An Introduction to Experimental Uncertainties and Error Analysis

for
Physics 22
Harvey Mudd College

T. W. Lynn

TABLE OF CONTENTS

No Information without Uncertainty Estimation!

LAX is about 59 minutes from Harvey Mudd by car. You can learn this from the driving directions on Google Maps, and it's a useful piece of information if you are checking out possible travel bargains. But what if you already have reservations out of LAX and need to know when to leave campus for the airport? Then you'd better know that the drive can easily take as little as 45 minutes or as much as an hour and a half, depending on factors you can't possibly determine in advance – like an accident in which a truckload of lemons is accidentally scattered across the 105 Freeway. You will wish you had been told (preferably before booking your tickets) that, while it may take you an hour and fifteen minutes in reasonable traffic, the westbound traffic is never reasonable between seven and ten o'clock on a weekday morning.

Similar comments apply to a vast array of numbers we measure, record, and trade back and forth with each other in our everyday affairs. I'll be off the phone in five minutes – or maybe two to eight minutes. The "freshman fifteen" could actually be three or twenty-three. We set the oven to 350^o, knowing that the actual temperature might be only 330^o when the preheat light goes off.

Just about every number in our lives is actually a stand-in for a range of likely values. Another way of putting this is to say that every value comes with an *uncertainty*, or an *error bar*. I'll be off the phone in 5±3 minutes, the oven temperature is 350±25 degrees, and so on. Sometimes a quantity does have zero error bars: I have exactly one brother. More often, though, numbers have error bars and we ignore them only through the ease of familiarity. I know my oven temperature is close enough to bake good cookies, so I don't care to remember (or to know in the first place) just how close it is.

When a situation is unfamiliar, though, suddenly it can be very important to ask about the error bars. Without error bars on the travel time to LAX, you may very well miss your flight. Your dorm room may be about seven feet wide, but lugging home that used couch will seem pretty dumb if the room is actually 7 feet ± 5 inches! And it's not just numerical values that can have uncertainties attached to them. You can afford to play video games tonight instead of studying for a chem exam… maybe. Your roommate thinks the person down the hall finds you attractive, but are you sure enough to act on it?

Original work in science and engineering will take you into unfamiliar situations where you absolutely must know how much confidence to place in a result – either a measured number or a final conclusion. Knowing the uncertainty of a measurement can tell you whether your levee will withstand a Category 5 hurricane with 95% or 45% confidence. You can decide whether a new treatment is effective in curing patients, or whether three test subjects just happened to have mild cases in the first place. You can tell whether your lab has discovered a new law of particle physics or just taken a few fluke readings with unreliable equipment.

Uncertainties – how to think about them, estimate them, minimize them, and talk about them – are an important aspect of Physics 22. We will learn a handful of statistical definitions and methods, but we will concentrate on whether they make sense rather than whether we can justify them rigorously. Our goal is to equip ourselves to talk and think reasonably about the experimental situations we encounter in the lab each week.

What Is an Error Bar?

In a laboratory setting, or in any original, quantitative research, we make our research results meaningful to others by carefully keeping track of all the uncertainties that might have an appreciable effect on the final result which is the object of our work. Of course, when we are doing something for the very first time, we don't *know* beforehand what the result is going to be or what factors are going to affect it most strongly. Keeping track of uncertainties is something that has to be done before, during, and after the actual 'data-taking' phase of a good experiment. In fact, the best experimental science is often accomplished in a surprisingly circular process of designing an experiment, performing it, taking a peek at the data analysis, seeing where the uncertainties are creeping in, redesigning the experiment, trying again, and so forth. But a good rule is to estimate and record the uncertainty, or error bar, for every measurement you write down.

What is an error bar and how can you estimate one? An error bar tells you how closely your measured result should be matched by someone else who sets out to measure the same quantity you did. If you record the length of a rod as 95.0±0.05cm, you are stating that another careful measurement of that rod is likely to give a length between 94.95cm and 95.05cm. The word "likely" is pretty vague, though. A reasonable standard might be to require an error bar large enough to cover a majority – over 50% – of other measurement results. On the other hand, if I am betting you $100 that your result will be within my range, maybe I ought to give a larger error bar so that I'll be covered 99% of the time. Error bars, then, should be larger when it is more crucially important for them to cover all the possibilities.

