook is a record of what works *and* what does not. Therefore, use the "worksheet" function in Excel to record the many analyses that you will perform on your data. Label graphs and mathematical calculations clearly, and write down your analysis for each descriptive statistic, graphical representation, or statistical test before moving on to another.

Introduction to Descriptive Statistics

Descriptive statistics are statistics that describe the most typical values and the variations that exist within a data set (Salkind 2008). The term *data set* refers to the numerical data you recorded as the results from your experiment. For example, a data set might include the measurements you recorded for the 20 trials of your experimental and control groups. The most common way to describe data is by using the measures of central tendency and the statistical variation.

Measures of Central Tendency

The measure of central tendency is the one value of a quantitative data set that is most typical (Cothron, Giese, and Rezba 2006). This number is used to best represent the entire data set. There are three types of central tendency that will be discussed in this section: *mode, arithmetic mean (average),* and *median.*

Each represents the entire set of values but highlights the central tendency of a distribution differently. Measures of central tendency can be used in isolation to analyze your data; however, being able to calculate mode, mean, and median are critical to performing inferential statistics that are introduced in Chapter 9.

Technologies for Calculating Statistics

Calculators

Texas Instruments *http://education.ti.com/educationportal/sites/US/homePage/index.html*
Casio *www.casio.com/products*

Spreadsheet Software

Open Office *www.openoffice.org*
Microsoft Excel *http://office.microsoft.com/en-us/excel*
Spreadsheet function in Google docs *https://docs.google.com*

Statistical Software

Key Press Fathom *http://keypress.com*
SPSS *www.spss.com*
Minitab *www.minitab.com*
PSPP (open source) *www.gnu.org/software/pspp*

Online Statistical Tutorials

Stat Tutorials *www.stattutorials.com*
Stat Tutorials for performing statistics in Excel *www.stattutorials.com/EXCEL/index.html*
Elementary Statistics and Probability Tutorials and Problems *www.analyzemath.com/statistics.html*

Mode

The *mode* is the value that appears most often in a data set (Salkind 2008). For example, your teacher may use this central tendency to know which of the items on a test was the most difficult for the class—it will be the item most students missed. For another example, suppose you collected temperatures for 11 days. You recorded the following temperatures in Celsius:

12, 12, 13, 14, 14, 15, 15, 15, 15, 37, 39

Both 12°C and 14°C appeared twice and 15°C appeared four times. Since 15°C appeared the most, this is called the mode of the data. The term *unimodal* means that there is one clear data point that is represented more often then the other data. The mode is not necessarily unique because two values can have the same frequency in the same data set. One of the disadvantages of using

mode is that many data sets have more than one value that is represented. For instance, in the data set {2, 3, 3, 4, 5, 5} both the number 3 and the number 5 are represented twice; therefore, this data set has two modes, 3 and 5 (which is referred to as *bimodal*). In addition, if the distribution is uniform, as in this data set {5, 5, 5, 5, 5, the mode is 5 but has little meaning and should not be used. The mode is best used with other measures of central tendency but not alone.

The mode should be used (1) when the data are categorical in nature or (2) when the data are not uniform. For example, if we want to know which tools and instruments are most used in a chemistry laboratory, each item would be tallied into a category (e.g., thermometers, beakers); the item with the most tallies would be the mode and would represent the most used items. Mode would also be appropriate to use in an experiment that has kilogram measurements of {65, 68, 69, 71, 72, 73, 73, 75, 77} because the data are not uniform and vary within the set. However, if the data set has no mode {65, 68, 69, 71, 72, 73, 75, 77}, then it should not be used as a measure of central tendency.

Arithmetic Mean (Average)

The arithmetic mean is commonly called an *average* or a *mean.* The mean is a measure of central tendency that is the centermost point when there is a symmetric distribution of values in the set (Weiss 2008). In other words, it is the centermost point because all the values on one side of the mean are equal in weight to all the values on the other side of the mean. Therefore, the arithmetic mean is very sensitive to extreme values, which will make it less representative of the set of values and less useful as a measure of central tendency for data sets with extreme values.

