## Using Microsoft Excel

Objective: Students will gain familiarity with using Excel to record data, display data properly, use built-in formulae to do calculations, and plot and fit data with linear functions. They will also understand the importance of the statistical concepts of the average, the standard deviation, and the idea of outliers.

Data analysis in the modern world relies upon computers and program suites to make the process faster and less prone to mistakes. Similarly, this is a much neater way to produce graphs and plots. Additionally, there are built-in functions which can provide statistical analyses that can give us estimates of the sizes of our error ranges. Whenever possible in this lab, we will use Microsoft Excel to help us process data.

## Important!!

The laptops in class will not save any new work permanently to the hard drive. If the machine shuts off, all of your work will be lost. Make sure your laptop is plugged in to prevent the battery from running out and causing you to lose your work. Be sure to save your work to your own USB drive, save it to a network drive (like Google Drive), or save it and e-mail it to yourself. As long as the computer remains on, your work will be available, and can be saved.

Procedure:
Download the Excel spreadsheet named "Excel_practice_lab" - ask your instructor from where you should get it. Open the spreadsheet, and click to enable editing a file that came from the Internet. Column A lists a reference to each individual run of an experiment, from 1 through 10 , for several different experiments. Column B gives you the values measured for that run of the experiment.

You may wonder why we run the experiment so many different times. Why not just five times? Or three times? One of the cornerstones of the scientific method is that results are repeatable. How do you know if your values repeat if you perform too few trials of the experiment? This also brings up another element of scientific measurements - the idea of an outlier. Repeated measurements, when an experiment is well-designed and executed carefully, should produce results which give about the same value. However, in some cases, you might get a value that is far away from what you expect from the other trials or from the theory. We need to decide if this result is showing us something we did not understand previously, or if it is an outlier. The best way to determine this is to run the experiment more times and do some statistics with the results. There are numerous ways to evaluate a set of data statistically. For this course, we will focus on taking the average of a set of data, and its standard deviation.

Look at column B in the "Excel_practice_lab" spreadsheet for the trials listed for each experiment. You can most likely get a feel for the average from looking through the list. It has to be higher than the lowest value, but also less than the highest value. The first way to consider it, before doing any calculations, is that it has to be "somewhere in the middle" of the data set. You probably already know how to calculate it, but to review:

- Add up all the values
- Then divide the sum by the number of values that you have.

Look at the temperatures as an example. If you wanted to do this by hand, you would add:

$$
95^{\circ} \mathrm{C}+103^{\circ} \mathrm{C}+98^{\circ} \mathrm{C}+\ldots+97^{\circ} \mathrm{C}+99^{\circ} \mathrm{C}
$$

We should introduce some more general mathematical notation for doing this, so that we can use an equation to represent this sum in the future. First, instead of using the values, we will represent each trial with a variable: $\mathrm{T}_{1}$ for the first trial, $\mathrm{T}_{2}$ for the second, up to $\mathrm{T}_{10}$ :

$$
T_{1}+T_{2}+\cdots+T_{10}
$$

where $T_{1}=95^{\circ} \mathrm{C}, T_{2}=103^{\circ} \mathrm{C}$, and so on, up to $T_{10}=99^{\circ} \mathrm{C}$. This is better, especially for a computer program, where you want different values assigned to each variable for each time you run the experiment for several trials. You just have as many $T_{i}$ as you need. What does the $i$ stand for that goes with our temperatures? It is an index. You see the trials are labeled from 1 to 10 -those are the indices to let us know which result comes from which trial. We want to sum the temperatures over their indices. The mathematical notation for this is to use a Greek capital letter 'S' to stand for sum. It looks like this:

$$
\Sigma
$$

and it is pronounced "sigma". The mathematical notation for adding up all the temperatures is

$$
\sum T_{i} .
$$

We can add to this notation to let someone else know we are adding from trial 1 up to trial 10:

$$
\sum_{i=1}^{10} T_{i}
$$

If there is not enough room to write it this way, you may also see it as $\sum_{i=1}^{10} T_{i}$.

