

Introduction to Data Analysis

Data Analysis Home

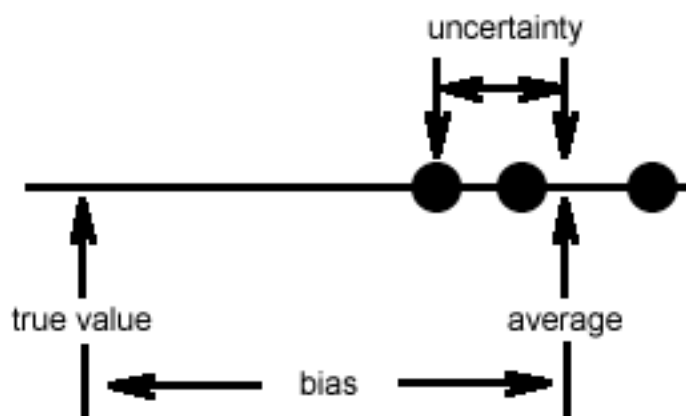
While working in lab this semester you will collect lots of data. If you are lucky your data will lead to a clear answer to the question you are investigating. At times, however, you will find your data ambiguous or, more interestingly, that something unexpected is hidden within the data. The modules on this site will guide you in exploring several important concepts in data analysis; these are:

- the **uncertainty** inherent in any measurement and how that uncertainty affects your results
- completing a **preliminary analysis** of your data by examining it visually and by characterizing it quantitatively
- **comparing data sets** to determine if there are differences between them
- modeling data using a **linear regression** analysis
- examining data for possible **outliers**

Every discipline has its own terminology. If you are unsure about a term's meaning, use the link on the left for the **On-Line Glossary**. The data sets in these modules are Excel files. The **How To...** link provides reminders on using Excel. Each of these resources opens in a new browser window so you can keep them open and available while working on a module.

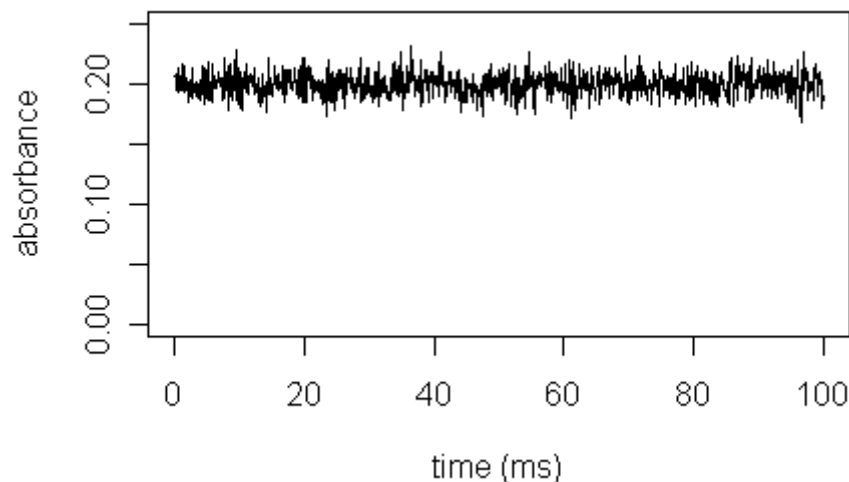
Uncertainty - Introduction

Suppose you are analyzing soil samples from a school playground near a busy intersection. After collecting three samples, you return to the lab, extract the soil samples with water and, using an atomic absorption spectrometer, determine the concentration of lead in the resulting extracts, obtaining an average result of 4.38 parts-per-billion (ppb). What kinds of errors might affect this result? As shown in this figure



the error has two parts. Any difference between the average result and the true value (there is a true value, even if we do not know it!) is a bias in our analysis. Possible sources of bias include, for example, failing to calibrate the instrument or contaminating the samples by collecting them in bottles that were not properly cleaned. Errors such as these are called determinate and they affect the accuracy of the result. Determinate errors are systematic, with a fixed value in a single direction relative to the true value.

The differences between the results for the three samples and their average (one example is labeled in the figure above) result from the uncertainty in making a measurement. An atomic absorption spectrometer, for example, works by drawing the sample into a flame where the flame's energy converts the aqueous lead ions into gaseous lead atoms. The flow of sample into the flame and the efficiency of atomization are not constant over time; instead, each varies around an average value. The result, as shown below, is a noisy signal.



If we divide the signal into five 20 ms periods, the resulting average signals will differ slightly from each other. Some of the results will be greater than 0.200 (the true value) and some results will be less than 0.200. This type of error is indeterminate and affects the precision (reproducibility) of our results, but not their accuracy. Indeterminate errors are random, having a value that is not constant and that may be either greater or smaller than the average value. The uncertainty of measurements is the focus of this module.

After you complete this module you should understand that:

- significant figures are one way to express a measurement's uncertainty
- for a result that depends on several different measurements, the uncertainty of each measurement contributes to the result's uncertainty

Begin with Problem 1.

Uncertainty - Problem 1

Let's begin with a simple exercise in which you will make two measurements and use them to complete a calculation. Click on the link labeled "Measurement Exercise," which opens in a new browser window. Complete the exercise and then proceed to the following tasks.

Task 1. Gather data from at least five classmates and compare their work with yours. Are your measurements of the rectangle's length and width exactly the same as those of your classmates, or are they different? If there are differences in the measurements, is this the result of determinate errors, indeterminate errors or both? Briefly explain your reasoning.

Task 2 . Compare your calculations for the rectangle's area to the results of your group of classmates. To what decimal place - hundreds, tens, ones, tenths, hundredths - are the areas in agreement? For example, the results 1413.2 and 1415.67 agree to the ten's place because, when moving from left to right, the ten's place is the last decimal place where the values are identical. Did you and your classmates report the rectangle's area to the same number of decimal places? If your answer is no, discuss some possible reasons for making different choices?

Task 3 . Based on your answers to Tasks 1-2, what is the your best estimate for the rectangle's area? What criteria did you use in arriving at your answer?

After completing these tasks, proceed to Problem 2.

Measurement Exercise

Print this page and, using a ruler with a millimeter scale, measure the rectangle's width and length in millimeters. In addition, calculate the rectangle's area.



Enter your values here:

Length = _____ Width = _____ Area = _____

Uncertainty - Problem 2

When measuring the rectangle's length and width and when calculating the rectangle's area you had to make several decisions:

- How do you define the rectangle? Should you measure its length and width from the outside edges of the border, from the inside edges or from the middle of the border?
- Should you estimate the rectangle's length and width to the nearest millimeter, or should you estimate between the millimeter markings and report the length and width to the nearest tenth of a millimeter?
- To how many decimal places should you report the area?

The first of these decisions is important because the border has a measurable thickness of approximately 1 mm. How you define the rectangle, therefore, affects your measurement of its length and width.

Task 1. Suppose a rectangle's "official" definition includes the entire border. If an analyst measures the rectangle's length using the border's inside edge, has he or she made a determinate or indeterminate error? Explain.

Task 2. Your second decision is important because estimating introduces uncertainty. If you report an object's length as 154 mm, you indicate absolute certainty for the digits in the hundred's and ten's place and uncertainty about the digit in the one's place. If the uncertainty in the last digit is ± 1 mm, for example, then the length might be as small as 153 mm and as large as 155 mm. On the other hand, if you reported an object's length as 154.3 mm and the uncertainty is ± 0.2 mm, then the length is between 154.1 mm and 154.5 mm.

You may recall that the we refer to the digits in our measurements as significant figures. Basically, a significant figure is any number in which we can express confidence, including those digits known exactly and the one digit whose value is an estimate. The lengths 154 mm and 154.3 mm have three and four significant digits, respectively. The number of significant figures in a measurement is important because it affects the number of significant figures in a result based on that measurement.