However, it's convenient to have some sort of standard definition of an error bar so that we can all look at each other's lab notebooks and quickly understand what is written there. One common convention is to use "one sigma" error bars; these are error bars which tell us that 68% of repeat attempts will fall within the stated range. The 68% figure is not chosen to be weird, but because it is easy to calculate and convenient to work with in the very common situation of something called 'Gaussian statistics.' We will not go into this in detail, but here's one example of how useful this error bar convention can be: for many, many situations, if 68% of repeat attempts are within one error bar of the initial result, 95% will be within two error bars.

The essential point here is that your error bars should be large enough to cover a majority, but not necessarily a *vast* majority, of possible outcomes. If someone does an independent measurement of your quantity and finds a value outside your error bars, you can be a little bit surprised. If someone finds a value different from yours by four error bars, you should be deeply disturbed.

Finally, an error bar estimates how confident you are in your own measurement or result. It represents how well you did in your experimental design and execution, not how well the group at the next bench did, or how well your lab manual or professor think you should be able to do. Error bars are part of your data and must follow logically from what *you* did and the observations *you* made; anything else is fraudulent data-taking.

Random Errors, Systematic Errors, and Mistakes

There are three basic categories of experimental issues that students often think of under the heading of experimental error, or uncertainty. These are random errors, systematic errors, and mistakes. In fact, as we will discuss in a minute, mistakes do *not* count as experimental error, so there are in fact only two basic error categories: random and systematic. We can understand them by reconsidering our definition of an error bar from the previous section.

An error bar tells you how closely your measured result should be matched by someone else who sets out to measure the same quantity you did. How is this mysterious second experimenter going to measure the same quantity you did? One way would be to carefully read your notes, obtain your equipment, and repeat your very own procedure as closely as possible. On the other hand, the second experimenter could be independent-minded and could devise an entirely new but sensible procedure for measuring the quantity you measured. Either way, the two results are not likely to be exactly the same!!

A careful repetition of your own procedure will give slightly different results because of ***random error***. There will be slight and uncontrollable differences from one trial to another. Of course, these uncontrollable differences may not be strictly random in their causes. Maybe the air conditioning happens to blow a slight puff of air on your setup the first time. Maybe a speck of lint falls on the second experimenter's ruler and causes them to slightly mis-estimate a string's length. But, however these differences arise, they cause different results when a single procedure is repeated several times. The differences don't trend in any particular direction, and their causes are subtle and hard to identify, let alone control, in the lab – so we call them random.

When a second experimenter designs her own, independent procedure to measure your quantity, the two of you can have differing results because of random error but also because of ***systematic error***. Systematic error arises when your experimental procedure and/or apparatus somehow cause all your measurements to be shifted away from the true value of the quantity you set out to measure. A systematic error happens in the same direction and the same (or similar) size in all your data, so its effect only shows up when an alternate measurement procedure is compared to yours.

An example: Suppose we want to know how tall a bean sprout is twenty days after planting. We plant ten sprouts, care for them all the same way, and then measure their height twenty days later, using a ruler. The sprouts, though planted and cared for identically, are not all measured at exactly the same height. We can calculate an average height, but it's fairly certain that a repeat trial of ten new sprouts won't give exactly the same average. There's *random error* because not all sprouts behave identically.

However, later on we realize that we didn't record (or remember) what time of day we did either the planting or the measuring. We meant to record height after twenty days, but perhaps it was actually 19.6 days or 20.5 days. The heights we measured were all off in the same way – either all too young (too short) or too old (too tall) – but we have no way of knowing which way they were off, or by how much. This is a *systematic error* in our measurement.