We do not have the average yet, though. We have to divide by the number of trials, or multiply by the reciprocal of the number of trials:

$$
\frac{1}{10} \sum_{i=1}^{10} T_{i}
$$

To write this as generally as possible, we do the following:

$$
\frac{1}{n} \sum_{i=1}^{n} T_{i}
$$

In our case, $\mathrm{n}=10$. We add together the values from all trials, 1 through 10 , then divide by 10 . If you run the experiment 26 times, then $\mathrm{n}=26$. In some cases, like at the Large Hadron Collider, they are looking for millions or tens of millions of runs of an experiment (or more). The
mathematical notation for the average value is a variable with a line above it. For our average temperature, it would be $\bar{T}$. Any time you see a variable with a bar over it in this course, you should think "average". The other name for the average value is the mean value. So, average and mean are interchangeable for our purposes.

To put this all together in one equation:

$$
\bar{T}=\frac{1}{n} \sum_{i=1}^{n} T_{i}
$$

We can use the same notation for average speed $(\bar{v})$, average mass $(\bar{m})$, or even average florbs ( $\overline{\text { florbs }}$ ).

How do you take an average in Excel? Excel has a built-in function for doing this. Go to an empty cell on the spreadsheet and type "=AVERAGE(" then, highlight all of the values you want to average, then close the parenthesis ")". When you hit the "Enter" key, the average value will appear in the cell. There is actually an even easier way to do this: highlight all the values and look in the lower right corner of the window for Excel.


The status area at the bottom displays the average of the values, how many values there are (the "Count"), and the sum of all the values.

Once we have an average value, how do we determine whether a value is an outlier? We have to evaluate how the data is spread out. Does it cover a narrow range of values, or do they go much higher and lower than average? What is most often computed to determine this is the standard deviation. This lets another person reading your report know, in one number, a rough idea of how your values were spread out, or distributed.

Standard deviations are computed for a set of data by doing the following:

- Find the average (or mean) value of the data
- Subtract the average from the value for each trial
- Square all the differences you got from the previous step
- Sum the squares of the differences
- Divide the sum by one less than the number of trials
- Take the square root of the result

Technically, this is known as the sample standard deviation - you could take an entire course on statistics if you want to know more about this. For our purposes, we'll call this our "spread".

Here are the steps for doing this the long way in Excel. Do this for the temperature values in the spreadsheet.

- First, find the average value using one the methods mentioned above - use either the "AVERAGE" function or highlight all the cells.
- In column C on your spreadsheet, do the following:
- Click on the cell to the right of the first trial
- Type an equals sign =
- Click on the value in column B; a reference to that cell should appear next to the equals sign. For example, it should now appear as $=\mathrm{B} 6$
- Type a minus sign - followed by the value for your average. This should now appear as $=$ B6-98.2 (or whatever you got for your average if it was not $98.2^{\circ} \mathrm{C}$ )
- Press "Enter" and the difference will appear
- Instead of repeating this for all the other cells, try this instead:
- Position the cursor over the lower right corner of the cell where you computed the difference. There is a small black square there; put the cursor directly over that. The square is circled in red in the image at the right.
- The cursor should change from an open cross to a solid black cross when it is over this square.
- When the cursor is a solid black cross, click and drag over the cells where you want to
 computation to be done. The same computation will be automatically repeated over all of the other cells.
- Now that we have the differences, use column D to compute the squares:
- Click on the cell to the right of the first trial in column C and type an equals sign =
- Click on the value in column C, then type an asterisk * for multiplication
- Click on the same value in column C again; the result should look something like $=\mathrm{C} 6 *$ C6
- Hit "Enter" to get the value squared
- Use the same trick with clicking and dragging on the little black square in the lower right to repeat the calculation for the other cells
- Find the sum of all values in column D
- Do not divide by the number of trials; divide by one less than that number (in this case, 9 instead of 10)
- Take the square root of the result; you can do this in Excel with =SQRT(47.2) (if 47.2 was your result from the last step)

What is produced from the last calculation is what we will use as the standard deviation, denoted by $\sigma$ (lowercase sigma). The general formula for the standard deviation is:

$$
\sigma=\sqrt{\frac{1}{n-1} \sum_{i=1}^{n}\left(x_{i}-\bar{x}\right)^{2}}
$$

This notation may vary slightly from what you learned in a statistics class, but the underlying concept is the same.

