CHAPTER 2

Data Analysis

When you complete a laboratory investigation, it is important to make sense of your data by summarizing it, describing the distributions, and clarifying "messy" data. Analyzing your data will allow you to do this.

Working with Data

Data analysis may involve calculations, such as dividing mass by volume to determine density or subtracting the mass of a container from the total mass to determine the mass of the contents. Using the correct rules for significant digits during these calculations is important to avoid misleading or incorrect results.

When adding or subtracting quantities, the result should have the same number of decimal places (digits to the right of the decimal) as the fewest number of decimal places in any of the numbers that you are adding or subtracting.

Table 2.1 presents examples and explains how the proper results should be written.

Table 2.1: Writing Your Results When Adding or Subtracting

Example	Explanation
3.7 cm + 4.6083 cm = 8.3 cm	The result is written with one decimal place because the number 3.7 has only one significant digit to the right of the decimal.
48.3506 m – 6.28 m = 42.10 m	The result is written with two decimal places because the number 6.28 has only two significant digits to the right of the decimal.
(8 km – 4.2 km) + 1.94 km = 6 km	The result is written with zero decimal places because the number 8 has zero significant digits to the right of the decimal.

Notice that the result of adding and subtracting has the correct number of significant digits if you consider significant digits to the right of the decimal.

When multiplying and dividing a set of numbers, look for the number with the fewest significant digits. Your result should have that number of significant digits. Table 2.2 explains how to apply this concept.

Table 2.2: Writing Your Results When Multiplying or Dividing

Example	Explanation
5.246 in. × 2.30 in. = 12.1 in.	The result is written with three significant digits because 2.30 has three significant digits.
0.038 cm ÷ 5.273 cm = 0.0072 cm	The result is written with two significant digits because 0.038 has two significant digits.
76.34 m × 2.8 m = 2.1×10^2 m	The result is written with two significant digits because 2.8 has two significant digits. [Note that scientific notation had to be used because writing the result as 210 would have an unclear number of significant digits.]

When calculations involve a combination of operations, you must retain one or two extra digits at each step to avoid any round-off errors; at the end of the calculation, you must round to the correct number of significant digits.

An exception to these rules occurs when a calculation involves count data, such as the number of times a ball bounces, or the number of waves that pass a point during a time interval. As shown in the following example, do not consider exact numbers when determining significant digits in a calculation.

Example

While performing the Millikan oil-drop experiment, you find that a drop of oil has an excess of three electrons. What is the total charge of the drop?

$$\text{Charge} = (\text{number of electrons})(\text{charge per electron})$$
$$q = ne$$
$$= (3 \text{ electrons})(1.6\times10^{-19}\,\text{C/electron})$$
$$= 4.8\times10^{-19}\,\text{C}$$

When determining the number of significant digits in the answer we ignore the number of electrons because it is an exact number.

Scientific Notation

When manipulating data, there will be many times when the numbers that you calculate will be either too large or too small to be conveniently expressed as decimals. To make it easier to work with these very large or very small numbers, scientists use scientific notation. In scientific notation, a number is written as a coefficient multiplied by the base 10 raised to some exponent. Let's look at Avogadro's number to better understand the components:

The coefficient must be between 1 and 10, and the exponent must be an integer. Very large numbers will have a positive exponent, while very small numbers will have a negative exponent; for example:

$$10000 = 1 \times 10^4$$

$$1000 = 1 \times 10^3$$

$$100 = 1 \times 10^2$$

$$10 = 1 \times 10^1$$

$$1 = 10^0$$

$$1/10 = 0.1 = 1 \times 10^{-1}$$

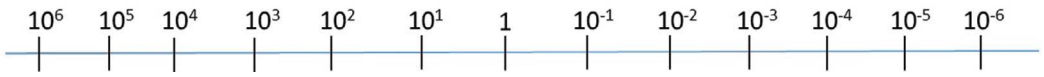
$$1/100 = 0.01 = 1 \times 10^{-2}$$

$$1/1000 = 0.001 = 1 \times 10^{-3}$$

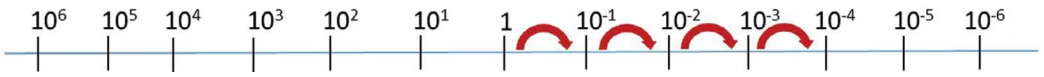
$$1/10000 = 0.0001 = 1 \times 10^{-4}$$

So, a number such as 0.00000000000757 would be written in scientific notation as 7.57×10^{-12}, while a number like 218,000,000 would be written as 2.18×10^8.