The example of the bean sprouts brings us to the ***mistake*** category. Mistakes are things that go wrong in an experiment that can and should be fixed. They don't count as experimental errors, since the experimentalist (you) fixes them before reporting a final result and uncertainty! Some mistakes are easier to fix than others. Suppose we conducted the bean sprout experiment but measured after only nineteen days by mistake. Obviously we'd realize our error that night, and go back and measure properly on the twentieth day. Or suppose we forgot to measure on the twentieth day, and could only get there one day late. This mistake would be harder to fix, but we could do it. We might measure all the heights on the twenty-first day, then measure them again on the twenty-second day. This would give us an idea of how much the sprouts grew per day, and we could estimate the actual twentieth-day heights by subtracting appropriately from the twenty-first day measurements.

Of course, if we measured the sprouts on the wrong day and then put the notebook away in a drawer while we went to Disneyland and the sprouts dried out and died… then the mistakes would take MUCH longer to correct. A cardinal rule of experimentation: the more you think about your results as you go, the easier it will be to correct your mistakes. It's tempting to say that you won't *make* mistakes in the first place, and therefore won't need to rethink midcourse and correct your mistakes… but it simply is not true. Consider the following quote:

> "Fast turnaround time has always been important to me. Mistakes are unavoidable, so I wanted an apparatus that would allow mistakes to be corrected as rapidly as possible."

The quote comes from Steven Chu's 1997 lecture on his acceptance of the Nobel Prize in Physics. Maybe *you* won't make mistakes, but the rest of us do it all the time.

How to Estimate Error Bars in Data

Since we are not going into Gaussian (let alone other) statistics, our definition of an error bar remains loose enough so that we should not be too concerned over the exact numerical value we assign to error bars in our experiments. However, we do want to base our error bars on experimental reality, so they can be useful in clarifying our data analysis and results in the end. The overall uncertainty of a result tells us how much trust to place in the specifics of the result. Beyond that, however, identifying the major source(s) of the final uncertainty can guide us in spending our time and effort productively, should we wish to redesign the experiment for better results in the future.

So: How do we assign an error bar to a measurement taken in the lab? Several specific but common situations are covered below. The zeroth rule of error estimation, though, is that we should always think about the meaning of an error bar… and assign an error bar that makes sense based on that meaning.

One of the simplest sources of uncertainty is the resolution or quoted accuracy of a measuring device. Many lab devices, such as electrical meters and mass balances, have resolutions specified by their manufacturers. These device uncertainties can be read off the device (sometimes on the bottom surface) or in its manual. However, something as simple as a meter stick also has an effective device resolution. If the stick is marked every millimeter, for example, then if an object ends between the 101- and 102-mm marks it is probably unreasonable to expect observers to do any better than choosing which mark is closer. In this way, an object that is truly 101.4mm long would be measured at 101mm, while an 101.8-mm object would be recorded as 102mm long. A reasonable error bar for the device resolution of the meter stick, then, would be +/- 0.5mm. A device-resolution uncertainty can be estimated for just about any measurement device by considering the device's construction and the reliability of a reasonable observer.

Another source of uncertainty, sample variation, becomes important when we measure a phenomenon that just doesn't quite come out the same every time. In the hypothetical bean sprout study discussed above, we conduct the experiment on more than one plant because we suspect there is random variation from one bean sprout to another. Measuring several plants and taking the mean of their heights seems like a natural way to find out something about average bean sprout growth. Just as importantly, though, measuring several plants gives us an idea of how strong the random variation might be – and thus how far off our several-plant average might still be from the "true" mean. If we measure twenty plants and all twenty are the same height to within a millimeter, we can be fairly certain that we know the average bean sprout height to better than a millimeter (barring systematic errors). On the other hand, if we measure two plants and their heights are 21.00cm and 22.00cm, we should be pretty wary of reporting the overall average to be 21.50cm. In the next section we will develop formulas for quantities called the *standard deviation* and *standard error* that can be used to find random uncertainty in a quantity based on repeat sampling like this.