Suppose the length and width of a rectangle are 154 mm and 32 mm, what is the rectangle's area to the correct number of significant figures? What if the length is 154.3 mm and the width is 31.8 mm? Click [here](#) for a review of significant figures.

When you are finished with these tasks, proceed to Problem 3.

Uncertainty - Problem 3

Before continuing, let's review the rules for including significant figures in calculations. When adding or subtracting, the result of the calculation is rounded to the last decimal place that is significant for all measurements. For example, the sum of 135.621, 0.33 and 21.2163 is 157.17 since the last decimal place that is significant for all three numbers (as shown below by the vertical line)

$$\begin{array}{r} 135.621 \\ 0.33 \\ 21.2163 \\ \hline 157.1673 \end{array}$$

is the hundredth's place. Note that rounding the answer to the correct number of significant figures occurs after completing the exact calculation.

When multiplying or dividing, the result of the calculation contains the same number of significant figures as that measurement having the smallest number of significant figures. Thus,

$$\frac{22.91 \times 0.152}{16.302} = 0.21361 \approx 0.214$$

because 0.152, with three, has the fewest number of significant figures.

The reason for these rules is that the uncertainty in a result cannot be less than the uncertainty in the measurements used in calculating the result. One way to appreciate these rules is to calculate the largest possible results and the smallest possible result (the "Worst Case Scenarios") by taking into account the uncertainties in each measurements.

Consider the first problem at the top of the page. If we ignore significant figures, the exact answer to the problem is 157.1673. Suppose that the uncertainty in each value is ± 1 in its last decimal place. For example, the measurement 135.621 could be as large as 135.622 or as small as 135.620. The largest possible sum of the three measurements, therefore, comes from adding together the largest possible measurements and the smallest possible sum comes from adding together the smallest possible measurements; thus

$$\begin{array}{r} 135.622 \\ 0.34 \\ 21.2164 \\ \hline 157.1784 \end{array} \qquad \begin{array}{r} 135.620 \\ 0.32 \\ 21.2162 \\ \hline 157.1562 \end{array}$$

Comparing the two worst case results and the exact result, we see that rounding to the hundredth's place is the first instance where there is no agreement between the three calculations. The result of the exact calculation, therefore, is rounded to the hundredth's place, giving 157.17 as the appropriate result of the calculation.

For the second problem, the exact answer is 0.213613 (to six decimal places). To obtain the largest possible answer we divide the largest possible numerator by the smallest possible denominator, and to obtain the smallest possible answer we divide the smallest possible numerator by the largest possible denominator. If the uncertainty in each measurement is ± 1 in its last decimal place, then the largest and smallest possible answers are

$$\frac{22.92 \times 0.153}{16.301} = 0.215125 \quad \frac{22.90 \times 0.151}{16.303} = 0.212102$$

Comparing the two worst case results and the exact result, we see that rounding to the thousandth's place is the first instance where there is no agreement between the three calculations. The result of the exact calculation, therefore, is rounded to the thousandth's place, giving 0.214.

Try your hand at these two problems.

Task 1. The mass of water in a sample is found by weighing the sample before and after drying, yielding values of 0.4991 g and 0.3715 g, respectively. How many grams of water are in the sample and what is the percent water by mass? To how many significant figures should you report your answer?

Task 2. Assuming that the uncertainty in measuring mass is ± 0.0001 g, what is the largest and the smallest possible mass of water in the sample? What is the largest and the smallest possible result for the percent water in the sample? To how many significant figures should you report your answer? Does this answer agree with the significant figure rules?

After you complete these tasks, read the module's summary.

Uncertainty - Summary

After completing this module you should more fully understand the importance of making proper use of significant figures in your calculations. You should also appreciate that the uncertainty of your experimental work is limited by the measurement with the greatest uncertainty. When planning an experiment, think carefully about how you will make use of your measurements. For example, suppose you are given the following directions:

Add 50 mL of water to a 250-mL volumetric flask. Transfer about 0.1 g of KCl to the volumetric flask, swirl until the solid dissolves and then dilute to volume. Calculate the exact concentration of KCl in your solution.

To calculate the concentration of KCl you need to know the mass of KCl. You will want, therefore, to measure this using a balance that weighs samples to the nearest ± 0.001 g or ± 0.0001 g. If available, the later choice is best as it provides a smaller uncertainty. Making a decision to replace the volumetric flask with an Erlenmeyer flask is a mistake since the greater uncertainty in an Erlenmeyer flask's volume increases your uncertainty in the concentration of KCl. On the other hand, the 50 mL of water does not appear in your calculation; for this reason it is not necessary to measure accurately this volume and a graduated cylinder will suffice.

Additional information on the topics covered in this module is available using the link on the left for further study.

Uncertainty - Further Study

Limitations to Significant Figures. Although significant figures are important, there are situations where they do not lead to a correct estimate of uncertainty. Consider the following example:

The mass of water in a sample is found by weighing the sample before and after drying, yielding values of 0.4991 g and 0.4174 g. What is the percent water by mass in the sample? The uncertainty in each mass is ± 0.0001 g.

Using significant figures to account for uncertainty gives the mass percent of water as

$$100 * \{(0.4991 \text{ g} - 0.4174 \text{ g})/0.4991 \text{ g}\} = 100 * \{0.0817 \text{ g}/0.4991 \text{ g}\} = 16.3695\%$$

which, allowing for three significant figures, rounds to 16.4%. Using worst case scenarios, the largest and smallest possible results are

$$100 * \{(0.4992 \text{ g} - 0.4173 \text{ g})/0.4990 \text{ g}\} = 16.4128\%$$

$$100 * \{(0.4990 \text{ g} - 0.4175 \text{ g})/0.4992 \text{ g}\} = 16.3261\%$$

Comparing the two worst case results and the exact result, we see that rounding to the hundredth's place is the first instance where there is no agreement between the three calculations. The result of the exact calculation, therefore, is rounded to four significant figures, giving 16.40% as the appropriate result of the calculation.

Worst Case Scenario. Four simple rules will help you make use of this approach to estimating uncertainty.

Rule One. If two values are added, the limits for the result are obtained by adding the upper limits and by adding the lower limits. For example, if X is 7.0 ± 0.1 and Y is 5.0 ± 0.1 , then the upper limit is

$$X + Y = 7.1 + 5.1 = 12.2$$

and the lower limit is

$$X + Y = 6.9 + 4.9 = 11.8$$

Rule Two. If two values are multiplied, the limits for the result are obtained by multiplying the upper limits and by multiplying the lower limits. For example, if X is 7.0 ± 0.1 and Y is 5.0 ± 0.1 , then the upper limit is

$$X * Y = 7.1 * 5.1 = 36.21$$

and the lower limit is

$$X * Y = 6.9 * 4.9 = 33.81$$

Rule Three. If two values are subtracted, the limits for the result are obtained by subtracting the lower limit of one number from the upper limit of the other number and by subtracting the upper limit of one number by the lower limit of the other number. For example, if X is 7.0 ± 0.1 and Y is 5.0 ± 0.1 , then the upper limit is

$$X - Y = 7.1 - 4.9 = 2.2$$

and the lower limit is

$$X - Y = 6.9 - 5.1 = 1.8$$

Rule Four. If two values are divided, the limits for the result are obtained by dividing the upper limit of one number by the lower limit of the other number and by dividing the lower limit of one number by the upper limit of the other number. For example, if X is 7.0 ± 0.1 and Y is 5.0 ± 0.1 , then the upper limit is

$$X/Y = 7.1/4.9 = 1.44898$$

and the lower limit is

$$X/Y = 6.9/5.1 = 1.35294$$

A useful article discussing this approach is Gordon, R.; Pickering, M.; Bisson, D. "Uncertainty Analysis the 'Worst Case' Method" *J. Chem. Educ.* **1984**, *61*, 780-781.