Another way of thinking about this is to use the following representation for the place values:



You can rewrite a number in scientific notation by simply using this representation to count how many decimal places to move the decimal point. If the number you are converting is greater than 10, then the decimal point is moved to the left on the line, while if it is less than 1, the decimal point is moved to the right. For instance, if you wanted to convert 0.000436 into scientific notation, you would start at the base unit—the 1—then count the decimal places you would have to move until the coefficient is between 1 and 10, as shown below.



This tells us that we need to move the decimal point four places to the right:

So, 0.000436 would be written in scientific notation as 4.36×10^{-4}.

For more detailed information on using scientific notation, watch the following tutorial:

Khan Academy: Introduction to scientific notation

Calculations Using Percentages

Percent Change

When working with data, sometimes we need to compare unequal quantities or scales; in order to do this we normalize the data. One way to do this is to compare the percent change over time. We use the following formula to calculate percent change:

$$\%\ change = \frac{final\ value - initial\ value}{initial\ value} \times 100$$

For example, in the AP Biology Diffusion and Osmosis lab investigation, dialysis bags are first filled with a sucrose solution and then placed in water for 30 minutes. We measure the mass of each bag before and after it sits in the water for 30 minutes, and report this as a percent change in mass. If the mass of a dialysis bag at the beginning of the experiment was 12.2 g and at the end of the experiment it was 16.7 g, the percent change is

$$\frac{16.7 - 12.2}{12.2} \times 100 = 36.9\%$$

Percent change can also be negative. What if in the previous example the mass at the beginning of the experiment was 16.7 g and the mass at the end of the experiment was 12.2 g? Let's look at this new calculation:

$$\frac{12.2 - 16.7}{16.7} \times 100 = -26.9\%$$

In the first calculation the positive result indicates that the dialysis bags gained mass. However, in the second calculation the negative result indicates that the dialysis bag lost mass.

Percent Difference

There are times when you may need to calculate the percent difference between two experimental results to see how they compare to each other. To calculate percent difference, we use the following formula:

$$\%\ difference = \left| \frac{x_1 - x_2}{(x_1 + x_2)/2} \right| \times 100$$

where x_1 is the first data point and x_2 is the second data point. The numerator is the difference between the measurements, and the denominator is the average of the measurements. The two vertical lines on either side of the fraction indicate that we are using the absolute value of the calculation.

Example

There are two cars traveling at different speeds: one at 25 mph and the other at 33 mph. We want to know the percent difference between the speeds of the two cars. The calculation would be

$$\begin{aligned}\%\text{ difference} &= \left|\frac{x_1 - x_2}{(x_1 + x_2)/2}\right| \times 100 \\ &= \left|\frac{25 - 33}{(25 + 33)/2}\right| \times 100 \\ &= -\left|\frac{8}{29}\right| \times 100 \\ &= 0.2759 \times 100 = 27.6\%\end{aligned}$$

This means that there is a 28% difference between the speeds of the two cars.

Percent Error

Percent error is a calculation that is done when you want to compare your results to a known or predicted theoretical value. We use the following formula to calculate percent error:

$$\%\,error = \frac{|experimental\,value - theoretical\,value|}{theoretical\,value} \times 100$$

Notice that we are using the absolute value of the difference between the experimental value and the theoretical value.

Example

Calculate the percent error of a titration of 3.0% hydrogen peroxide (H_2O_2) with potassium permanganate ($KMNO_4$), as in AP Chemistry Investigation 8: Oxidation-Reduction Titration. If we performed this investigation and calculated the concentration of H_2O_2 in our sample to be 2.74%, our calculation would be:

$$\begin{aligned}\%\,error &= \frac{|2.74 - 3.0|}{3.0} \times 100 \\ &= \frac{0.26}{3.0} \times 100 \\ &= 8.67\%\end{aligned}$$

This means that our titration yielded data that was in error by 8.67% relative to what was expected.