The error estimation techniques we have just discussed apply primarily to random errors. How can we estimate systematic errors? First we must consider possible causes of systematic error, then estimate reasonably – from theoretical knowledge, additional experiments, or prior experience – how much effect these causes might have. If we are measuring the length of a metal rod, the length might reasonably depend on temperature. Perhaps the temperature in the room could be as much as three degrees different from standard 'room temperature' definitions. How much shift could that cause in the rod's length? If we have no experience or reference materials to guide us, we could deliberately cool the rod in a refrigerator, measure the new length, and estimate roughly how much length change occurs per degree. This technique of deliberately exaggerating an effect to estimate its significance is often useful in dealing with systematic errors.

There is one more cardinal rule of error sources: "human error" is *never* a legitimate source of error. That phrase is completely uninformative, and should never be used as an insurance or catch-all in discussing an experiment. Humans cause error, of course, but in specific ways that can be described and quantified.

Sample Mean, Standard Deviation, and Standard Error

In this section we develop formulas to quantify a measurement and its random error, based on taking the measurement repeatedly in what is supposed to be the same way (this is sometimes called sampling). This is probably the most mathematical section of our error analysis discussion, but even here we will give reasons why our formulas are reasonable without actually rigorously deriving them.

Imagine we sample a quantity repeatedly, yielding measurements $(x_1, x_2, ..., x_N)$. While we try to make all the measurements identical, random variation shows up in our list, so to estimate an overall result we quite naturally take the mean:

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}. \qquad \text{(Eq. 1)}$$

Perhaps we have done *N=10* repetitions. If we kept going to *N=20* how would the value of $\bar{x}$ change? What if we kept going even longer? In other words, how much uncertainty is left in our measurement because of our limited sampling of the random variation? To answer this question, it's useful to step back a bit first.

When we want to combine all *N* measurements into a single representative result x_{rep}, it's easy and natural to take the mean: $x_{rep} = \bar{x}$. But why is $\bar{x}$, as defined in Equation 1, really the best candidate for x_{rep}? It would be nice to come up with some measure of deviation which is minimized, sample-wide, by this choice. Perhaps we should be trying to minimize the distance between the individual data points and x_{rep}. That is, maybe we should minimize $\sum_{i=1}^{N} \left| x_i - x_{rep} \right|$. This is a nice thought, but it turns out that $x_{rep} = \bar{x}$ does *not* minimize this particular deviation measure… so this must not be the right deviation measure to think about if we are taking sample means. On the other hand, it turns out that $\sum_{i=1}^{N} \left(x_i - x_{rep} \right)^2$ *is* minimized by taking $x_{rep} = \bar{x}$. To see this, we can differentiate the expression with respect to x_{rep} and set the derivative equal to zero:

$$\frac{d}{dx_{rep}}\left(\sum_{i=1}^{N}(x_i - x_{rep})^2\right) = 0$$

$$\sum_{i=1}^{N} 2(x_i - x_{rep})(-1) = 0$$

$$\sum_{i=1}^{N}(x_i - x_{rep}) = 0$$

$$\left(\sum_{i=1}^{N} x_i\right) - Nx_{rep} = 0$$

$$x_{rep} = \frac{\sum_{i=1}^{N} x_i}{N} = \bar{x}.$$

Indeed, the sample mean is the representative value that minimizes the *sum of the squares of the individual deviations*. So if the sample mean is a good measure of the overall result, something related to this summed-squared deviation should be a good measure of the overall result's uncertainty!

Let's begin by imagining that we take an *(N+1)*th measurement. How far from the previous mean is this single, new measurement likely to be? Well, we can use the summed-squared deviation to help us guess, but probably we should divide the sum by *N* first to turn it into a mean-squared deviation: $\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$. This still isn't a good measure of deviation, since it is still squared – if the measurement is a length in centimeters, for example, this thing is in cm^2 so it can't be a deviation. Therefore we'll take the square root: $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$ is called the *root mean square* deviation, or *r.m.s.* deviation for short, and it's a useful measure of how far from the mean a single measurement is likely to fall. It turns out that, by doing proper statistics, one comes up with a slightly more generous (*i.e.*, larger) estimate of individual deviation from the mean. Thus we define a quantity called the ***standard deviation***:

$$std.deviation = \sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}\ . \qquad \text{(Eq. 2)}$$

The standard deviation is used to estimate how far from the mean a single measurement is likely to fall.