Propagation of Uncertainty. A more rigorous approach to determining the uncertainty in a result is called a propagation of uncertainty. You can read more about this approach [here](#) and click [here](#) for a summary of equations. The following [applet](#) provides a useful calculator for determining uncertainty (note that you can download this applet to your computer).

Preliminary Analysis of Data - Introduction

When you gather data in lab you typically are trying to explain the relationship between a particular response and one or more factors. For example, you might be interested in how the concentration of lead in soil (the response) varies as a function of the soil's distance from a highway (the factor). After collecting samples from different locations and completing the analysis you have a data set consisting of lead concentrations and distances from the highway.

How might you begin to analyze your data. Two common approaches are to examine the data visually, looking for interesting relationships between the response and the factors, and to summarize quantitatively the data.

After you complete this module you should:

- understand the difference between a population and a sample
- appreciate the importance of visually examining your data
- be able to summarize your data quantitatively

To work through this module we need a question to investigate, which for us will be:

What is the mass of a US penny in circulation?

Your goal is to report the best estimate for a US penny's mass, as well as a sense of the variability in that mass. Begin with Problem 1.

Preliminary Analysis of Data - Problem 1

One approach to this problem is to collect a sample of pennies (perhaps from your change jar) and to measure the mass of each penny. Open Data Set 1, using the link on the left, which is an Excel file containing the masses for 32 pennies. Note that the data consists of one response (the mass of the penny) and one factor (each penny's ID number).

Task 1. Characterize the data quantitatively by calculating the mean and standard deviation. (Use the link on the left for help with using Excel.) In several sentences, explain the meaning of these two values and what they suggest about this set of data.

Task 2. Characterize the data visually by creating a scatterplot with mass on the y-axis and the ID number on the x-axis. Be sure to scale the axes so that the data occupies most of the available space. In several sentences, explain what information this graph conveys about the data. In what ways do your quantitative and visual characterizations of the data provide similar and/or different information about the data?

Task 3. The visual presentation of the data should strike you as interesting and unusual. Look carefully at your graph of the data. What does it suggest about this sample of pennies? (If you are stuck, try this [hint](#)). Can you think of factor(s) that might explain the variation in the mass of these 32 pennies?

Task 4. Estimate the uncertainty in measuring the mass of a single penny. For example, if a penny has a mass of 2.512 g, is the uncertainty in its mass 0.1 g, 0.01 g, 0.001 g or 0.0001 g? Compare your estimate for the uncertainty in a penny's mass with your calculated standard deviation. Is this comparison consistent with your conclusions from Task 3? Explain.

After completing these tasks, proceed to Problem 2.

Hint

Where along the y-axis do you find the masses of pennies?

Where along the y-axis do you not find the masses of pennies?

Is this pattern what you might expect for a random sample of pennies?

Preliminary Analysis of Data - Problem 2

A preliminary analysis of the data from Problem 1 suggests that the pennies come from two populations, one clustered around a mass of 2.5 g and the other clustered around a mass of 3.1 g. A possible factor that might help in further analyzing this data is the age of the pennies. Open Data Set 2, using the link to the left, which is an Excel file containing the masses and years of minting for the 32 pennies in Data Set 1.

Task 1. Characterize the data visually by creating a scatterplot with mass on the y-axis and the year of minting on the x-axis. Be sure to scale the axes so that the data occupy most of the available space. In several sentences, explain what information this graph conveys about the data. As part of your answer, be sure to include the terms population and sample, and to offer at least one plausible explanation for any trends you see in this data.

Task 2. Based on your results from Task 1, divide the data into two groups and determine the mean and standard deviation for each. You first might find it helpful to sort the data. Compare your estimate for the uncertainty in a penny's mass (Problem 1 - Task 4) to your calculated standard deviations. Do these comparisons support your conclusions from Task 1? Explain. Does dividing the pennies into two groups explain all the uncertainty in your data? Explain. If no, then suggest some other sources of uncertainty.

Task 3. Clearly there is an interesting trend to this data. How might you redesign this experiment to provide a more complete answer to the original question: What is the mass of a US penny?

After completing these tasks, proceed to Problem 3.

Preliminary Analysis of Data - Problem 3

One limitation to Data Sets 1 and 2 is the absence of any information on pennies minted in the years 1980 to 1983, a gap that includes the period when the mean mass of a penny changed from approximately 3.1 g to approximately 2.5 g. A better experimental design would first divide the population of all pennies into separate populations organized by the year of minting. A random sample can then be drawn from each of these populations.

Task 1. Open Data Set 3, using the link on the left, and examine the data on Sheet 1, which provides the mean mass for samples of 10 pennies drawn from the populations of pennies minted in the years 1977 to 1987. Examine the data and explain what it suggests about the change in a US penny's mass over time.

Task 2. Examine the data on Sheet 2, which includes the standard deviations for the data on Sheet 1. Does this additional information alter your conclusions from Task 1? If so, what are your new conclusions?

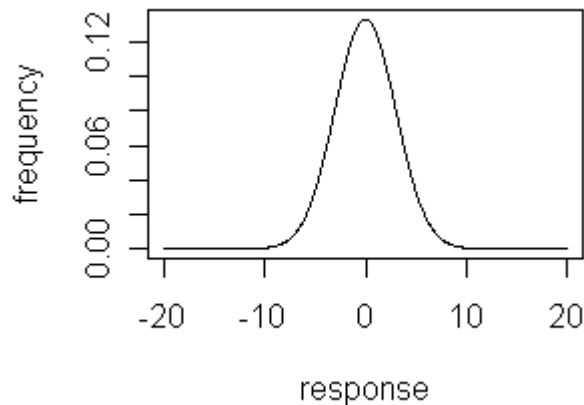
Task 3. Examine the data on Sheet 3, which provides the mass for all 110 pennies included in this study. Does this additional information confirm (or alter) your conclusions from Tasks 1 and 2? Explain.

After completing these tasks, proceed to the module's summary using the link on the left.

Preliminary Analysis of Data - Summary

After completing this module you should more fully understand the importance of considering the population from which a sample is drawn. The population of US pennies is not singular. Our data, which is limited to pennies minted between 1972 and 1992, show that there are two distinct populations and that a change in the composition of a penny must have occurred in 1982. There have been other changes in the penny, such as different designs, so there might be other populations that we have not considered in this module.

Trying to reach a conclusion about the mass of a penny using samples drawn from more than one population leads to an error in our analysis. The mean of 2.769 g reported in Task 1 of Problem 1, which did not consider the year of minting, is meaningless because no single penny can have this mass. For data drawn from a single population that follows a normal distribution (the classic "bell-shaped curve"),



which often is typical of the data we collect in lab, the sample's mean approximates the population's mean and the sample's standard deviation approximates the population's standard deviation.

Another important lesson from this module is the importance of examining your data visually. Much useful information and possibly some surprising trends may be evident in such plots.

Additional information on the topics covered in this module is available using the link on the left for further study.

Preliminary Analysis of Data - Further Study

Descriptive Statistics: In addition to the mean and standard deviation, there are other useful descriptive statistics for characterizing a data set. The median, for, example, is an alternative to the mean as a measure of a data set's central tendency. The median is found by ordering the data from the smallest-to-largest value and reporting the value in the middle. For a data set with an odd number of data points, the median is the middle value (after reordering); thus, for the following data set

3.086, 3.090, 3.097, 3.103, 3.088

the reordered values are

3.086, 3.088, 3.090, 3.097, 3.103

and the median is the third value, or 3.090. For a data set with an even number of data points, the median is the average of the middle two values (after reordering); thus, for the following data set

2.521, 2.511, 2.530, 2.506, 2.487, 2.532

the reordered values are

2.487, 2.506, 2.511, 2.521, 2.530, 2.532

and the median is the average of 2.511 and 2.521, or 2.516.