## Important Excel note:

If you change the fomula for one row of a computation, you need to drag that box (using the small black square) over all other computations for it to take effect. Excel does not automatically change the other computations.

Follow along with these steps as we do this for the temperature data, then do these steps on your own with the speed of light data. This will get you more accustomed to using Excel for calculations.

Once these calculations are in your spreadsheet, compare them with the result from the shortcut:

- In an empty cell, type =STDEV(
- Then, drag over the cells that contain the values for the trials
- Close the parentheses ) and hit "Enter"

Did you get the same result as you did when you did it the long way? From now on, use the built-in function, but keep in mind what that function is doing with your data.

- Questions to answer in your lab write-up:
- Why do we square the differences after we subtract the average from each trial? What purpose does this serve?
- List another example from mathematics where we take the square root of the sum of quantities that are squared.

How should you report these average values with standard deviations in your lab reports? There are two primary ways you may see this done. The first is to use a plus/minus sign, which is what your text uses in the homework from Chapter 1:

The accepted value for the radius of a proton is $0.879 \pm 0.008 \mathrm{fm}$.
The second way is to include the standard deviation in parentheses after the last significant digit:
The accepted value for the radius of a proton is $0.879(8) \mathrm{fm}$.

The method with the plus/minus ( $\pm$ ) sign seems much clearer to me, especially for people who are just learning physics. This is the version you should use. To write this in general for the radius of a proton, we would use:

$$
\overline{r_{p}} \pm \sigma_{r_{p}}
$$

The $r_{p}$ subscript on the $\sigma$ reminds us that this is a standard deviation of a measurement of the radius of a proton. (Yes, even subscripts can have subscripts!)

Examine the data for the mass of a mole of bismuth (chemical symbol, Bi ) atoms. A mole of something is $602,214,082,000,000,000,000,000$ of whatever thing you are counting. This is a bad standard to use for things like books or puppies, but good for things like atoms or molecules. All of those extra zeroes on that number (after the last "2")? Those are not significant. A much better way of writing the most precise value measured ${ }^{3}$ for a mole is $6.02214082 \times 10^{23}$.

- Examine the 10 trials for the mass of bismuth. Can you identify an outlier from looking at the data? Which trial is it? Compute the average mass, $\bar{m}$, and the standard deviation, $\sigma_{m}$, if you include the outlier. Record this value to include in your report.
- Change the ranges on your AVERAGE and STDEV so they do not include the outlier. You can do this by changing =AVERAGE(B78:B87) (or whatever the range is) to $=A V E R A G E(B 78: B 79, B 81: B 87)$ where the example skips the value in $B 80$ (if that is the outlier). Record the new values for $\bar{m}$ and $\sigma_{m}$.


## Be very careful with outliers:

Never assume that because one run of an experiment produces a value that is different from the others that it is automatically an outlier. The best thing to do is run the experiment for additional trials. If the procedure calls for 5 separate measurements, and you suspect one is an outlier, run 6 or 7 (at least) so that you can be sure. Another approach is, once you have run the extra trials, compute the average and standard deviation ( $\sigma$ ) without the outlier. If the outlier is more than three standard deviations away from the average, you may be safe in omitting it.

In the case of the three data sets we have worked with, there is an accepted value for these quantities. It should only take a little bit of searching to find these accepted values. Since you will have access to a computer during lab, you should get accustomed to using it to look up information from reliable sources. We can compute a percent error for each of the values as follows:

$$
\frac{\mid \text { accepted value }- \text { experimental value } \mid}{\text { accepted value }} \times 100 \%
$$

[^0]You may also see "accepted value" listed as "theoretical value", if we have a valid theory (or hypothesis) for determining this value. If there is an accepted or theoretical value for a quantity that you measure in this lab, you should include the percent error.