To calculate the mean, you add all the values and divide by the total number of values. The mathematical formula for mean is as follows:

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

where $\sum x$ represents the sum of all the values within the data set and n represents the total number of values within the data set.

If you had multiple entities or trials within a group, you may want to calculate a mean so that the data in each group are combined for comparison. For example, if the trial measurements for density of a control group were 8 g/cm^3, 10 g/cm^3, 7 g/cm^3, and 9 g/cm^3, the arithmetic mean would be calculated by adding all the measurements (or data points) together and dividing by the total number of measurements (or data points) of the entire group.

$$\frac{8\ g/cm^3 + 10\ g/cm^3 + 7\ g/cm^3 + 9\ g/cm^3}{4} = \frac{34\ g/cm^3}{4} = 8.5\ g/cm^3$$

Means are a good statistic to use if your data are unimodal and approximately symmetrical on each side. For example, in the following ordered list, the single data point that is represented more often than any other number is the number 51.

12, 20, 33, 42, 48, 49, 50, 51, 51, 51, 56, 58, 64, 77, 83, 90

Having symmetrical data means that the data are divided in half—that the two sides are identical. Therefore, the numbers above are unimodal and symmetric.

If your data are not unimodal or symmetric, calculating means could produce results that are misleading. For example, suppose you collected temperatures for 11 days. You recorded the following temperatures in Celsius:

12, 12, 13, 14, 14, 15, 15, 15, 15, 37, 39

The mean, or average, is calculated as follows:

$$\frac{12 + 12 + 13 + 14 + 14 + 15 + 15 + 15 + 15 + 37 + 39}{11} = \frac{201}{11} = 18.27°C$$

What temperature was most common in this data set? Was it 18°C? No! Clearly from the data above, the most common temperatures were in the early to midteens. If you were to look at a histogram or dot plot of this data, you could see that it is bimodal. That is, there are two obvious peaks of the data, one at 15 and the other at 39. Therefore, the mean, in this case, is *not* the best measurement to use to calculate and display your data. Because of the extreme value, calculating the median would be a more accurate way to represent this data.

Median

The *median* is the number that appears in the middle of an *ordered* finite list of data (Cothron, Giese, and Rezba 2006). A median is the number that separates the higher half of the data set from the lower half. Median can be found by ordering the list from lowest value to highest value and then choosing the middle number. Therefore, the median of an odd-numbered data set is the center value; but for an even-numbered data set the median is the calculation of the mean of the two middle values. The median should be used when the data set has extreme values. Unlike the arithmetic mean, the median is insensitive to extreme values. The median is the centermost value of the data set even when the data set has one or more extreme values.

The difference between mean and median is that the arithmetic mean is the middle point of the set of values, and the median is the middle point of the number of data in the data set. Therefore, the median is an indication about how many data there are in the data set, not the values of those cases. In the data set below, there are 11 values. The middle number is the sixth number because there would be five data points on each side. Therefore, the median of this data set is 15.

12, 12, 13, 14, 14, 15, 15, 15, 15, 37, 39

Comparing the Three Types of Central Tendency

We have looked at three different types of central tendency: mode, arithmetic mean, and median. In a set of data with normal distribution, all three are the same value. However, when data are not distributed normally, the values of central tendency will vary. The mean is used with many inferential statistical tests like the *t*-test and ANOVA. (See Chapter 9 for more information about inferential statistics.) The mean is not the centermost value of a data set when the data set has extreme scores, in which case you would use the median. However, when the data set is categorical, mode should also be used for central tendency. Remember that the mode should not be used as the only measure of central tendency; it should always be used with another measure of central tendency.