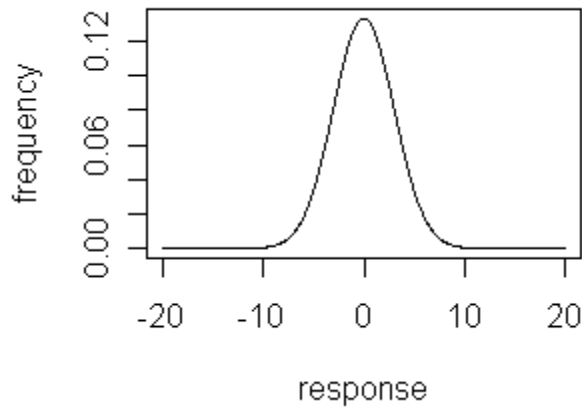
The range, which is the difference between the largest and smallest values in a data set, provides an alternative to the standard deviation as a measure of the variability in the data set. For example, the range for the two data sets shown above are 0.017 and 0.045, respectively.

Return to [Data Set 1](#) and report the median and range for the 32 pennies.

Follow this [link](#) for a further discussion of descriptive statistics. This [link](#) provides instructions on using Excel to report descriptive statistics. An on-line calculator can be accessed [here](#).

Alternative Plots: A scatterplot is one of several ways to display data graphically. Other useful plots are boxplots and histograms. This [site](#) provides a very brief introduction to these plots. The link on the site to "Questions on this subject" provides you with an opportunity to view boxplots, histograms and summary statistics for a data set consisting of 19 properties for 120 organic molecules. Experiment with creating different plots of the data. This [site](#) provides instructions for creating boxplots in Excel.

Normal Distributions: The summary to this module notes that the data we collect in lab typically follows a normal distribution (the classic "bell-shaped curve").



It is not intuitively obvious, however, that this is true. The applet at this [site](#) demonstrates that regardless of the shape of the underlying population's distribution, a normal distribution occurs if we gather samples of sufficient size.

Population 1, for example, consists of 101 possible responses (0, 1, 2, ..., 98, 99, 100) each of which is equally probable. The mean for the data set is 50.0. Set the sample size to 1, which means that we will draw one sample at random from this population, record its value and then continue in this fashion. Click on the Draw button and note that the distribution for the results is similar to the distribution for the underlying population and that the mean is approximately 50.

Next, set the sample size to 2, which means that we will draw two samples from the population and record their average. Click on the Draw button and note that the distribution for the results no longer looks like the distribution for the underlying population, but that the mean value continues to be approximately 50. Repeat using sample sizes of 3, 4 and 5 and note how the distribution for results increasingly takes on the shape of a normal distribution. Try using some of the other distributions as well.

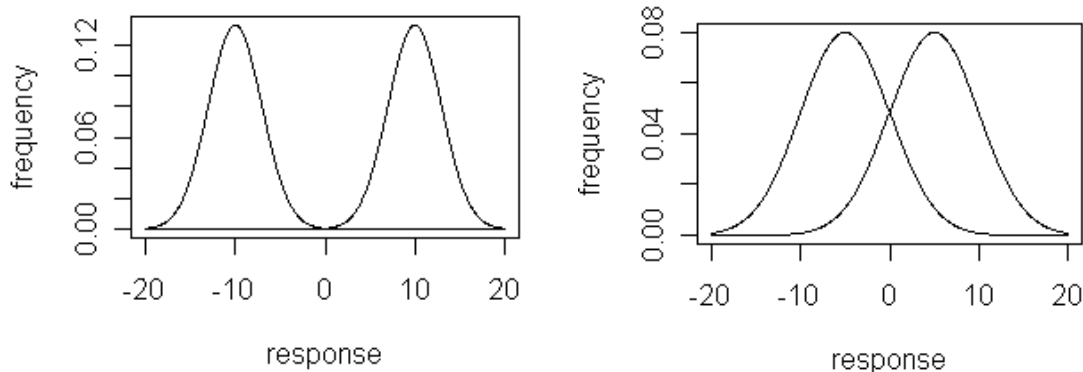
The tendency of results to approximate a normal distribution when using large samples is known as the Central Limit Theorem. The samples we analyze in lab generally contain large numbers of individual particles. Consider, for example, a sample of soil, which is not a single particle, but a complex mixture of many individual particles. When determining the concentration of Pb in a soil sample, the result that we report is the average of the amount of Pb in each particle. The large number of particles in a single soil sample tends to favor a normal distribution for our results.

You should be aware that there are other types of distributions for data, such as the binomial and Poisson distributions, but they are less commonly encountered.

Comparing Data Sets - Introduction

In the Preliminary Analysis module you inferred that pennies minted between 1977 and 1987 come from two populations, with the dividing line occurring in 1982. How confident are you, however, that a mean mass of 3.093 g for a penny minted in 1979 is significantly different than a mean mass of 2.513 g for a penny minted in 1985?

The answer to this question depends not only on the mean values, but also on their respective standard deviations. Consider the two figures shown below:



Each figure shows two normally distributed populations. For each population the distribution's maximum frequency corresponds to the population's mean and the distribution's width is proportional to the population's standard deviation.

The two populations in the figure on the left clearly are well separated from each other and we can state confidently that a sample drawn from one population will have a mean that is significantly different than the mean for a sample drawn from the other population. The two populations on the right, however, are more problematic. Because there is a significant overlap between the two populations, the mean for a sample drawn from one population may be similar to that for a sample drawn from the other population. Fortunately, there are statistical tools that can help us evaluate the probability that the means for two samples are different.

After you complete this module you should:

- appreciate how the mean and standard deviations for two data sets affect our ability to determine if they come from different populations
- be able to use an applet to carry out a t-test for comparing two samples and understand how to interpret the result of this statistical test

Before tackling some problems, read a description of the t-test by following the link on the left.

Comparing Data Sets - The t-Test

The standard approach for determining if two samples come from different populations is to use a statistical method called a t-test. Although we will not worry about the exact mathematical details of the t-test, we do need to consider briefly how it works.

Recall that a population is characterized by a mean and a standard deviation. Suppose that for the population of pennies minted in 1979, the mean mass is 3.083 g and the standard deviation is 0.012 g. Together these values suggest that we will not be surprised to find that the mass of an individual penny from 1979 is 3.077 g, but we will be surprised if a 1979 penny weighs 3.326 g because the difference between the measured mass and the expected mass (0.243 g) is so much larger than the standard deviation. In fact, we can express this probability as a confidence interval; thus:

- 68.3% of 1979 pennies will have a mass of $3.083 \text{ g} \pm 0.012 \text{ g}$ (± 1 std dev)
- 95.4% of 1979 pennies will have a mass of $3.083 \text{ g} \pm 0.024 \text{ g}$ (± 2 std dev)
- 99.7% of 1979 pennies will have a mass of $3.083 \text{ g} \pm 0.036 \text{ g}$ (± 3 std dev)

The probability of finding a 1979 penny whose mass is outside the range of 3.047 g - 3.119 g, therefore, is 0.3%. These probabilities hold for a single sample drawn from any normally distributed population. Thus, there is a 99.7% probability that a measurement on any single sample will be within ± 3 standard deviation of the population's mean.

We also can extend the idea of a confidence interval to larger sample sizes, although the width of the confidence interval depends on the desired probability and the sample's size. As we did above, let's assume that the population of 1979 pennies has a mean mass of 3.083 g and a standard deviation of 0.012 g. This time, instead of stating the confidence interval for the mass of a single penny, we report the confidence interval for the mean mass of 4 pennies; these are:

- 68.3% of 1979 pennies will have a mass of $3.083 \text{ g} \pm 0.006 \text{ g}$ (± 1 std dev)
- 95.4% of 1979 pennies will have a mass of $3.083 \text{ g} \pm 0.012 \text{ g}$ (± 2 std dev)
- 99.7% of 1979 pennies will have a mass of $3.083 \text{ g} \pm 0.018 \text{ g}$ (± 3 std dev)

Note that each confidence interval is half of that for the mass of a single penny.