Originally, though, we were trying to answer a different question. We wanted to know how far our calculated mean was likely to be from the true, or ideally-and-infinitely-well-sampled, mean. *This* is the uncertainty of our final (mean) result, and we call it the ***standard error*** or ***standard deviation of the mean***. If we increase the number of samples *N*, the standard deviation defined in Equation 2 will not in general get smaller.

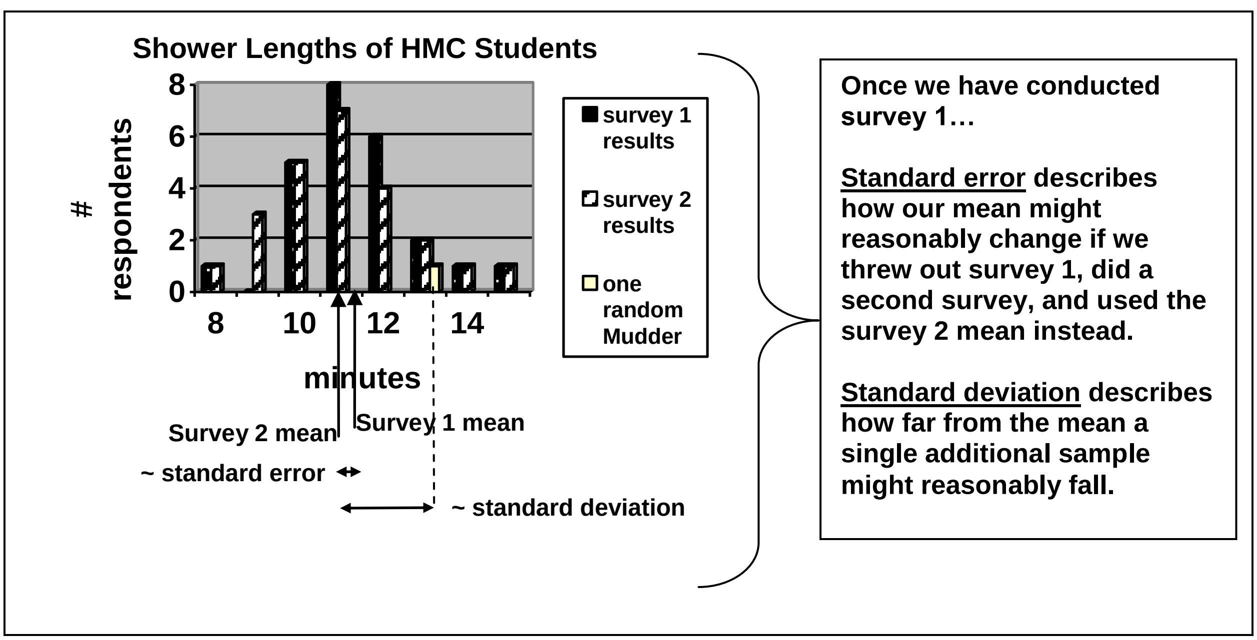
But certainly taking more measurements in our sample ought improve the standard error. Each new measurement we add won't necessarily make $\bar{x}$ closer to the ideal, but in general we'll creep and wander towards the ideal value. Therefore, the standard error is given by:

$$Std.Err. = \frac{\sigma}{\sqrt{N}} = \frac{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}}{\sqrt{N}}. \qquad \text{(Eq. 3)}$$

To sum up this rather lengthy discussion of repeated trials or samples:

In the presence of sample variation, the true value of a quantity can often be calculated by taking the mean of N repeated trials. In that case, the standard error as defined in Equation 3 is a good estimate for the uncertainty of this mean.