Statistical Variation

Another way to describe the data is by using statistical variation. A measure of *statistical variation* reflects how scores differ from one another (Salkind 2008). These differences are also called *variation, statistical dispersion, spread*, or *dispersion.* This measure indicates how different scores are from one another. Even if two sets of data have the same mean, the statistical variation can be different and describe how different the data sets are. Variation represents the average difference from the mean. The variation increases as the data become more diverse. Below, I discuss the four measures of variation: *range, interquartile range, standard deviation,* and *variance.*

Range

Range is the most general measure of variation. It is a type of statistical dispersion that indicates the difference between the lowest and highest values in a data set (Triola 2001). It is a measure of how a set of measurements or data is spread out and is sometimes called the *spread of the data.* In other words, the range gives an idea of how far apart values are from the lowest to the

highest. Range is calculated by taking the difference between the highest and the lowest number in the data set.

For example, in the temperature data on page 98, the range would be the highest temperature measured minus the lowest temperature measured. The highest temperature measured is 39°C and the lowest temperature measured is 12°C, then the range is 39°C – 12°C = 27°C. Therefore, the range, or spread, of this data is 27°C.

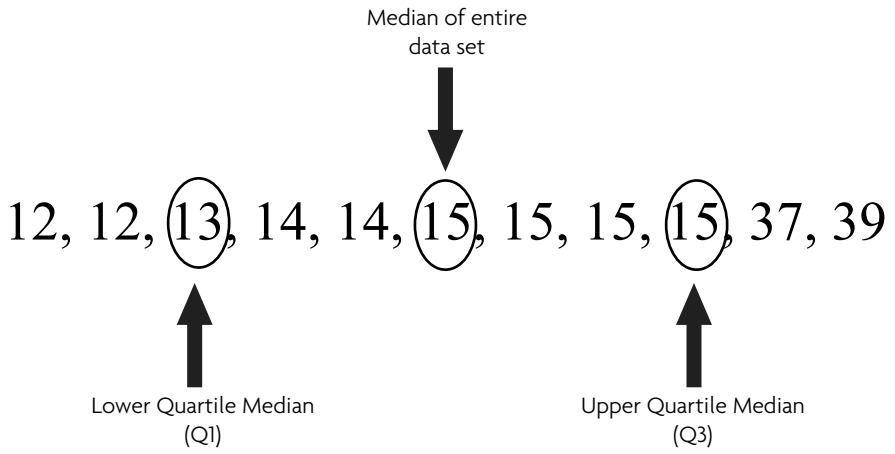
Interquartile Range

Another measure of statistical dispersion is interquartile range (IQR), also called the *midspread* or *middle fifty.* IQR is a graphical representation of a data set that indicates the range of the data organized into four quartiles (Cothron, Giese, and Rezba 2006). To calculate the IQR, you will need to determine the upper and lower quartiles. These numbers will separate your data into four quartiles, or quarters, each containing 25% of the data. To do this, find the median of the entire data set. Next, identify the median of the lower half of the data. This is the lower quartile median. For the upper quartile, find the median of the upper half of the data.

The median of the data set in Figure 7.1 is 15°C. The lower quartile median is 13°C and is labeled with the notation, Q1. The upper quartile is 15°C, and is labeled Q3. To calculate the interquartile range, subtract the upper and lower quartiles, 15 – 13 = 2°C. Therefore, IQR = 2. Calculating the interquartile range is very helpful for determining outlier data, which I discuss next, and for constructing a box and whisker plot (which I discuss in Chapter 8).

Figure 7.1

Finding the Interquartile Range



Dealing With Outlier Data

Outlier data, or *outliers,* are data points that lie outside the data set and appear not to belong with that data set (Salkind 2008). Outlier data can occur within experiments due to either observational or recording error. If this is your first time performing the particular methods in your experiment, you can expect to make mistakes in the data collection process. Even professional scientists make errors when performing new methods. This is normal. If you confidently know that the outlier data is due to your own error, instrument error, or some other known cause *not* associated with your independent variable, you may consider throwing out outlier data when performing your calculations for data analysis.

Of course, these outliers must still be mentioned in the Results and explained in the Analysis and Conclusions sections of your paper or poster. Because outliers can taint the data, it is better to do calculations without them, but you must be confident that the outlier data is *not* related to the independent variable.