Now we are ready to consider how a t-test works. Suppose that we want to determine if two samples are different and that we want to be at least 95% confident in reaching this decision. We analyze each sample and determine their respective means and standard deviations. For each sample we can represent the confidence interval using a solid circle to represent the sample's mean and a line to represent the width of the sample's 95% confidence interval. If the 95% confidence intervals for the two samples do not overlap, as shown in case 1 below, then we can state that we are at least 95% confident that the two samples come from different populations. Note that there is no more than a 5% probability that this conclusion is incorrect.



On the other hand, if the 95% confidence intervals overlap, then we cannot be 95% confident that the samples come from different populations and we conclude that we have insufficient evidence to determine if the samples are different. Note that we are not 95% confident that the samples are the same; this is a subtle, but important point.

When you are ready, proceed to Problem 1 using the link on the left.

Comparing Data Sets - Problem 1



(Applet courtesy of Prof. C. E. Efstathiou, http://www.chem.uoa.gr/applets/Applet_Index2.htm)

In this problem you will use an applet to explore how the means, standard deviations and sizes for two samples affects the ability of a t-test to discriminate between the samples.

Task 1. With the "Demo data" radio button selected, enter three data points for Data Set A at approximately 48, 48.5 and 49. (To place a point, position the cursor and left-click with your mouse; each point appears as a red dot and a red line.) Add three data point for Data Set B at approximately 51, 51.5 and 52.

Click on the CALCULATE button to display the results. The number of data points, the mean and the standard deviation for each data set are in shown in the box to the right, and the mean and standard deviation are superimposed on the data as vertical and horizontal blue lines. Of particular interest to us is the confidence levels (CL) for the t-test . The statement "The means ARE different at CL 95%" indicates that that there is at least a 95% probability that the two samples are from different populations. A statement such as "The means ARE NOT different at CL 95%" means that we cannot be 95% sure that the samples are different. The value for P(type 1 error) also provides useful information; the exact confidence level at which we can show a difference between the two sample's is

$$100 * \{1.000 - P(\text{type 1 error})\}$$

For example, if P(type 1 error) is 0.212, then we can be

$$100 * (1.000 - 0.212) = 78.2\%$$

78.2% confident that the samples are different (and there is a 21.2% probability that this conclusion is incorrect).

Examine the calculated mean and standard deviation for each data set and compare the numerical results to the visual picture provided by the horizontal and vertical blue lines. Do the two data sets overlap? Does it appear that the data sets represent different populations? What does the t-test report state?

Task 2. Clear the data sets by clicking on the button labeled CLEAR. Create two new data sets by adding points at approximately 49, 49.5 and 50 for Data Set A and at approximately 50, 50.5 and 51 for Data Set B. How confident can you be that these two data sets come from different populations?

Add an additional data point to Data Set A between 49 and 50. How does this additional point change your analysis? Add an additional data point to Data Set B between 50 and 51. How does this additional point change your analysis? Continue adding data points to

the two data sets (between 49 and 50 for Data Set A and between 50 and 51 for Data Set B) until the difference between the two data sets is significant at the 99% CL. What happens to the means and standard deviations as you add more samples? How many samples did you need?

Task 3. Clear the data sets and create two new data sets by adding points at approximately 49.5, 50 and 50.5 for Data Set A and at approximately 50, 50.5 and 51 for Data Set B. Continue to add data points to each set (between 49.5 and 50.5 for Data Set A and between 50 and 51 for Data Set B) until you can show that the samples are different at the 99% CL. How many samples did you need?

Task 4. Briefly summarize your general conclusions from these three tasks. In your answer, consider how factors such as the location of the means, the size of the standard deviations and the size of the samples affect the results of the t-test.

When you are done, proceed to Problem 2.

Comparing Data Sets - Problem 2



(Applet courtesy of Prof. C. E. Efstathiou, http://www.chem.uoa.gr/applets/Applet_Index2.htm)

Task 1. Now that you are familiar with the applet and with the t-test, let's use it to analyze some real data. Click on the radio button labeled "User data," which replaces the two axes with two small data sets. Click on Calculate and interpret the results of this t-test. Add two additional points with values of 3.5 and 3.6 to the data set on the left (just click in the box and enter the values on separate lines). Do these additional points change the results of the t-test.

Task 2. Clear the data and use the link on the left to open a spreadsheet containing the masses for three samples of pennies. Is there any evidence for a difference between the samples from 1978 and 1980? Between 1978 and 1984? Are these results consistent with your conclusions from the Preliminary Analysis module?

Proceed to the module's summary using the link on the left.

Comparing Data Sets - Summary

After completing this module you should understand that there is more to determining if two samples are different than simply comparing their respective means. Although the difference between the means is important, so, too, are the standard deviations and the sizes of the samples. You also should understand that there are statistical methods for determining if two samples are different, but that such methods can establish only a level of confidence in the conclusion and are subject to error. Finally, you should feel comfortable using the t-test applet to compare data you collect in lab.

Additional information on the topics covered in this module is available using the link on the left for further study.

Comparing Data Sets - Further Study

Mathematical Details: Our treatment of the t-test has been limited. In particular, the mathematical details have been omitted. This [site](#) provides the general equations for using a t-test to compare two data sets, as well as for other types of t-tests. This [site](#) also provides useful details and an on-line calculator.

For further information, including a discussion of more esoteric (but important) considerations, such as one-tailed vs. two-tailed t-tests, type 1 vs. type 2 errors and paired t-tests consult the following textbook:

Miller, J. C.; Miller, J. N. Statistics for Analytical Chemistry, Ellis Horwood: Chichester

What if I have more than two data sets?: When you have more than two data sets, using a t-test to make all possible comparisons is not a good idea. Each t-test has a certain probability of yielding an incorrect result and when used multiple times, the probability increases that at least one conclusion will be incorrect. Instead, you should use an analysis of variance (ANOVA). This [site](#) provides a brief discussion of ANOVA and an on-line calculator.

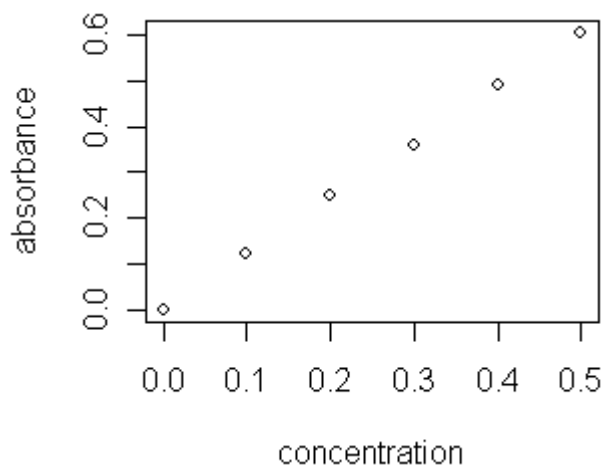
Using Excel for a t-Test: Links are provided here for instructions on using Excel to:

- [compare the mean value for a data set to a known value](#)
- [compare the mean values for two data sets \(unpaired data\)](#)
- [compare the mean values for two data sets \(paired data\)](#)

This [handout](#) provides an explanation of the difference between paired and unpaired data.