Rate Calculations

You may occasionally have to determine a rate of change when you are processing data from an experiment. Examples include a rate of reaction, growth rate, speed, and acceleration. Each of these describes how a quantity changes over time. The change over time can be expressed as

$$\frac{\Delta Y}{\Delta t} = \frac{\text{the change in the dependent variable}}{\text{the change in time}}$$

where ΔY represents the change on the *y*-axis and Δt represents the change on the *x*-axis (time).

Example

Suppose you were doing an AP Physics lab and wanted to calculate the magnitude of the average velocity (speed) of an object. You would do this by calculating the displacement traveled during a particular period of time. So, if you pushed a toy car across the floor and it traveled in a straight line from 1.0 meter to 4.0 meters in 8 seconds, you would calculate the speed as follows:

$$\text{average velocity} = \frac{\text{displacement}}{\text{time}} = \frac{(4.0-1.0)\,\text{meters}}{(8-0)\,\text{seconds}} = \frac{3.0\,\text{meters}}{8\,\text{seconds}} = 0.375\,\text{meters/sec}$$

Example

Suppose you are doing an AP Chemistry lab and needed to calculate the rate of the decompostion of NO_2 from 60 to 120 seconds:

$$NO_2(g) \rightarrow 2NO(g) + O_2(g)$$

Time (seconds)	$[NO_2]$	$[NO]$	$[O_2]$
0	0.0150	0	0
60	0.0085	0.0027	0.0018
120	0.0071	0.0041	0.0024

$$Rate = \frac{\Delta A}{\Delta t} = \frac{\Delta[NO_2]}{\Delta t}$$

$$Rate = \frac{0.0071\,M - 0.0085\,M}{120\,\text{sec} - 60\,\text{sec}}$$

$$Rate = \frac{-0.0014\,M}{60\,\text{sec}} = -2.33\times10^{-5}\,M/\text{sec}$$

Note that the negative sign indicates that the NO_2 is being consumed in the reaction. We could also calculate the rate using the one of the products:

$$Rate = \frac{\Delta A}{\Delta t} = \frac{\Delta[NO]}{\Delta t}$$

$$Rate = \frac{0.0041\,M - 0.0027\,M}{120\,\text{sec} - 60\,\text{sec}}$$

$$Rate = \frac{0.0014\,M}{60\,\text{sec}} = 2.33\times10^{-5}\,M/\text{sec}$$

Notice that the rate of product formation is the same as the rate of consumption of the reactant. The rate is positive in magnitude because product is being formed.

The following tutorials can help you review how to do rate calculations:

Khan Academy: Intro to rates

Khan Academy: Introduction to average rate of change

Linear Relationships and Curve Fitting

Graphing Data as a Straight Line

When you plot data on *x*-*y* axes, a straight line is the simplest relationship that data might have. Graphing data points as a straight line is useful because you can easily see where data points belong on the line.

You can represent data as a straight line on a graph as long as you can identify its slope (*m*) and its *y*-intercept (*b*) in a linear equation: $y = mx + b$. The slope is a measure of how *y* varies with changes in x, $m = \Delta y / \Delta x$. The *y*-intercept is where the line crosses the *y*-axis (where $x = 0$).

Linearizing Data

Even if the data you measure do not have an apparent linear relationship, you may be able to represent the data as a straight line by revising the form of the variables in your graph. One method is to transform the equation to represent the relationship so that it has the linear form of $y = mx + b$ by substitution. For powers of *x*, the data would be in the form $y = Ax^c + b$. To linearize this data, substitute x^c for the *x* in the linear equation. Then you can plot $y \, vs. \, x^c$ as a linear graph. For example, graphing kinetic energy, *KE*, and velocity, *v*, for the function $KE = \frac{1}{2}mv^2$, yields a parabola, as shown in figure 2.1a. But if we set the horizontal axis variable equal to v^2 instead, the graph is linear, as shown in figure 2.1b, and the slope is equal to $1/2m$ [Note that "1/2" should be a built-up fraction, with *m* setting next to it.].