Consider the temperature data again. Suppose you noticed after recording temperatures of 37°C and 39°C that the thermometer had been broken. Therefore, instrument error is the known cause of the outlier data, and the two inaccurate readings can be left out of the calculations. If calculations are done only on the data set {12, 12, 13, 14, 14, 15, 15, 15, 15}, the results will not only be different but will show a more accurate picture of the scenario. On the other hand, if the two values 37°C and 39°C represent actual temperature measurements, they should remain in the data set.

How can you determine whether an outlier is a piece of data that does or does not belong? Mathematically speaking, there is an outlier "rule of thumb" to follow. Let y be a piece of data; y is an outlier if

$$y < Q1 - (1.5 \times IQR)$$
$$\text{OR}$$
$$y > Q3 + (1.5 \times IQR)$$

For example, in our temperature data:

$$Q1 = 13$$
$$Q3 = 15$$
$$IQR = 2$$

To determine whether or not 39 is an outlier, we would use the second equation with the Q3 number because our "outlier-in-question" is on the high end of our data set, not the lower end. Therefore

$$15 + (1.5 \times 2) = 18$$

Since 39 > 18, we have mathematical confidence that 39 is an outlier (as is 37). Use this calculation to help you determine which data you can confidently throw out when performing your statistical analysis. But remember, outliers still must be mentioned and addressed in your final paper or poster.

Standard Deviation

Standard deviation (SD) is the most frequently used measure of variation. SD is a measure of how close data are to the mean (Weiss 2008). It shows how much variation or dispersion there is from the arithmetic mean and represents the average amount of variation in a set of scores. The larger the standard deviation, the larger the average distance each data point is from the mean of the distribution. Therefore a low standard deviation indicates that the data points tend to be very close to the mean. If, on the other hand, the standard deviation is high, the data is spread out over a large range of values from the mean; scores can be close and still far away from the mean.

There are two types of standard deviation: one is calculated for a population and the other is calculated for a sample. As first introduced in the Chapter 2, "Research Design," *population* means the complete collection of every item that has the same characteristics. For your purposes, the population is the entire group from which you want to collect data. The *sample* is the subset of the population from which you actually collect data. Therefore, the sample is meant to be representative of the population. Statistically, all tests are more powerful with larger sample sizes (Triola 2001). For projects you design, you should have between 4 (the minimum) and 10 entities within each of your experimental and control groups. If your sample has more than 30 data points, additional, more powerful statistical tests can be used (Gonzalez-Espada 2007).

Getting information from the entire population is impractical and usually impossible in STEM research. Most likely, your research data represents a sample coming from a larger population. On the following page note the mathematical differences in the calculations of the standard deviation of a population compared to the standard deviation of a sample.

There are two important differences between the population and sample equations in the box. One of the differences is their symbolic notation: the standard deviation notation for *sample* is s and for *population* is σ. The other differences are the denominators of the fraction in the square root in each standard deviation. The standard deviation for a population divides by n and the standard deviation for a sample divides by $n - 1$.

The standard deviation for a sample, which divides by $n - 1$, is an unbiased estimate. By subtracting 1 in the denominator, the number is smaller and therefore the approximation of standard deviation is larger. Therefore, in case

Mathematical Notation for Standard Deviation in a *Population*

$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$, where

σ is the symbolic notation for standard deviation,
∑ is sigma, which tells you to find the sum of what follows,
x_i is each individual value in the data set,
μ is the arithmetic mean of all the values, and
n is the number of values the data set has.

Note that $(x_i - \mu)$ tells you to find the differences between each individual value and the arithmetic mean.

Mathematical Notation for Standard Deviation in a *Sample*

$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$, where

s is the symbolic notation for standard deviation (another symbolic notation is SD),
∑ is sigma, which tells you to find the sum of what follows,
x_i is each individual value in the data set,
$\bar{x}$ is the arithmetic mean of all the values, and
n is the number of values the data set has.

Note that $(x_i - \bar{x})$ tells you to find the differences between each individual value and the arithmetic mean.

there is an error, the standard deviation will be an overestimation and will compensate for the possibility of errors, and therefore the standard deviation for a sample equation is a good choice for STEM research.