Linear Regression - Introduction

In lab you frequently gather data to see how a factor affects a particular response. For example, you might prepare a series of solutions containing different concentrations of Cu^{2+} (the factor) and measure the absorbance (the response) for each solution at a wavelength of 645 nm. A scatterplot of the [data](#)



shows what appears to be a linear relationship between absorbance and $[\text{Cu}^{2+}]$. Fitting a straight-line to this data, a process called linear regression, provides a mathematical model of this relationship

$$\text{absorbance} = 1.207 * [\text{Cu}^{2+}] + 0.002$$

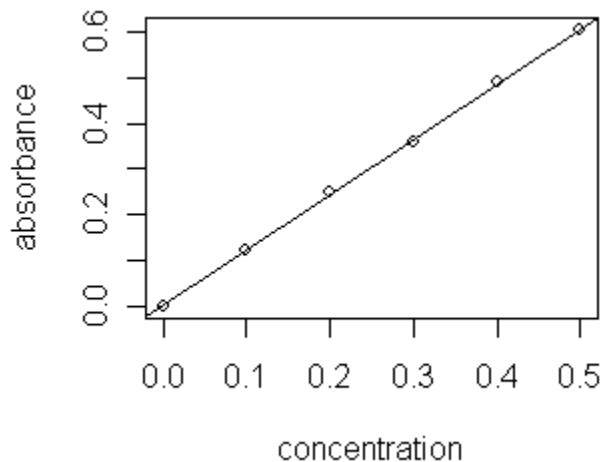
that can be used to find the $[\text{Cu}^{2+}]$ in any solution by measuring that solution's absorbance. For example, if a solution's absorbance is 0.555, the concentration of Cu^{2+} is

$$0.555 = 1.207 * [\text{Cu}^{2+}] + 0.002$$

$$0.555 - 0.002 = 0.553 = 1.207 * [\text{Cu}^{2+}]$$

$$0.553 / 1.207 = [\text{Cu}^{2+}] = 0.458 \text{ M}$$

A scatterplot showing the data and the linear regression model



suggests that the model provides an appropriate fit to the data. Unfortunately, it is not always the case that a straight-line provides the best fit to a data set. The purpose of this module is to emphasize the importance of evaluating critically the results of a linear regression analysis.

After you complete this module you should:

- appreciate how a linear regression analysis works
- understand that it is important to examine visually your data and your linear regression model

Before tackling some problems, use the link on the left to read an explanation of how linear regression works.

Data

Here is the data used in this example:

| [Cu ²⁺] (M) | Absorbance |
|-------------------------|------------|
| 0.000 | 0.000 |
| 0.100 | 0.124 |
| 0.200 | 0.248 |
| 0.300 | 0.359 |
| 0.400 | 0.488 |
| 0.500 | 0.604 |

Linear Regression - How It Works

Suppose that you measure a response, such as absorbance, for several different levels of a factor, such as the concentration of Cu^{2+} . If the data are linear, then you should be able to model the data using the equation

$$\text{absorbance} = \text{slope} * [\text{Cu}^{2+}] + \text{intercept}$$

If you assume that errors affecting the concentrations of Cu^{2+} are insignificant, then any difference between an experimental data point and the model is due to an error in measuring the absorbance. For each data point, the difference between the experimental absorbance, A_{expt} , and the predicted absorbance, A_{pred} , is a residual error, RE.

$$\text{RE} = (A_{\text{expt}} - A_{\text{pred}})$$

Because these residual errors can be positive or negative, the individual values are first squared and then summed to give a total residual error, RE_{tot} (note: this is the reason that a linear regression is sometimes called a "least-squares" analysis).

$$\text{RE}_{\text{tot}} = \sum (A_{\text{expt}} - A_{\text{pred}})^2$$

Different values for the slope and intercept lead to different total residual errors. The best values for the slope and intercept, therefore, are those that lead to the smallest total residual error.

This [applet](#) provides an excellent visualization of how the slope and intercept affect the total residual error. Give it a try and see if you can achieve a total residual error that is lower than my best effort of 843.

When you are done, use the link on the left to proceed to Problem 1.

Linear Regression - Problem 1

Open Data Set 1, using the link on the left, which is an Excel file containing 11 paired X and Y values.

Task 1. Calculate the mean and standard deviation for the 11 X values and for the 11 Y values. Briefly explain the meaning of these values.

Task 2. Prepare a scatterplot of the data and model the data by adding a linear trendline (a linear regression); be sure to display both the equation and the R-squared value on your scatterplot. How well does this model appear to explain the data?

After completing these tasks, proceed to Problem 2.

Linear Regression - Problem 2

A linear model of the data in Problem 1 gives the following results:

$$Y = 0.500 * X + 3.00$$

$$R^2 = 0.6665$$

$$R = 0.8164$$

The coefficient of determination (R^2) tells us that about 2/3 of the variation in the values of Y can be explained by assuming that there is a linear relationship between X and Y. There is no compelling visual evidence that another mathematical model is more appropriate, so the model seems reasonable. The significant scatter in the points, however, suggests that there is substantial uncertainty in X, in Y or in both X and Y.

Let's consider some additional data sets.

Task 1. The table below contains X and Y values for the data from the previous problem (Data Set 1), and for two additional data sets. All three data sets have identical values for X but have different values for Y. Examine the data sets without using any mathematical approaches. Do the data sets appear to show a similar relationship between X and Y? Explain.

| Data Set 1 | | Data Set 2 | | Data Set 3 | |
|------------|-------|------------|------|------------|-------|
| X | Y | X | Y | X | Y |
| 10.00 | 8.04 | 10.00 | 9.14 | 10.00 | 7.46 |
| 8.00 | 6.95 | 8.00 | 8.14 | 8.00 | 6.77 |
| 13.00 | 7.58 | 13.00 | 8.74 | 13.00 | 12.74 |
| 9.00 | 8.81 | 9.00 | 8.77 | 9.00 | 7.11 |
| 11.00 | 8.33 | 11.00 | 9.26 | 11.00 | 7.81 |
| 14.00 | 9.96 | 14.00 | 8.10 | 14.00 | 8.84 |
| 6.00 | 7.24 | 6.00 | 6.13 | 6.00 | 6.08 |
| 4.00 | 4.26 | 4.00 | 3.10 | 4.00 | 5.39 |
| 12.00 | 10.84 | 12.00 | 9.13 | 12.00 | 8.15 |
| 7.00 | 4.82 | 7.00 | 7.26 | 7.00 | 6.42 |
| 5.00 | 5.68 | 5.00 | 4.74 | 5.00 | 5.73 |

Task 2. Data Sets 2 and 3 have statistical characteristics that are identical to those for Data Set 1. All three data sets have the same means for X and Y (9.00 and 7.50, respectively), the same standard deviations for X and Y (3.32 and 2.03, respectively), the

same linear models ($Y = 0.500 \cdot X + 3.00$), and nearly identical values for R^2 (0.6665, 0.6662 and 0.6663). Given this additional information, reconsider your answer to the question in Task 1: Do you expect that all three data sets are described equally well by the same linear model? Explain.

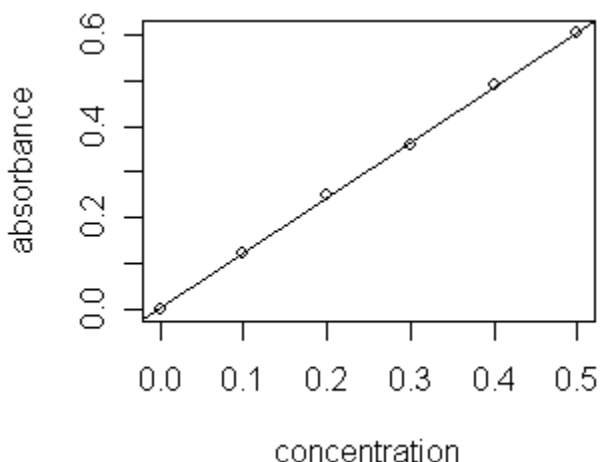
After completing this task, proceed to Problem 3.

