AP® Physics 1 and 2 Lab Investigations

Student Guide to Data Analysis

Peter Sheldon, Randolph College, Lynchburg, VA

CollegeBoard
New York, NY

Contents

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 correct 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* describes how close a measurement is to a known or accepted value. Suppose, for example, the mass of a sample is known to be 5.85 grams. A measurement of 5.81 grams would be more accurate than a measurement of 6.05 grams. *Precision* describes 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 grams. Suppose several students each measure the sample's mass, and all of the measurements are close to 8.5 grams. 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. Examples include errors due to the calibration of instruments and errors due to faulty procedures or assumptions. These types of errors make measurements either higher or lower than they would be if there were no systematic errors. An example of a systematic error can occur when using a balance that is not correctly calibrated. Each measurement you make using this tool will be incorrect. A measurement cannot be accurate if there are systematic errors.

Random errors are errors that cannot be predicted. They include errors of judgment in reading a meter or a scale and errors due to fluctuating experimental conditions. Suppose, for example, you are making temperature measurements in a classroom over a period of several days. Large variations in the classroom temperature could result in random errors when measuring the experimental temperature changes. If the random errors in an experiment are small, the experiment is said to be *precise*.

Significant Digits

The data you record during an experiment should include only significant digits. Significant digits are the digits that are meaningful in a measurement or a calculation. They are also called **significant figures**. The measurement device you use determines the number of significant digits you should record. If you use a digital device, record the measurement value exactly as it is shown on the screen. If you have to read the result from a ruled scale, the value that you record should include each digit that is certain and one uncertain digit.

Return to
Table of Contents

Figure 1, for example, shows the same measurement made with two different scales. On the left, the digits 8 and 4 are certain because they are shown by markings on the scale. The digit 2 is an estimate, so it is the uncertain digit. This measurement has three significant digits, 8.42. The scale on the right has markings at 8 and 9. The 8 is certain, but you must estimate the digit 4, so it is the uncertain digit. This measurement is 8.4 centimeters. Even though it is the same as the measurement on the left, it has only two significant digits because the markings are farther apart.

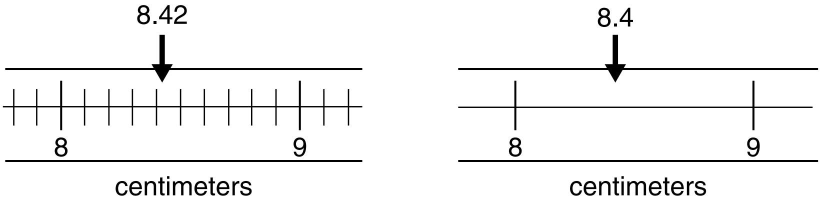


Figure 1

Uncertainties in measurements should always be rounded to one significant digit. When measurements are made with devices that have a ruled scale, the uncertainty is half the value of the precision of the scale. The markings show the precision. The scale on the left has markings every 0.1 centimeter, so the uncertainty is half this, which is 0.05 centimeter (cm). The correct way to report this measurement is $8.43 \pm 0.05\,\text{cm}$. The scale on the right has markings every 1 centimeter, so the uncertainty is 0.5 centimeter. The correct way to report this measurement is $8.4 \pm 0.5\,\text{cm}$.

The following table explains the rules you should follow in determining which digits in a number are significant:

Rule	Examples
Non-zero digits are always significant.	4,735 km has four significant digits. 573.274 in. has six significant digits.
Zeros before other digits are not significant.	0.38 m has two significant digits. 0.002 in. has one significant digit.
Zeros between other digits are significant.	42.907 km has five significant digits. 0.00706 in. has three significant digits. 8,005 km has four significant digits.
Zeros to the right of all other digits are significant if they are to the right of the decimal point.	975.3810 cm has seven significant digits. 471.0 m has four significant digits.
It is impossible to determine whether zeros to the right of all other digits are significant if the number has no decimal point.	8,700 km has at least two significant digits, but the exact number is unknown. 20 in. has at least one significant digits, but the exact number is unknown.
If a number is written with a decimal point, zeros to the right of all other numbers are significant.	620.0 km has four significant digits. 5,100.4 m has five significant digits. 670. in. has three significant digits.
All digits written in scientific notation are significant.	6.02×10^4 cm has three significant digits.

Analyzing data

Analyzing data may involve calculations, such as dividing mass by volume to determine density or subtracting the mass of a container to determine the mass of a substance. 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 least number of decimal places in any of the numbers that you are adding or subtracting.