An example of how standard deviation can be calculated for an experiment is recorded in an article by Shaefer et al. (2000). The authors reported that Hurricane Hugo had a significant impact on stream water chemistry on tropical streams located in the El Yunque National Forest (which is located in Puerto Rico and is part of the U.S. Forest Service). Table 7.1 (p. 106) shows a sample of 10 randomly collected ammonia measurements (kg/hectare per year) in the first year after Hurricane Hugo (the hurricane occurred in September 1989).

Table 7.1

Ten Ammonia Measurements Taken in the El Yunque National Forest in the First Year After Hurricane Hugo

57	66	88	96	116
147	147	154	154	175

Find the arithmetic mean:

$$\bar{x} = \frac{57 + 66 + 88 + 96 + 116 + 147 + 147 + 154 + 154 + 175}{10} = \frac{1{,}200}{10} = 120$$

x	$x - \bar{x}$	$(x - \bar{x})^2$
57	(57 – 120) = -63	$(-63)^2 = 3969$
66	(66 – 120) = -54	$(-54)^2 = 2916$
88	(88 – 120) = -32	$(-32)^2 = 1024$
96	(96 – 120) = -24	$(-24)^2 = 576$
116	(116 – 120) = -4	$(-4)^2 = 16$
147	(147 – 120) = 27	$(27)^2 = 729$
147	(147 – 120) = 27	$(27)^2 = 729$
154	(154 – 120) = 34	$(34)^2 = 1156$
154	(154 – 120) = 34	$(34)^2 = 1156$
175	(175 – 120) = 55	$(55)^2 = 3025$
Total = 1,200		Total = 15,296