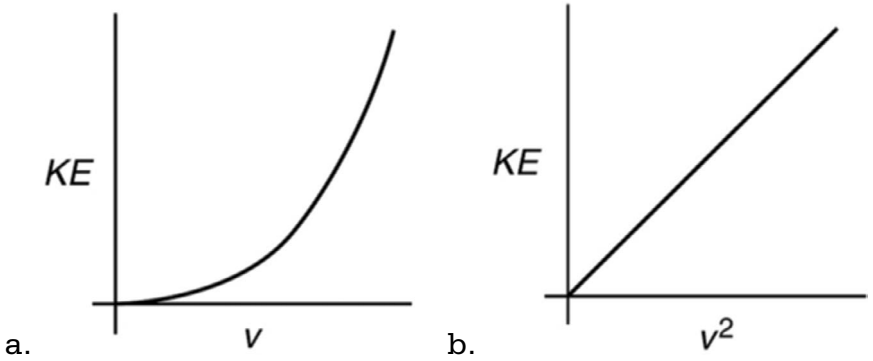


Figure 2.1: Changing the Variable on the *x*-axis to Produce a Linear Graph

If the data is exponential, as in $y = Ae^{bx}$, or is a power of x, as in, $y = ax^n$, taking the log of both sides of the equation will linearize them. For exponential data, the equation you obtain is $ln(y) = ln(A) + bx$. The data will approximate a line with y-intercept ln(A) and slope b.

Similarly, for an equation with a power of x, taking the log of both sides of $y = ax^n$ results in $log(y) = log(a) + nlog(x)$. If you plot log($y$) versus log($x$), the data will approximate a line with y-intercept log(a) and slope n, as shown in figures 2.2a and 2.2b.

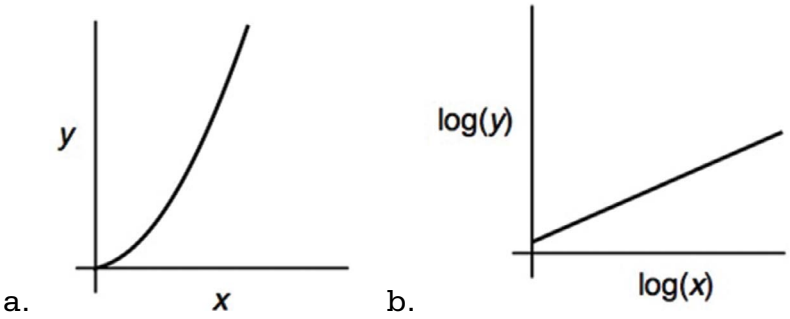


Figure 2.2: Linear Graphs of Equations with a Power of x

Curve Fitting

A useful way to analyze data is to determine whether it corresponds to a certain mathematical model. A mathematical relationship or function will allow you to make a prediction if you know the function and an initial condition. The first step is to plot the points and see if they follow a recognizable trend, such as a linear, quadratic, or exponential function. Figure 2.3 shows examples of each of these types.

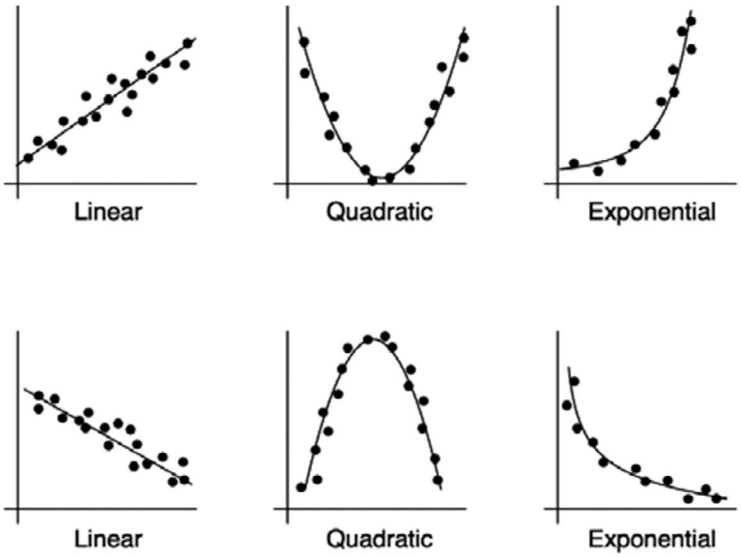


Figure 2.3: Common Mathematical Models

The general equation of a **linear function** is $y = mx + b$, as noted above, in which m is slope and b is the y-intercept. For example, a linear function in physics is the time dependence of the velocity of an object undergoing constant acceleration, $v = v_0 + at$, where the acceleration, a, is the slope and the initial velocity, v_0, is the y-intercept. An example of a linear function in biology is the amount of oxygen consumption by an endotherm over time at a constant temperature. In chemistry, an example of a linear function is the relationship between the concentration of a solution and the amount of light that is transmitted through the solution.