## Something students often get wrong in this course:

When you subtract two values that have a limited number of significant figures, the result often has fewer significant figures than when you start. We will discuss more about this in the next lab, but you should already have done some work with this in class. This problem arises most often when students are determining percent errors. Pay close attention to significant figures whenever you subtract two quantities.

The last exercise we will do as part of this lab is to create a plot. Look at the data that represents the altitude of an object falling toward the surface of Mars. First, let's plot the altitude (in meters) vs. the time (in seconds).

1. Click "Insert" at the top, then click the symbol for "Scatter" under "Charts."
2. Choose the option that shows "Scatter with only markers" - it's the top left option.
3. You now need to select the data to use. "Chart Tools" should now be highlighted in green. Click the button from the row at the top that says "Select Data."
4. From the window that appears, click "Remove" on the left side of the window until the area under it is empty.
5. Click the "Add" button.
6. Enter "Altitude vs. time" for the "Series name."
7. Click on the box for "Series $X$ values," then drag the cursor over all of your values for the time from your table.
8. Click on the box for "Series Y values," delete any numbers that are present, then drag the cursor all of your values for altitude from your table.
9. Click OK on this dialog and the one where you originally selected the data. This should complete the data for your graph.

Although this is not the final graph that we want, take a look at what these steps produce.

- What shape would you say the points on the graph make?

What we would prefer to plot is a linear graph - it is easier to determine what relationships are present if the graph is a straight line. To do this, we will plot altitude vs. time squared ( $\mathbf{T}^{2}$ ). In another column next to the altitude measurements, populate the cells with the square of the time data. You should know how to do this from our practice with standard deviations. Run the steps above to plot altitude vs. time squared.

The last thing you will need to produce for this lab is a copy of the plot shown below. It should have a title, labels on the axes with the proper units, a trend line showing the best line fit to the data, and the equation of the trend line. Here are some guidelines for accomplishing this:
10. "Chart Tools" should still be highlighted in green. Click "Layout" from this highlighted section.
11. Click "Axis Titles", highlight "Primary Vertical Axis Title," and choose "Rotated Title."
12. Click in the "Axis Title" area that appears on your graph and add the appropriate variable and include the units.
13. Click "Axis Titles", highlight "Primary Horizontal Axis Title," and choose "Title Below Axis." Add the appropriate variable and units to the area that appears on your graph.
14. Click on the button that says "Trendline" and choose "Linear Trendline."
15. Right click on the line that appears on your chart and click "Format Trendline."
16. Click the box for "Show Equation on Chart."

You can also use "Gridlines" from the "Layout" menu to add those. To get superscripts or subscripts, right click on the letter or numbers you want to change, then select "Font" from the menu that appears. This will give you the option to change something to a superscript or subscript. You will be graded on how closely your plot resembles the following:


## Your report for this lab should include:

- You do not have to write a formal report for this lab. Include a cover sheet, and make sure the following points are addressed.
- Having me sign off on your work in Excel before you leave the lab
- Email your completed Excel file, including your graph, as an attachment to the e-mail address provided by your instructor:
- For all repeated measurements of a single value, such as the boiling point of water or the speed of light, include them in your report in the form:

$$
\bar{x} \pm \sigma_{x}
$$

with proper units, scientific notation, and significant figures. Explicitly state what physical quantity the numerical value represents.

- For measurements which are compared with a well-established quantity, compute the percentage error, and include it.
- State whether the accepted value falls within the error range for each measurement.
- The Excel spreadsheet lists the trials as "accurate but not precise", "precise but not accurate", and "precise and accurate." Explain in your own words why each phrase is associated with its set of data.
- Write the radius of a proton in meters (m) rather than in femtometers (fm) by using scientific notation and correctly including the error in the value.
- In the equation in the corner of the "Altitude above Mars vs. time squared" plot above, what is the significance of the $y$-intercept value? What point in the fall does this refer to?
- Your report should also include using an equation editor to type out the following expressions in the conclusions of your report. This is to get practice with the equation editor.

1) 


2) $v_{f}^{2}-v_{i}^{2}=2 a \Delta x$


[^0]:    ${ }^{3}$ See efforts to redefine the kilogram at www.nist.gov