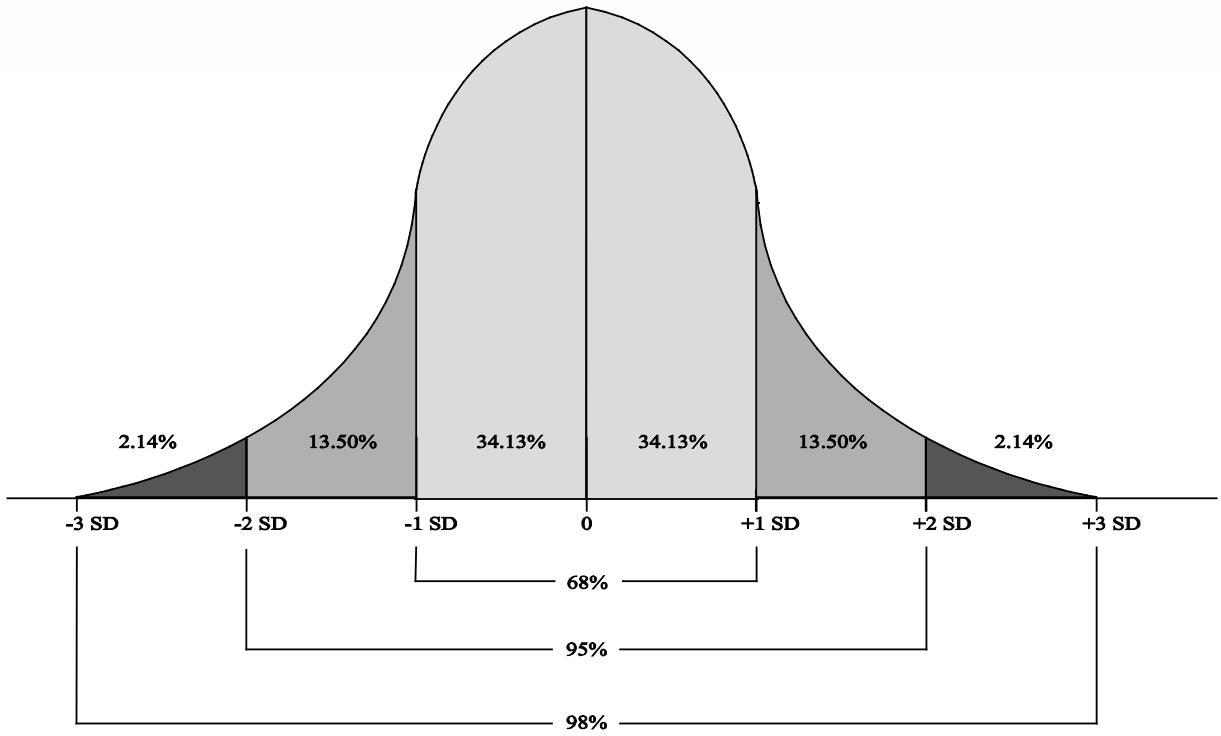
Then the standard deviation is

$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{15296}{10-1}} = \sqrt{\frac{15296}{9}} \cong 41.23$ kg/hectare per year. This indicates that there is an average distance from the arithmetic mean to the data of 41.23 kg/hectare per year. A standard deviation curve can be used to further explain the data.

As long as the distribution of data in the population is normal (which can be determined by a bell-shaped histogram), the addition of the mean and the standard deviation will provide additional insight into the data. The mean is represented in the highest peak of a normal standard deviation curve and is labeled as zero, as shown in Figure 7.2. Most of the data are found within 1 SD of the mean (34.13% on either side).

Figure 7.2

Standard Deviation Curve



In the previous stream water chemistry example, ranges can be calculated by subtracting one standard deviation from the mean to show where 34% of the data are found. Since the mean is 120, we add or subtract 41.23 from each side to find the range of the first standard deviation.

$$120 + 41.23 = 161.23$$
$$120 - 41.23 = 78.77$$

These calculations indicate that the 34% of the data is in the range of 120 and 161.23 kg/hectare per year and that 68% of the data set is between the values 78.77 and 161.23 kg/hectare per year. Notice how the standard distribution helps describe the data. With the standard deviation, you can answer a lot of questions of your data, for example:

- What values were above the mean? What values were below the mean?
- What is the percentage of the data that were in a range of values? What 50% of the values are above the mean?
- What range of values would you expect for 95%?

The standard deviation is the average distance from the arithmetic mean (do not use median or mode). If every number is the same, the standard deviation is 0. Therefore, the larger the standard deviation, the more spread out the values are and the more different they are from one another. The standard deviation, like the mean, is sensitive to extreme scores.

Variance

The *variance* is the standard deviation squared (Weiss 2008). Once you have calculated standard deviation, you can just square it to get the variance.

The Mathematical Notation for Variance

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

s^2 is the symbolic notation for variance,
$\sum$ is sigma, which tells you to find the sum of what follows,
x_i is each individual value in the data set,
$\bar{x}$ is the arithmetic mean of all the values, and
n is the number of values the data set has.

Note that $(x_i - \bar{x})$ tells you to find the differences between each individual value and the arithmetic mean.

Notice that the variance and standard deviation are closely related and almost the same formula. The variance should always be reported with the standard deviation because the variance informs how far a set of numbers are spread out from one another in addition to the average distance from the arithmetic mean.

Standard deviation and variance can be calculated using spreadsheet software like Excel, or free online calculators such as *www.easycalculation.com/statistics/standard-deviation.php*.

Additional Calculations

There are additional calculations that can be done with your data, including total change and rate of change. If these calculations apply to your data, you should run a statistical test (like the *t*-test or ANOVA) to compare each experimental group's change to the control group in order to determine if the change is significant or not. See Chapter 9 for more information on inferential statistics.

Total Change

Total change measurements are easy to calculate, either by hand or in spreadsheet software such as Excel. These measurements help you compare the results of your experimental groups to one another as well as to the control group. The total change is calculated by taking the final measurement and subtracting the starting measurement.

$$\text{Final measurement} - \text{starting measurement}$$

For example, if day 1 had a measurement of 45 mm and the last day's measurement was 52 mm, the total change is calculated:

$$52 \text{ mm} - 45 \text{ mm} = 7 \text{ mm}$$

If total change increased in this group, the number will be positive, and if the measurement decreases, the number will be negative. If some entity's change varied (increased and decreased) throughout the experiment, total change may not be the most accurate description of your data.