The table below explains how the proper results should be written:

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 just one decimal place.
48.3506 m − 6.28 m = 42.10 m	The result is written with two decimal places because the number 6.28 has just two decimal places.
(8 km – 4.2 km) + 1.94 km = 6 km	The result is written with zero decimal places because the number 8 has zero decimal places.

Notice that the result of adding and subtracting has the correct number of significant digits if you consider decimal places. With multiplying and dividing, the result should have the same number of significant digits as the number in the calculation with the least number of significant digits.

The table below explains how the proper results should be written:

Example	Explanation
5.246 in. × 2.30 in. = 12.1 in.^{2}	The result is written with three significant digits because 2.30 has three significant digits.
0.038 cm ÷ 5.273 cm = 0.0072	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^2	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 round-off error. At the end of the calculation, round to the correct number of significant digits.

An exception to these rules is when a calculation involves an exact number, such as numbers of times a ball bounces or 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?

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

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

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 (a much larger data set), then the mean you calculate is an estimate of the mean of a population. The mean, $\overline{x}$, is determined using this formula:

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

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 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. If you make a large number of measurements, then the majority of the measurements are within one standard deviation above or below the mean. (See "Confidence Intervals" on page 6 for a graph of the standard deviation ranges.)

Since standard deviations are a measure of uncertainty, they should be standard using only one significant digit. Standard deviation is commonly represented by the Greek symbol sigma, σ, for data that is from a sample of a population; and by the symbol, s, for data that is from a sample.

You calculate standard deviation using this formula:

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

Return to
Table of Contents

When you make multiple measurements of a quantity, the **standard error**, SE, of the data set is an estimate of its precision. It is a measure of the data's uncertainty, but it reduces the standard deviation if a large number of data values are included. You calculate standard error using this formula:

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

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:

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

The standard deviation of the data is:

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

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

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

The standard error is:

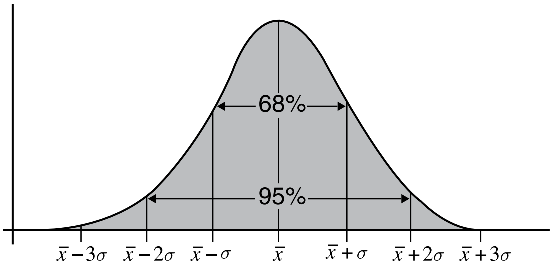
$$SE = \frac{s}{\sqrt{n}} = \frac{1.63}{\sqrt{4}} = 0.8 \text{ (rounded to one significant digit)}$$

Using the standard deviation, we would report the temperature as 21.2 ± 2°C. 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. If, however, we had taken a large number of measurements, the standard deviation would show that the majority (specifically, 68%; see "Confidence Intervals" below) 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.

Confidence Intervals

A **confidence interval** is a range of values within which the true value has a probability of being. 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, so 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. The following graph illustrates a method commonly used in physics:



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%, and so on) has been determined by calculating the area under the curve.

A wide variety of data types in various subjects follow a **bell curve** distribution. In physics, bell curves apply to repeated measurements of a single value, such as measuring fluorescence decay time. A bell-shaped distribution is not appropriate when more than one central value is expected, or when only a few measurements are made.

Propagation of Error

If calculations involve the results of two or more measurements, you must state the **combined uncertainty** of the measurements.

The combined uncertainty of quantities that are added or subtracted is the square root of the sum of the squares of their individual uncertainties. If, for example, you calculate a quantity $K = F + G - H$, where F, G, and H are measured values, and their uncertainties are ΔF, ΔG, and ΔH, where the Δ symbol, in this case, means "the uncertainty of." The uncertainty of K, then, is:

$$\Delta K = \sqrt{(\Delta F)^2 + (\Delta G)^2 + (\Delta H)^2}$$

Return to
Table of Contents

Example:

Suppose you measure the masses of two objects as 3.18 ± 0.01 kilograms and 2.184 ± 0.001 kilograms. The combined uncertainty is:

$$\Delta m_{combined} = \sqrt{(\Delta m_1)^2 + (\Delta m_2)^2} = \sqrt{(0.01)^2 + (0.001)^2} = 0.01$$

The sum of the masses would have three significant figures and their combined uncertainty should be recorded as 5.36 ± 0.01 kilograms.