CAUTION: Students are often tempted to take a set of repeated trials and summarize them as a mean value plus or minus the standard deviation, rather than plus or minus the standard error. Perhaps this happens because 'standard deviation' is a more familiar term than 'standard error,' and many calculators and software packages have built-in functions for standard deviation but not for standard error. If you find yourself falling into this trap, contemplate the following example which illustrates the rough significance of standard deviation vs. standard error:



How and When to Throw Out Data

Yes, there are times when it is legitimate to throw out a data point. From time to time, one repetition of an experiment gives a result so completely out of line with the other trials that we know there must have been an unidentified problem of unusual size. When our data looks like a tight cluster with one (or two) faraway outliers, it is all right to throw away the outliers just because they are so far from all the rest of the data.

Likewise, when our repetitions produce results that cluster in two different places, we can sometimes think carefully and figure out what we changed or did wrong in half the trials. If we have good reason to think half the data is contaminated and the other half is good, we can throw out the contaminated half.

HOWEVER, it is not legitimate to throw away data points based on a comparison between experimental results and what we *expected* to get, either based on theory or on someone else's reports. This kind of throwing out is habit-forming and very dangerous, since a habit of this sort will prevent us from ever discovering anything new or surprising on our own. It's easy to fall into the trap of fudging data (or inflating error bars) to match an experiment to theory. The best of scientists have done it, as evidenced by another quote from Steve Chu's Nobel Lecture, which is remarkable for its candid discussion of his experiences in atomic physics:

> "Our first measurements showed a temperature of 185μK, slightly lower than the minimum temperature allowed by the theory of Doppler cooling. We then made the cardinal mistake of experimental physics: instead of listening to Nature, we were overly influenced by theoretical expectations. By including a fudge factor to account for the way atoms filled the molasses region, we were able to bring our measurement into accord with our expectations."

Chu's group, which had already demonstrated several milestones of atomic physics, was actually observing an effect now known as sub-Doppler cooling. Bill Phillips and his group measured the same effect but believed in their result, and Chu and Phillips each received 1/3 of the Nobel in 1997. (Theorist Claude Cohen-Tannoudji was awarded the final 1/3 of the Prize that year.)

Habits are hard to break. Make good habits now, and don't adjust your data to match your expectations.

Combining Unrelated Sources of Error

In most experimental situations, if we look hard enough, there are many different sources of uncertainty. The simplest measurement example we have considered so far is that of finding the room-temperature length of a metal rod. If we use a meter stick with millimeter markings, an uncertainty of ±0.5mm is associated with the measurement device. We might estimate an uncertainty of ±0.01mm from unknown temperature variations. An uncertainty of ±0.8mm might be estimated from doing repeated trials – presumably they're different because we have difficulty holding the rod straight against the meter stick each time, or because the rod is slightly longer on one edge than on the other.

What can we report for the overall uncertainty in the length of our simple metal rod? We must somehow come up with a rule for how to combine several uncertainties which are unrelated to each other, but which all influence a single outcome. We could add these uncertainties together, for an overall uncertainty of ±1.31mm, but this is too pessimistic. Adding the uncertainties assumes a kind of worst-case scenario in which the unrelated error sources all end up producing errors in the same direction. More likely, one cause makes the measurement too small, another makes it too large, etc.

We could simply use the largest single uncertainty and neglect all the others, giving us a length uncertainty of ±0.8mm. This, however, is too optimistic. Surely the errors do sometimes combine to make the overall result worse than any one contributing factor.

To combine unrelated error sources, we need a way to add them together without neglecting any of them, and without forcing them to be in the same direction as each other (or opposite each other, either). But wait! In another area of math we are already familiar with adding things that are not in the same direction as each other: We know how to add together mutually perpendicular vectors. If a vector **a** is perpendicular to a vector **b**, then the vector sum **c** has a length given by the Pythagorean theorem:

$$c = \sqrt{a^2 + b^2}\ .$$

And this is how we add together unrelated uncertainties as well. If a measurement has two unrelated sources of uncertainty δ_1 and δ_2, then the overall uncertainty is given by $\delta = \sqrt{\delta_1^2 + \delta_2^2}$. The method extends to deal with more than two unrelated error sources as well.