The general equation of a **quadratic function** is $y = ax^2 + bx + c$, where a, b, and c are constants. An example of a quadratic function in physics is the potential energy of a spring, $U = \frac{1}{2}kx^2$, where x is the distance the spring is stretched from equilibrium, k is the spring constant, and in this case the constants b and c are zero. Another example of a quadratic function is the position as a function of time for a constantly accelerating object, $x = \frac{1}{2}at^2 + v_0t + x_0$, where a is acceleration, v_0 is initial velocity, and x_0 is initial position.

The general equation of an **exponential function** is $y = Ae^{bx}$, where A and b are arbitrary constants. An example of the exponential function in physics is the number of radioactive particles left after a certain time of radioactive decay: $N = N_0e^{-\lambda t}$, where N_0 is the original number of particles, and λ is the decay rate. Population growth is an example of an exponential function in biology and environmental science (see the section on population growth later in this chapter).

If the pattern of the data is clearly linear, or if you can plot the data using linearization, you can use a straightedge to draw a **best-fit line** that has approximately the same number of data points above and below the line. You can then determine an equation of the line by identifying the slope and y-intercept of the best-fit line.

If a more exact equation is desired, or if the data do not clearly follow a linear pattern, you can use a graphing calculator or a computer to fit the data to a mathematical model. In this case, you input the data and choose the model that you think will best fit the data. This is called **regression analysis**. Regression analysis is a common curve-fitting procedure. An analysis using this procedure provides parameters for the equation you have chosen for the fit, as well as parameters that describe how well the data fit the model. Figure 2.4 shows the same data using a linear model and a quadratic model. The value r^2 is the **coefficient of determination**. It indicates how well the model fits the data. A value closer to 1 indicates a better fit. In the examples in figure 2.4, both models are a good fit for the data, but the r^2 values show that the quadratic model is better because 0.9826 is closer to 1 than 0.95492 is.

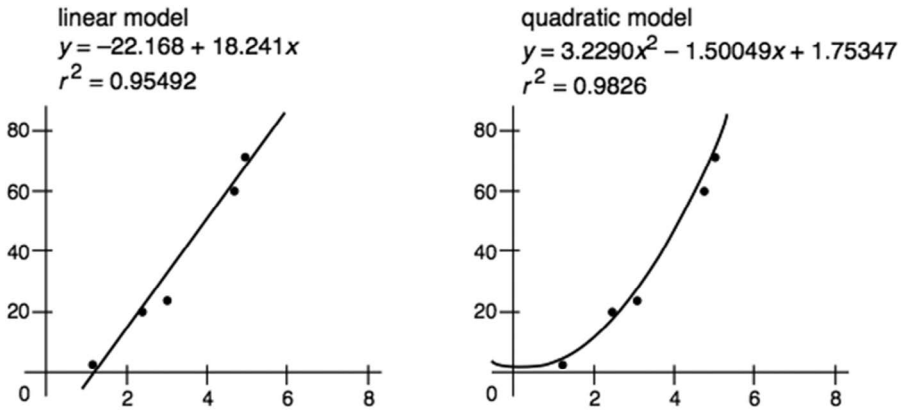


Figure 2.4: Different Models Have Different r^2 Values

For more detailed information on linear functions, watch the following tutorial:

Khan Academy: Comparing linear functions word problem

Descriptive Statistics

Mean, Standard Deviation, and Standard Error

You can describe the uncertainty in data by calculating the mean and the standard deviation. The **mean** of a set of data is the sum of all the measurement values divided by the number of measurements. If your data is a sample of a population, then the mean you calculate is an estimate of the mean of a population. The mean, $\overline{X}$, is determined using the formula

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + x_3 + \ldots}{n}$$

where x_1, x_2, etc. are the measurement values and n is the number of measurements.

Standard deviation is a measure of how spread out the data values are. If your measurements have similar values, then the standard deviation is small: each value is close to the mean. If your measurements have a wide range of values, then the standard deviation is high: some values may be close to the mean, but others are far from it. In general, if you make a large number of measurements, then the majority of them are within one standard deviation above or below the mean. (See the section on confidence intervals for a graph of the standard deviation ranges later in this chapter.)