To calculate the combined uncertainty of quantities that are multiplied or divided, the uncertainties must be divided by the mean values. Suppose that now $K = F \times G \div H$. The combined uncertainty when multiplying or dividing is:

$$\Delta K = |K| \sqrt{\left(\frac{\Delta F}{F}\right)^2 + \left(\frac{\Delta G}{G}\right)^2 + \left(\frac{\Delta H}{H}\right)^2}$$

Example:

Suppose you want to calculate the magnitude of the acceleration of an object. You measure the net force on the object, F = 1.49 ± 0.03 N, and the mass of the object, m = 3.42 ± 0.01 kilograms. The acceleration without the uncertainty is:

$$a = \frac{F}{m} = \frac{1.49\,\text{N}}{3.42\,\text{kg}} = 0.436\,\text{m/s}^2$$

The combined uncertainty is:

$$\Delta a = |a| \sqrt{\left(\frac{\Delta F}{F}\right)^2 + \left(\frac{\Delta m}{m}\right)^2} = |0.436| \sqrt{\left(\frac{0.03}{1.49}\right)^2 + \left(\frac{0.01}{3.42}\right)^2} = 0.009$$

The acceleration should then be recorded as $0.436 \pm 0.009\,\text{m/s}^2$.

Comparing Results: Percent Difference and Percent Error

If two lab groups measure two different values for an experimental quantity, you may be interested in how the values compare to each other. A large difference, for example, might indicate errors in measurements or other differences in measurement procedures. A comparison of values is often expressed as a **percent difference**, defined as the absolute value of the difference divided by the mean, with the result multiplied by 100:

$$\text{Percent difference} = \left| \frac{\text{value 1} - \text{value 2}}{\frac{1}{2}(\text{value 1} + \text{value 2})} \right| \times 100$$

You may instead want to compare an expected or theoretical value to a measured value. Knowing that your value is either close to or far from a known value can suggest whether your experimental procedure is reliable. In this case you can calculate the **percent error**, defined as the absolute value of the difference divided by the expected value, with the result multiplied by 100:

$$\text{Percent error} = \left|\frac{\text{measured value} - \text{expected value}}{\text{expected value}}\right| \times 100$$

Note that when the expected value is very small, perhaps approaching zero, the percent error gets very large because it involves dividing by a very small number. It is undefined when the expected value is zero. Percent error may not be a useful quantity in these cases.

Return to
Table of Contents

Graphs

Graphs are often an excellent way to present or to analyze data. When making graphs, there are a few guidelines you should follow to make them as clear as possible:

- Each axis should be labeled with the variable that is plotted and its units.
- Each axis should include a reasonable number of labeled tick marks at even intervals. Having too many tick marks will make the graph crowded and hard to read. Having too few will make the value of data points difficult to determine.
- Typically, graphs should be labeled with a meaningful title or caption.

Independent and Dependent Variables

When you graph data, you most often choose to plot an independent variable versus a dependent variable. The independent variable is plotted on the *x*-axis, and the dependent variable is plotted on the *y*-axis.

An **independent variable** is a variable that stands alone and isn't changed by the other variables you are trying to measure. For example, time is often an independent variable: in kinematics, distance, velocity, and acceleration are dependent on time, but do not affect time.

A **dependent variable** is something that depends on other variables. For example, in constant acceleration motion, position of a body will change with time, so the position of the body is dependent on time, and is a dependent variable.

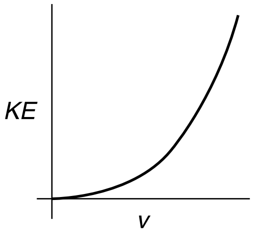
Graphing Data as a Straight Line

When you make a plot on *x*-*y* axes, a straight line is the simplest relationship that data can have. Graphing data points as a straight line is useful because you can easily see where data points belong on the line. A line makes the relationships of the data easy to understand.

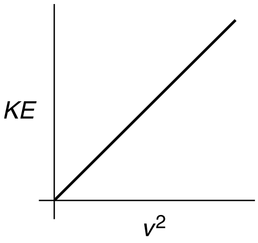
You can represent data as a straight line on a graph as long as you can identify its slope, m, and its *y*-intercept, b, in a linear equation: $y = mx + b$. The slope is a measure of how y varies with changes in x: $m = \Delta y / \Delta x$. The *y*-intercept is where the line crosses the *y*-axis (where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

Linearizing Data

Even if the data you take do not have a linear relationship, you may be able to plot it as a straight line by revising the form of the variables in your graph. One method is to change a relationship so that is 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 versus 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 Graph 1 below. But if we we set the horizontal axis variable equal to v^2 instead, the graph is linear, as shown in Graph 2:



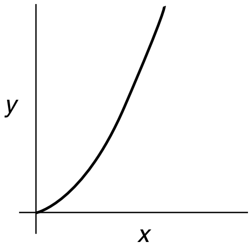
Graph 1



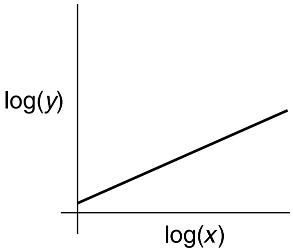
Graph 2

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 $\log(y) = \log(a) + n\log(x)$. 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 Graphs 3 and 4 below.