> If a quantity has *n* unrelated (or *independent*) sources of uncertainty $(\delta_1, \delta_2, ...\delta_n)$, then the overall uncertainty is given by $\delta = \sqrt{\delta_1^2 + \delta_2^2 + ... + \delta_n^2}$. (Eq. 4)

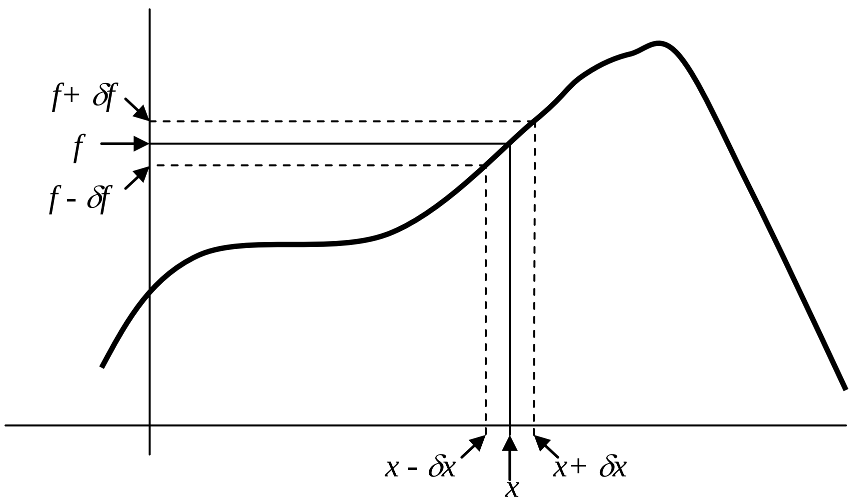
This way of combining independent errors is known as *adding in quadrature*.

Returning to the example of our simple metal rod, we can see that adding our three errors in quadrature gives an overall uncertainty of ±0.9mm, or ±0.94mm if we keep one more decimal place. It's also clear that the temperature-related uncertainty of ±0.01mm is completely unimportant compared to the other two. It is often true that one or two error sources are much more important than all the others, and dominate the overall uncertainty of an experimental result. When this is the case, it's not very important to carry out a heroic error calculation that includes all error sources! Each identified error source should be recorded and estimated, but as soon as it can be clearly labeled as unimportant, it can be dropped from calculations in the interest of time and sanity.

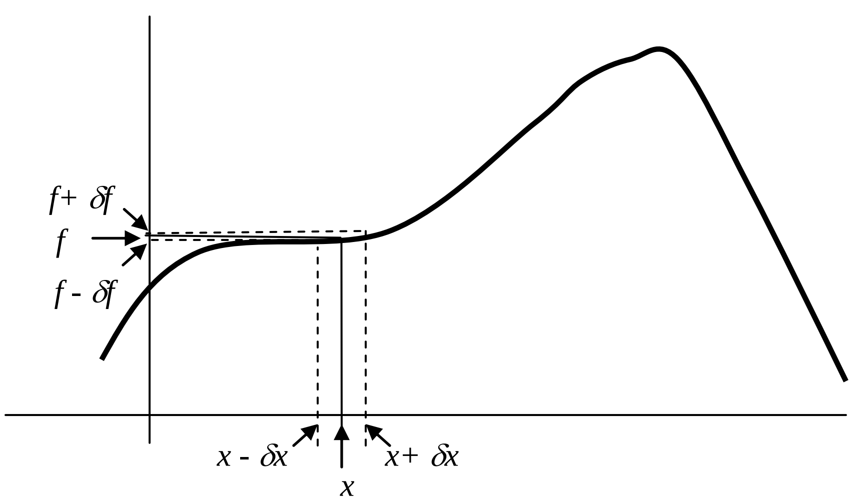
Error Propagation in Calculations: Functions of a Single Measured Quantity

We have discussed methods for finding the uncertainty for a direct measurement. Often, however, we must do some calculations with our raw data to arrive at the result we are actually interested in. The calculation may be as simple as measuring the diameter of a circle and dividing by two to find its radius… but once we use any measured quantity in a calculation, we have to keep track of the uncertainty in our calculated result due to uncertainty in the original measurement. Keeping track in this way is called ***error propagation***. There is really only one basic formula that governs error propagation, and we will develop it right now.

