## *Data Analysis*

Once you've finished performing calculations with your data, you may need to do a few more calculations to assess your data. First, we'll cover what to do with repeated measurements of the same thing, then how to compare your results to expected values.

### *Averages*

If you make repeated measurements, you will probably want to find a single value that reflects all your measurements. The simplest way to do this is with an average. Suppose you have 3 values of some quantity $X$ ($X_1$, $X_2$, $X_3$).

$$X_{average} = (X_1 + X_2 + X_3)/3$$

In general, to find an average, add all the individual results together, then divide by the number of results you added. It's usually best to average your final calculated results, not your raw data (in example A above, it would be better to average the number of moles $n$, not the temperatures and pressures).

### *Outliers*

When you go to average your results, you might notice that some of your trials are very different from the others. For example, if you measure the mass of 10 mL of water 5 times, and get 10.003 g, 9.998 g, 9.986 g, 6.581 g, and 10.005 g, one of those values is really different from the others. It looks like maybe you spilled a lot of the water in trial 4, making it an outlier. Four of the results are within 0.019 g, while the other is off by 3.5 g. In this case, it makes sense to leave out trial 4, which would significantly change your average. However, although trial 3 is also a little farther off (the other results are within 0.007, and it's 0.012 off from the next nearest), probably it's best to leave this one in. There isn't really enough data to be sure trial 3 is an outlier.

### *Average deviation*

If you do repeated measurements of a single quantity, you can calculate a better uncertainty by comparing the measurements to each other, rather than just using sig figs.

1. Calculate the **average** of the trials
2. Find the **deviations**: calculate the difference between each trial and the average. Use the absolute value (drop any minus signs). Use 2 sig figs for the deviations.
3. Take the **average of the deviations for all the trials** (same way you'd calculate any average). Use 2 sig figs.

Your uncertainty for the measurement is roughly twice the average deviation. Usually you use 1 sig fig for the uncertainty, unless the first digit in your uncertainty is a 1, in which case save 2 sig figs. For example, 0.05, 6, or 12 have appropriate numbers of sig figs for uncertainties.

### *Summary: reporting results of multiple trials*

In your post-labs, you may be asked for the "reported value". This means the answer you measured, to the best of your knowledge. You should report values as the average of your trials, with the uncertainty equal to two times the average deviation. You should round your average so that it includes one uncertain digit. Thus, if the uncertainty is 0.006, you should round to the third decimal place, and your reported value would be something like 0.123 ± 0.006 mol/L. If the

uncertainty is 20, you should round to the tens place, and the reported value would be something like 2340 ± 20 J.

***Example: Reporting results of multiple trials***
*if you measure the density of water 3 times and get 1.002, 0.998 and 0.999 g/ml,*
*Average density = (1.002 + 0.998 + 0.999)/3 = 0.9997 g/ml*
*Deviation = average — individual measurement (but use the absolute value)*
*so the deviations are 0.9997 — 1.002 = 0.0023, 0.9997 — 0.998 = 0.0017, 0.9997 — 0.999 = 0.0007*
*Average Deviation = (0.0023 + 0.0017 + 0.0007)/3 = 0.0016 ~ 0.002*
*(Notice that I save extra significant figures during the calculation, and round the average deviation to 1 significant figure at the end)*

*Average: 0.9997 g/ml*
*Average Deviation: 0.0016*
*Average Deviation x2 (use 1 sig fig): 0.003*
*uncertainty is in the third decimal place, so round 0.9997 to 1.000.*
*Reported value of density: 1.000 ± 0.003 g/ml*

***Comparing your results***
Several comparison measures are commonly used in the lab.

**% Yield**: When you do an experiment to prepare and isolate a sample of a chemical, you usually compare the amount of the chemical you got to what you expected to get using % yield. Ideally, the % yield is close to 100% (meaning you didn't lose much of the stuff).

$$\%\ \text{yield} = \frac{\text{actual amount}}{\text{theoretical amount}} \times 100\ \%$$

**% Error**: When you measure a number and want to compare it to the accepted value, you use % error. In this case, a very small % error is ideal. For the measured value, if you did multiple trials, use the average value.

$$\%\ \text{error} = \frac{\text{measured value} - \text{accepted value}}{\text{accepted value}} \times 100\ \%$$

**% Difference**: You might use this to measure the range of values when you repeat a measurement. Ideally, it would be small.

$$\%\ \text{difference} = \frac{\text{biggest result} - \text{smallest result}}{\text{average result}} \times 100\ \%$$

## *Using Spreadsheets for data analysis and graphing*

It may have occurred to you that calculating averages and average deviations will be a pain if you have more than a few trials. In this case, it's a good idea to use a spreadsheet program to do it. The results will be more reliable, it's faster, and it's useful to know how to do this. The following instructions are for using Google Sheets, which is free and easily does the things you'll want. You can also use Microsoft Excel, Apple Numbers, or other programs which are similar but perhaps slightly less useful (Numbers is terrible for adding fit lines, and Excel is clumsy for histograms).