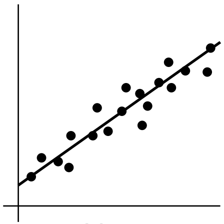
Graph 3



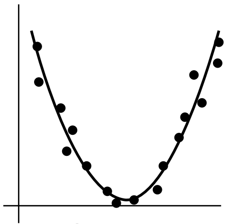
Graph 4

Curve Fitting

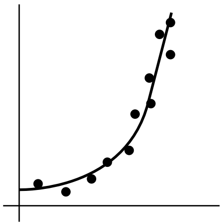
A useful way to analyze data is to determine whether it corresponds to a certain mathematical model. The first step is to plot the points and see if it follows a recognizable trend, such as a linear, quadratic, or exponential function. The graphs below show examples of each of these types.



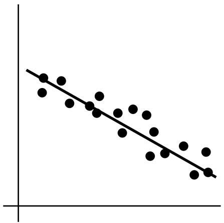
Linear



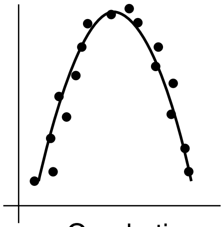
Quadratic



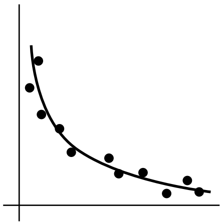
Exponential



Linear



Quadratic



Exponential

The general equation of a **linear function** is $y = mx + b$, where 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.

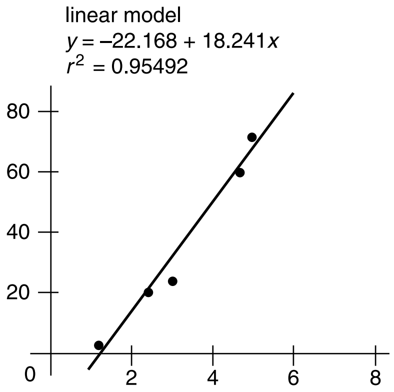
The general equation of a **quadratic function** is $y = ax^2 + bx + c$, where a, b, and c are arbitrary constants. An example of a quadratic function in physics is spring potential energy, $U = \frac{1}{2}kx^2$, where x is the distance the spring is stretched from equilibrium, k is the spring constant, 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.

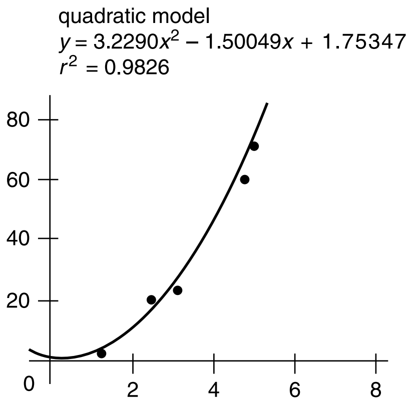
Return to
Table of Contents

If the pattern 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 points above and below the line. You can then determine an equation of the line by identifying the slope and *y*-intercept from 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. Graphs 5 and 6 below show 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 below, both models are a good fit for the data, but the r^2 values show that the quadratic model is better.



Graph 5



Graph 6

Helpful Links

"Averaging, Errors and Uncertainty." Department of Physics and Astronomy. University of Pennsylvania. Accessed January 6, 2015. https://www.physics.upenn.edu/sites/www.physics.upenn.edu/files/Error_Analysis.pdf.

"Excel 2013 Training Course, Videos and Tutorials." Microsoft Office. Accessed January 6, 2015. http://office.microsoft.com/en-us/excel-help/training-courses-for-excel-2013-HA104032083.aspx.

"Functions and Formulas." Google Help (for Google Sheets). Accessed January 6, 2015. https://support.google.com/docs/topic/1361471?hl=en&ref_topic=2811806.

"Intro to Excel." Department of Physics and Astronomy. University of Pennsylvania. Accessed January 6, 2015. https://www.physics.upenn.edu/sites/www.physics.upenn.edu/files/Introduction_to_Excel.pdf.

"Useful Excel Commands for Lab." Department of Physics. Randolph College. Accessed January 6, 2015. http://physics.randolphcollege.edu/lab/IntroLab/Reference/exchint.html.

Return to
Table of Contents