Since standard deviations are a measure of uncertainty, they should be standard using only one significant digit. Standard deviation is commonly represented by the letter *s*. You calculate sample standard deviation using this formula:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

When you make multiple measurements of a quantity, the **standard error** (SE) of the data set is an estimate of the precision to which you know the mean of the quantity. The standard error is related to how spread out the data is, but it also includes the fact that when you measure a quantity many times and take the average of your measurements, you get a more precise value than if you only measure the quantity a few times. You calculate standard error using this formula:

$$\mathrm{SE} = \frac{s}{\sqrt{n}}$$

Because the number of measurements, n, is in the denominator, the more measurements you take, the smaller the standard error.

Example

Suppose you measure the following values for the temperature of a substance:

Trial	1	2	3	4
Temperature (°C)	20.5	22.0	19.3	23.0

The mean of the data is

$$\bar{x} = \sum_{i=1}^{4} \frac{x_i}{4} = \frac{20.5 + 22.0 + 19.3 + 23.0}{4} = 21.2^{\circ}\text{C}$$

The standard deviation of the data is

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{4} (x_i - \bar{x})^2}{4-1}}$$

$$= \sqrt{\frac{(20.5-21.2)^2 + (22.0-21.2)^2 + (19.3-21.2)^2 + (23.0-21.2)^2}{3}}$$

$$= 1.63 = 2 \text{ (rounded to one significant digit)}$$

The SE is

$$\text{SE} = \frac{s}{\sqrt{n}} = \frac{1.63}{\sqrt{4}} = 0.8 \text{ (rounded to one significant digit)}$$

Using one standard deviation, we would report the temperature of the substance as 21.2 ± 2°C, meaning the typical temperature is in a range that is 2° above or 2° below the mean temperature. Since we only have a few data values, a standard deviation of 2°C shows that most of the data values were close to the mean. However, if we had taken a large number of measurements the standard deviation would show that the majority (specifically, 68%; see the Confidence Intervals section later in this chapter) of the data values were between 19.2°C and 23.2°C. Alternatively, the data could be reported using the standard error as 21.2 ± 0.8°C. This tells us how our data compare to the true population mean with 95% confidence. In other words, because we took four measurements we have 95% confidence that the average temperature is within 0.8°C of 21.2°C. The standard error tells us how confident we are in our determination of the mean, while the standard deviation tells us how far we expect any individual measurement to be from the mean.

A graph of your data showing the statistics can clearly summarize the data in a way that is easy to understand and interpret.

Example

Suppose you conducted an investigation to determine if English ivy leaves in a shady area have a greater width than English ivy leaves in a sunny area. Table 2.3 shows the raw data from your experiment.

Table 2.3: Leaf Measurement Data

Shady Leaves (in cm)	Sunny Leaves (in cm)
3.7	3.2
5.2	3.5
5.4	4.1
5.7	4.3
5.8	4.4
5.8	4.6
6.0	5.0
6.1	5.0
6.5	5.2
6.5	5.2
6.6	5.3
6.8	5.4
7.0	5.6
7.3	5.7
7.3	5.7
7.4	5.8
7.7	6.0
7.9	6.0
8.0	6.4
8.1	6.5
8.1	6.7
8.2	6.7
8.3	7.1
8.9	7.1
9.0	7.1
9.4	7.3
9.9	7.5

Table 2.3: Leaf Measurement Data *(continued)*

Shady Leaves (in cm)	Sunny Leaves (in cm)
9.9	7.9
9.9	8.0
10.4	8.2

The statistics from each experimental group can be calculated and shown in a table such as table 2.4.