***Data Analysis: average and average deviation***:

| | A | B | C | D |
|---|---|---|---|---|
| 1 | trial 1 | 1.002 | "=ABS($B$4 - B2)" | |
| 2 | trial 2 | 0.998 | | |
| 3 | trial 3 | 0.999 | | |
| 4 | | "=AVERAGE(B1 : B3)" | "=AVERAGE(C1 : C3)" | |

First, you'll need to enter data. You can do this by hand, or you might be able to import a .csv file provided by your instructor in some cases. Usually you want to put your trials in rows, and the different quantities you're calculating in columns. For example, enter the sample results from the density example above something like this (A1: A3, B1:B3).

Then go to the density column, below the last result. Type "=average(", select the three results and type ")". Hit enter or return. You'll see the average appear.

In the next column over, type "=ABS(". Select the average, type "-", then select the individual trial. Type ")" and hit enter or return. You'll see the first deviation appear.

Then double-click that cell (C1) and insert the $ symbols into the "address" for the average. This tells it not to to change this part. Then select C1:C3, and hit "Meta D" on PC or "⌘+D" on Mac. This "fills down" the formula for the deviation, so the second and third deviations appear. Now you can calculate the average deviation just as you did the average.

***Graphing: using histograms to find outliers***

If you want to decide which values are outliers, a good way to do this is by looking at a histogram. Select the column or cells with your data (it should all be in the same column). Hit the "insert chart" button, or use "Insert … Chart" in the menu. Choose histogram from the type options (only available if you have more than a few data points), select the customize tab, and and adjust the bucket size appropriately (use the preview to decide what's good; probably small enough to separate the data well). Now you should be able to see clearly if any results are really outliers.

***Graphing: XY scatter plot and linear fit***

To make a traditional graph (XY or scatter type) you'll need two columns of paired data points. For example, you might have a column with mass data and a second column with the volume that corresponds to each mass. In this case, the mass data will be on the *x*-axis and the volume data will be on the *y*-axis (volume vs mass). If you want it the other way, switch the order of the columns. Select both columns and click the chart button. In the Charts tab, choose Scatter. Then in the Customize tab, if you want, you can scroll all the way down and choose trendline. If you think your points should all be on a straight line, choose linear, and then choose "use equation" for label. Now the graph will show the best fit line for your data, and the equation of the line. You can see how well you think it fits.

You can also set the title, axis labels, etc. If you're going to print your graph and hand it in, you should definitely include labels on the axes, with units. Probably it will choose sensible ranges for the axes, but you can change these if you want by double-clicking on the axes. Ideally, your data points should fill the whole graph on both axes so you can see them clearly.

***Graphing by hand***

Graphing by hand is not recommended. Hopefully you remember how to do it in a pinch, but in the modern world you should know how to graph with a spreadsheet. This is much more precise and accurate, and it's much faster and easier once you get the hang of it. You'll also get much better best fit lines from a computer.

## *Errors and Their Effects*

### *Types of Errors*

There are three important types of errors that can affect your experiments.

**Blunders**: these are the avoidable mistakes you make, like spilling your sample, misreading an instrument, not following instructions, etc. Since blunders are avoidable, they are never interesting. In real life, if you make a blunder, you'll always repeat the experiment until you do it without blunders. In your post-lab sheets, you should usually not mention blunders.

**Random error**: all measurements have some amount of random error, meaning if you repeat a measurement many times, it might be slightly different each time because of tiny changes that we can't keep track of. For example, if you take the mass of an object five times on a digital balance, you'll probably find that the mass is slightly different in the last decimal place. This is fine and normal: the last decimal place is the uncertain digit, and it will change a little depending on air currents or other small effects. If you repeat the measurement many times, the average of all the measurements should be the true value, assuming that the instruments are calibrated correctly and only random errors are present. Random errors will probably increase if you are careless, but no matter how careful you are there will be some variation in repeated results.

**Bias**: this means errors that always affect the result in the same direction even when the procedure is followed carefully. For example, suppose you are measuring density of a liquid. You measure the volume in a tall graduated container that doesn't fit on the balance, then pour it into another shorter container to find the mass. This could introduce bias, because in each trial you will leave a little bit of the liquid in first container when you pour it. Your mass measurements will always correspond to a slightly smaller amount of the liquid than your volume measurements, meaning that your calculated densities (mass/volume) will always be a little low and your average density will be a little wrong, no matter how many trials you do. In this case, you aren't making a mistake, because you're following the procedure in a sensible way. (Probably the tall container reduces random error in measuring the volume.) However, it would be better to change the procedure to avoid this error, if possible. A bias in your experiment won't necessarily cause your repeated measurements to have a larger uncertainty, but it does make your result inaccurate, so if you don't pay close attention to possible bias you might be badly mislead.

Notice that although bias sometimes refers to a person's preference for a particular result, human bias is not the only source of bias in experiments. In your work, human bias is not likely to be all that important, assuming that you carefully read your instruments.

### *Accuracy and Precision*

In an **accurate** experiment, the true value is within the uncertainty range of the measured result. For example, if the reported result is measured to be $1.89 \pm 0.3$, then the uncertainty range is 1.86 - 1.92. If the true value is between 1.86 and 1.92, the result is accurate. If enough trials are made, random error will not make an experiment inaccurate, because the average will still match the true value.

In a **precise** experiment, repeated measurements are very close together. In other words, the % difference is small. The uncertainty in your reported results is a measure of the precision,