Linear Regression - Problem 3

The identical statistical results for Data Sets 1, 2 and 3 certainly suggests that all three data sets may be equally well explained by the same linear model. Let's look at this more closely. Using the link on the left, open the Excel file containing the three data sets. Each data set is on a separate Excel worksheet.

Task 1. For each data set, create a separate scatterplot with a linear trendline. How well does this linear model explain the relationship between X and Y for each data set? What does this imply about the usefulness of using R^2 or R as the sole measure of a model's appropriateness?

Task 2. For all three data sets, the value of R^2 is relatively small. The linear model of data from the module's introduction



on the other hand, has an R^2 of 0.9997. Look carefully at Data Set 2. Is there a mathematical model that might better explain the relationship between X and Y? Remove the linear trendline and try a more appropriate model. You should be able to find a model with an R^2 value of nearly 1.

Look carefully at Data Set 3. It appears that the data are linear, so what is the reason for the relatively small value for R^2 ? Edit the data to remove the problem and examine how this changes your model (if necessary, replot the data). You should be able to find a linear model with an R^2 value of nearly 1.

After completing these tasks, proceed to the module's summary.

Linear Regression - Summary

A regression analysis provides us with the ability to mathematically model the data we collect in lab. In turn, this allows us to make predictions about the results of additional experiments. Blindly accepting the results of such an analysis without carefully examining the data and the model, however, can lead to serious errors.

By now, you know that Data Set 1 can be explained adequately using the model

$$Y = 0.500 * X + 3.00$$

although there appears to be substantial uncertainty in the values for X, for Y, or for both X and Y. The data in Data Set 2 are nonlinear and adequately modeled using the following 2nd-order polynomial equation

$$Y = -0.1276 * X^2 + 2.7808 * X - 5.9957$$

with an R² of 1. With the exception of one data point, which appears to have an unknown source of determinate error, the data in Data Set 3 are linear. Removing this data point (X = 13.00, Y = 12.74) and fitting a linear trendline gives the following model equation

$$Y = 0.3454 * X + 4.0056$$

with an R² of 1.

Additional information on the topics covered in this module is available using the link on the left for further study.

Linear Regression - Further Study

The data sets in this module were created by F. J. Anscombe and were first published in the article "Graphs in Statistical Analysis," *American Statistician*, Vol. 27, pp 17-21 (1973). An alternative data set created by J. M.. Bremmer can be found [here](#).

Mathematical Details. Our treatment of linear regression has been quite limited. In particular, the mathematical details have been omitted. For further information, including a discussion of more esoteric (but important) considerations, such as the standard deviation of the regression, the confidence intervals for the slope and intercept, and the use of a weighted linear regression, consult the following textbook:

Miller, J. C.; Miller, J. N. *Statistics for Analytical Chemistry*, Ellis Horwood: Chichester

A handout showing the derivation of the equations for a linear regression is available [here](#).

Effect of Outliers. This [applet](#) demonstrates how an outlier affects the results of linear regression. The original data consists of a five-point linear model. Click to add a sixth point and observe its effect on the regression line. Try placing the sixth point at various places both near and far from the line, as well at low, middle and high values for X.

Residuals. As you have seen, it is not a good idea to rely on the value of R^2 or R as the sole measure of a model's appropriateness. An additional tool for evaluating a regression model is to examine a plot of the residual error in Y as a function of X. If a model is appropriate, then the residual errors, should be randomly scattered around a value of zero. To calculate the residual error for each value of X, use your regression line to calculate the predicted Y; that is

$$Y_{\text{pred}} = \text{slope} * X + \text{intercept}$$

Next, calculate the residual errors

$$RE = (Y_{\text{expt}} - Y_{\text{pred}})$$

Finally, plot the residual errors vs. the values of X and examine the plot. [Here](#) are the three data sets that you worked with in this module. Create residual plots for each and see if the results agree with your earlier determination about the validity of fitting a straight-line to the data. For Data Set 2, try both the straight-line model and a 2nd-order polynomial model. For Data Set 3, try the straight-line model with and without the apparent outlier.

Outliers - Introduction

When we collect and analyze several replicate portions of a material we do so with the intent of characterizing that material in some way. Suppose, for example, that we gather seven replicate sediment samples from a local stream and bring them back to the lab with the intent of determining the concentration of Pb in the sediment. After analyzing each replicate, we obtain the following results (in ppb)

4.5, 4.9, 5.6, 4.2, 6.2, 5.2, 9.9

and report the average (5.8 ppb) as an estimate of the amount of Pb in the sediment and the standard deviation (1.9 ppb) as an estimate of the uncertainty in that result.

The mean and standard deviation given above provide a reasonable summary of the sediment **if** the seven replicates come from a single population. The last value of 9.9 ppb, however, appears unusually large in comparison to the other results. If this sample is not from the same population as the other six samples, then it should not be included in our final summary. Such data points are called outliers.

Discarding the last result changes the mean from 5.8 ppb to 5.1 ppb and the standard deviation from 1.9 ppb to 0.73 ppb, which are not insignificant changes. Is discarding the suspect data point justifiable? Is it ever justifiable to discard a data point? If so, what criteria should influence a decision to reject an apparent outlier?

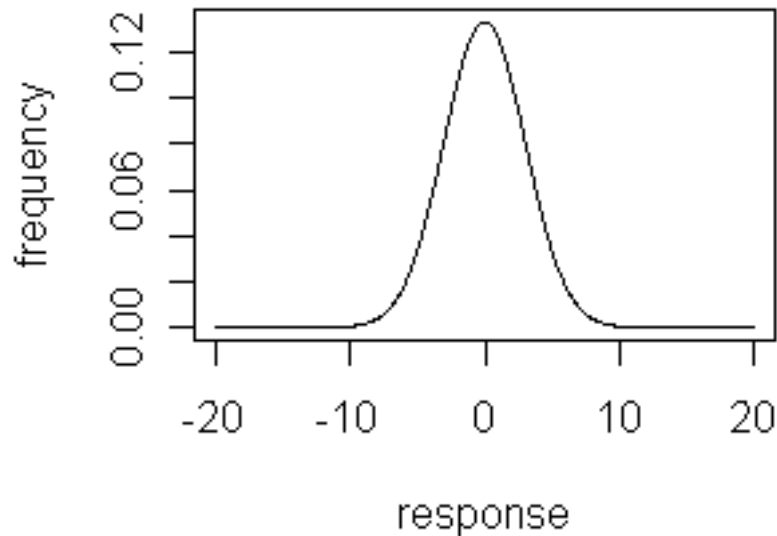
When you complete this module you should:

- appreciate how the relative position of one data point from the remaining data affects your ability to determine if it is an outlier
- be able to use the Q-test to identify and reject a possible outlier
- understand how paying attention to your data as it is collected can help you in identifying possible outliers

Before tackling some problems, read an explanation of how the Q-test works by following the link on the left.

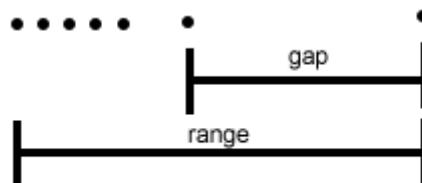
Outliers - The Q-Test

In earlier modules we introduced the idea of a normal distribution (the so-called "bell-shaped" curve)



whose shape is symmetrical, and is centered on the population's mean with a width that is proportional to the population's standard deviation. We also learned that the probability of obtaining a particular response for any single sample is greatest when its value is close to the mean and smallest when its value is far from the mean. For example, there is a probability of only 0.15% that a single sample will have a value that is larger than the mean plus three standard deviations. It is unlikely, therefore, that the result for any replicate will be too far removed from the results of other replicates.