Table 2.4: Descriptive Statistics

	Shady Leaves	Sunny Leaves
Mean	7.43	5.88
Standard Deviation	1.63	1.32
N	30	30
Standard Error	0.30	0.24

Using this information, you can graph your data to visually compare the means of the two groups of leaves:

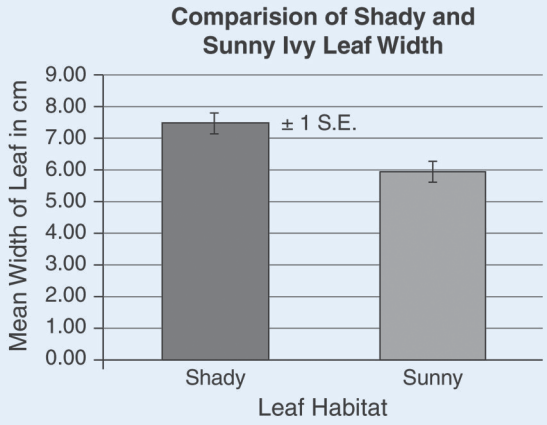


Figure 2.5: Comparison of Shady and Sunny Ivy Leaf Width

We see that the mean width of leaves gown in a shady environment is greater than the mean width of the leaves grown in a sunny environment. We also see that the error bars (± 1 SE) for the two means do not overlap. This supports our claim that the two populations are different, in other words, that English ivy leaves grown in a shady environment have a greater width than English ivy leaves grown in a sunny environment.

Confidence Intervals

A **confidence interval** is a range of values that the true value of the population has a probability of being within. If you measure a single quantity such as the mass of a certain isotope multiple times, you would expect a small standard deviation compared to the mean: the confidence intervals would be narrow. A wide confidence interval in this case would indicate the possibility of random errors in your measurements.

Confidence intervals can be presented in different ways. Figure 2.6 illustrates a commonly used method.

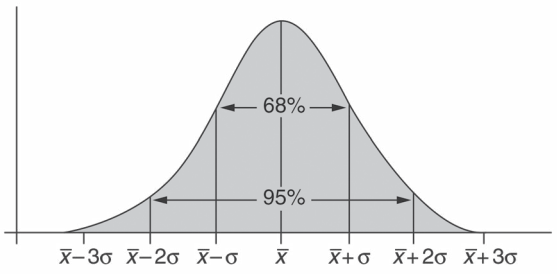


Figure 2.6: Confidence Intervals for a Normal Distribution

This method applies only to data that has a normal (bell-shaped) distribution. The mean lies at the peak of the distribution. Confidence intervals on either side of the peak describe multiples of the standard deviation from the mean. The percentage associated with each confidence interval (68%, 95%) has been determined by calculating the area under the curve.

A wide variety of data types in various subjects follow a **normal distribution** (i.e., a bell curve). In science, normal distributions apply to repeated measurements of a single value, such as multiple measurements of fluorescence decay time. A normal distribution is not appropriate when more than one central value is expected, or when only a few measurements are made.

If you need more information, the following tutorials can help to further explain these concepts:

Khan Academy: Measures of spread: range, variance & standard deviation

Khan Academy: Standard error of the mean

Khan Academy: Confidence interval 1

Accuracy, Precision, and Experimental Error

Communication of data is an important aspect of every experiment. You should strive to analyze and present data that is as accurate as possible. Keep in mind that in the laboratory neither the measuring instrument nor the measuring procedure is ever perfect. Every experiment is subject to experimental error. Data reports should describe the experimental error for all measured values.

Experimental error affects the accuracy and precision of data.

- **Accuracy:** how close a measurement is to a known or accepted value. For example, suppose the mass of a sample is known to be 5.85 g. A measurement of 5.81 g would be more accurate than a measurement of 6.05 g because 5.81 g is closer to actual value of the measurement.
- **Precision:** how close several measurements are to each other. The closer measured values are to each other, the higher their precision.

Measurements can be precise even if they are not accurate. Consider again a sample with a known mass of 5.85 g. Suppose several students each measure the sample's mass, and all of the measurements are close to 8.5 g. The measurements are precise because they are close to each other, but none of the measurements are accurate because they are all far from the known mass of the sample.

Systematic errors are errors that occur every time you make a certain measurement.

- They result in measurements that can be inaccurate or incorrect by making measurements that are consistently either higher or lower than they would be if there were no systematic errors.
- Examples include errors due to the calibration of instruments and errors due to faulty procedures or assumptions; for example, using a balance that is not correctly calibrated.

Random errors are errors that cannot be predicted.

- This includes errors of judgment in reading a meter or a scale and errors due to fluctuating experimental conditions.
- If the random errors in an experiment are small, the experiment is said to be *precise*.
- For example, when having numerous groups of students making temperature measurements of a classroom at the same time, they will have random variations due to local variation and instrument fluctuation.