If you had multiple entities or trials in an experimental group, you may want to calculate total change for each entity/trial, as well as an average total change for the entire group. For example, if beginning pressure measurements for an experimental group was 80 Pascals (Pa), 77 Pa, 85 Pa, and 76 Pa, and the ending measurements were 60 Pa, 77 Pa, 63 Pa, and 68 Pa, the average total change is calculated:

$$\frac{(60 + 77 + 63 + 68)}{4} - \frac{(80 + 77 + 85 + 76)}{4} = 67 - 80 = -13 \text{ Pa}$$

Rate of Change

Rate of change is useful when you want to compare the speed at which changes occurred in a specific period of time. This is calculated by dividing the total change in this period of time by how much time is in this period.

$$\frac{\text{Final measurement} - \text{starting measurement}}{\text{Total time}}$$

For example, if an entity's weight changed from 7.2 kg to 6.1 kg in a span of 26 days, the rate of change is calculated:

$$\frac{7.2 - 6.1 \text{ kg}}{26 \text{ days}} = 0.042 \text{ kg/day}$$

Notice that the units used in the numerator and the units in the denominator remain in the final answer to indicate how much change occurred per unit of time. This is helpful when comparing groups to one another.

Using Descriptive Statistics to Explain Experimental Results

The purpose of calculating descriptive statistics, such as central tendency and variation, is to highlight specific characteristics of your data. Descriptive statistics will help you determine and discuss what differences, if any, exist between your experimental and control groups. Organizing these results into tables is a good way to present the data for the Data and Results section of your paper or presentation. For example, a table could contain the mean for each group as well as the standard deviation or variance. Ask yourself questions about the calculations, such as those listed in the "Data Interpretation" section of Chapter 9 on pages 140–142. Write down your observations and analysis about these descriptive statistics in your laboratory notebook.

However, any differences highlighted by descriptive statistics may not necessarily be a direct result of the changes you made in the independent variable (Valiela 2001). It is possible that these changes are due to experimental error or to chance. For example, if you determine a difference in the mean between the control group and one of the experimental groups, at this point you do not know if the difference is due to the random nature of data collection or if the differences are due to the treatment of the independent variable. This is why additional statistics called *inferential statistics* (see Chapter 9) may be necessary. The descriptive statistics that you calculated in this chapter will come in handy should you eventually decide to use inferential statistics to analyze your data. The next chapter, Chapter 8, is about describing your data visually using graphical representations. Organizing your raw data and descriptive statistics into graphs and tables will further help you determine what your data mean. Then, Chapter 9 will introduce you to the basics of inferential statistics—or how to determine whether your results are statistically significant.

Chapter Applications

Once you have collected data from your experiment, you will need to calculate various descriptive statistics explained in this chapter. Do this in your laboratory notebook, making notes to yourself about what the data may be saying about your results.

References

Cothron, J. H., R. N. Giese, and R. J. Rezba. 2006. *Science experiments and projects for students: Student version of students and research.* Dubuque, IA: Kendall/Hunt.

Gonzalez-Espada, W. 2007. Using simple statistics to ensure science-fair success. *Science Scope* 8 (30): 48–50.

Salkind, N. J. 2008. *Statistics for people who think they hate statistics*. Los Angeles: Sage Publications.

Shaefer, D., W. H. McDowell, F. N. Scatena, and C. E. Asbury. 2000. Effects of hurricane disturbance on stream water concentrations and fluxes in eight watersheds of the Luquillo Experimental Forest, Puerto Rico. *Journal of Tropical Ecology* 16 (2): 189–207.

Triola, M. F. 2001. *Elementary statistics.* New York: Addison-Wesley.

Valiela, T. 2001. *Doing Science: Design, analysis, and communication of scientific research.* New York: Oxford University Press.

Weiss, N. A. 2008. *Elementary statistics.* San Francisco: Pearson Addison-Wesley.