There are several ways for determining the probability that a result is an outlier. One of the most common approaches is called Dixon's Q-test. The basis of the Q-test is to compare the difference between the suspected outlier's value and the value of the result nearest to it (the gap) to the difference between the suspected outlier's value and the value of the result furthest from it the range).



The value Q is defined as the ratio of the gap to the range

$$Q = \frac{\text{gap}}{\text{range}}$$

where both "gap" and "range" are positive. The larger the value of Q , the more likely that the suspected outlier does not belong to the same population as the other data points. The value of Q is compared to a critical value, Q_{crit} , at one of three common confidence levels: 90%, 95%, and 99%. For example, if we choose to work at the 95% confidence level and find that

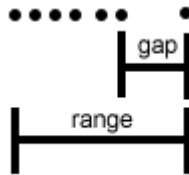
$$Q \leq Q_{\text{crit}}$$

then we cannot be 95% confident that the suspected outlier comes from a different population and we retain the data point. On the other hand, if

$$Q > Q_{\text{crit}}$$

then we can be at least 95% confident that the suspected outlier comes from a different population and can discard it. Of course, there also is as much as a 5% chance that this conclusion is incorrect.

Data that are more tightly clustered, as in this example,



are less likely to yield a value for Q that favors the identification of an outlier.

When you are ready, proceed to Problem 1 using the link on the left.

Outliers - Problem 1



(Applet courtesy of Prof. C. E. Efstathiou, http://www.chem.uoa.gr/applets/Applet_Index2.htm)

In this problem you will use an applet to explore how the relative position of one data point relative to the remaining data points affects your ability to classify it as an outlier.

Task 1. Begin by placing two points about 1 cm apart somewhere in the middle of the one of the lines and placing one point (the possible "outlier") near the right end of the same line. Click on the CALC button and wait for the results. In the box to the right you will find the calculated value for Q and the probability (P) that rejecting the data point is an error. A P-value of 0.22, for example, means that there is a 22% probability that the "outlier" might come from the same population as the remaining data. At the bottom of the applet is an indication of whether the "outlier" can be rejected at the 90%, 95% and 99% confidence levels. Is the "outlier" really an outlier? What happens if you add an additional point to the middle of the line? How about if you add another data point in the middle of the line?

Consider how the results of this task might affect how you perform an experiment? A common choice in lab, where time is limited by the length of the lab period, is to repeat an analysis three times. What would you do if two of your results are in good agreement, but the other result is much larger or much smaller? [Hint](#).

Task 2. Starting with a new line in the applet, place three data points in the middle of the line, click on the CALC button and note the results of the Q-test. Next, add a point on the far right hand side of the line. Click on the CALC button again. Is this data point an outlier? Finally, add a point on the far left hand side and click on the CALC button. How does this new data point affect your ability to reject a data point? What can you conclude about the importance of precision in rejecting a data point?

Task 3. Use the CLEAR ALL button to remove any data points already entered into the applet. Add three data points on the left hand side of each line in the applet; try to arrange these so that their positions on each line are similar. On each line add one additional data point, arranged such that on each successive line the new data point is further from the first three data points. Beginning with the first data set, use the CALC button to obtain the results of the Q-test. What can you conclude about how the distance of an outlier from the remaining data affects a Q-test? Does this make sense?

When you finish these tasks, move on to problem 2.

Hint

If your answer is to throw out the data point, rethink your answer!

Outliers - Problem 2

Open Data Set 1, using the link on the left, which is an Excel file containing the mass and the year of minting for 47 pennies.

Task 1. Plot this data in Excel and consider whether there are any possible outliers. Why did you select the results that you did and for what years are they? Propose several reasons for why the mass of an individual penny might be an outlier?

Task 2. Use this [link](#) to learn more about the history and production of the U. S. penny. Does this information allow you to reach a conclusion about any of the possible outliers? Does it explain all possible outliers? If not, is there any additional information that might help you evaluate any of the remaining possible outliers?

After completing these tasks, proceed to Problem 3.

Outliers - Problem 3

In Problem 2 you should have identified two possible outliers. You can reject one of the possible outliers, the penny from 1943, without a Q-test because you know, from external evidence, that its composition is different from that of the other pennies. The 1943 penny must, therefore, come from a different population and should not be included with the remaining pennies. This is an important point. If you know that there is a significant error affecting one data point that does not affect other data points, then you should eliminate that data point without regard to whether its value is similar to or different from the remaining data. Suppose, for example, that you are titrating several replicate samples and using the color change of an indicator to signal the endpoint. The indicator is blue before the endpoint, is green at the endpoint and is yellow after the endpoint. You should, in this case, immediately discard the result of any titration in which you overshoot the endpoint by titrating to the indicator's yellow color.

The other apparent outlier in this data set, one of the three pennies from 1950, has no simple explanation. Examining the penny might show, for example, that it is substantially corroded or that it is coated with some sort of deposit. In this case, we would again have good reason to reject the result. But what if the 1950 penny does not appear different from any other penny? Without a clear reason to reject the penny as an outlier, we must retain it in our data set.

In a situation such as this, the Q-test provides a method for evaluating the suspected outlier. For small data sets consisting of 3 to 10 data points, the critical values of Q are given here:

| | Confidence Level | | |
|---------|------------------|-------|-------|
| Samples | 90% | 95% | 99% |
| 3 | 0.941 | 0.970 | 0.994 |
| 4 | 0.765 | 0.829 | 0.926 |
| 5 | 0.642 | 0.710 | 0.821 |
| 6 | 0.560 | 0.625 | 0.740 |
| 7 | 0.507 | 0.568 | 0.680 |
| 8 | 0.468 | 0.526 | 0.634 |
| 9 | 0.437 | 0.493 | 0.598 |
| 10 | 0.412 | 0.466 | 0.568 |

Task 1. In the introduction we examined results for the concentration of Pb in seven samples of a sediment. Given the following data (in ppb):

4.5, 4.9, 5.6, 4.2, 6.2, 5.2, 9.9

is there any evidence at the 95% confidence level that the result of 9.9 ppb is an outlier?
Is your answer different at the 90% or 99% confidence levels?

After completing this task, proceed to the module's summary using the link on the left.

Outliers - Summary

After completing this module you should understand what it means for a result to be a possible outlier and understand how to consider whether it is possible to reject the apparent outlier.

If you suspect that a result is an outlier you should first look carefully at the sample and your work. Is there convincing evidence that the sample, or your analysis of the sample, is fundamentally different from other samples? A penny coated with a green discoloration due to oxidative corrosion should be discarded if it being compare to pennies that are free from such corrosion. A sediment sample containing relatively large solid particles, such as pebbles, should not be included with samples that consist of only fine grain particles. If you overshoot the endpoint during a titration, you should discard the result. Do not retain samples or results in these, or similar, circumstances just because the result happens to make your overall results look better.

If there is no convincing visible evidence for discarding an outlier, consider using the Q-test. Be cautious in its use, however, since you may be eliminating a valid result.

Additional information on the topics covered in this module is available using the link on the left for further study.

Outliers - Further Study

Larger Data Sets. One limitation to our treatment of the outliers is that the Q-table is limited to data sets consisting of 3 to 10 samples. Q-tables are available for larger sample sizes, although, interestingly, the test as defined here is less reliable for samples larger than 10. There are alternative equations, which use different definitions for the gap and range. For details, consult the following paper:

Rorabacher, D. B. *Anal. Chem.* **1991**, *63*, 139-146.

Multiple Outliers. Suppose your data set has an apparent outlier on each end or two apparent outliers on one end. For small data sets, the Q-test as defined will be unable to detect an outlier. For larger data sets, those containing more than 11 samples, there are alternative forms of the Q-test that provides some discriminating ability. For details, consult the following paper:

Rorabacher, D. B. *Anal. Chem.* **1991**, *63*, 139-146.