Let's make this problem general by saying we have a quantity, x, which we can measure directly with uncertainty δx. There is a function $f(x)$ we are interested in knowing. Being uncertain about x will clearly cause some uncertainty in f, so we will call this uncertainty δf. We can sketch a graph below that will give us some inspiration about how to calculate δf:



From the sketch above, it is tempting to suggest a simple rule like $\delta f = \delta x$. But we can make a second hypothetical sketch:



In the second sketch, δf is clearly much smaller than in the first sketch, even though δx is exactly the same in both cases. Why is δf small in the second case? Because the function *f(x)* is flat, or very nearly so, near the value of *x* we care about in the second drawing. Aha! The uncertainty in *f* depends on the uncertainty in *x*, but also on the steepness of the function *f* in the spot where we are evaluating it. We can express "the steepness of the function *f* " in more precise and mathematical terms – it is the function's derivative, $\frac{df}{dx}$. Thus we have an error propagation rule for any function of a single variable:

$$\delta f = \delta x \left| \frac{df}{dx} \right|. \qquad \text{(Eq. 5)}$$

The absolute value signs are there because error bars give the size of errors, not their direction, so all error bars are expressed as positive numbers.

Example: if we measure the diameter of a circle as $d = 1.0 \pm 0.1 cm$, the radius is $r = \frac{d}{2} = 0.5 cm$, with an uncertainty of $\delta r = \delta d \left| \frac{1}{2} \right| = 0.05 cm$. However, the area of that circle is $A = \pi d^2 / 4 = 0.79 cm^2$, with an uncertainty of $\delta A = \delta d \left| 2\pi d / 4 \right| = 0.16 cm^2$. Notice that we write the final result as $\delta A = 0.16 cm^2$ rather than $0.1570796... cm^2$, since with an uncertainty in the first decimal place it is a clear waste of space to write down many more. By writing all numbers ± uncertainties we can afford to be lax about significant digits, but we shouldn't offend common sense with long strings of irrelevant numerals.

Describing a Data Set with a Function: Graphing and Chi-Square Fitting

Often several different measurements, with different individual uncertainties, must be combined to answer a final question in an experiment. For example, imagine we are conducting our bean sprout experiment (again), but a colleague has suggested that growth in bean sprouts is linear over time. We now wish to test whether growth is linear or not, and if it is, find a value for the height increase per day. A general plan of action is to take average heights of our sprouts at 10, 20, 30, and 40 days, then make a graph of average height vs. time, see if the data look like a straight line, and find the slope of that line if so.

However, with our new understanding of uncertainties we can make our plan much more specific and quantitative. First of all, we have a data set $\{h_i \pm \delta h_i\}$ of heights measured at times $\{t_i\}$, so any graph we make should show the error bars as well as just the data points. Now, we need some way to choose the *best* possible line $H(t) = mt + b$ to go with our data. That is, we need to find the best possible values of the constants (m, b) based on all four of our data points. In this situation, (m, b) are called ***parameters*** – they are constants in the formula for the line, but they are constants whose values we will adjust to suit ourselves.

Earlier, we developed the concept of standard deviation by considering a data set $\{x_i\}$ of values which were "supposed" to be the same. We wanted to represent the whole data set by a single value, x_{rep} (which for us was $\bar{x}$). Now, we have a data set $\{h_i\}$ in which the individual values are not supposed to be the same, but are supposed to be represented by a single function $H(t_i)$. In our earlier discussion, the mean value $\bar{x}$ was found to minimize the standard deviation σ. In a similar spirit, what quantity do we want to minimize by our choice of the function H – that is, our choice of (m, b) – now? Perhaps we should choose (m, b) to minimize a summed-squared deviation, like $\sum_{i=1}^{N} (h_i - H(t_i))^2$. This is a lot like what we did before, so it seems like a good starting place. In fact, many fit routines (including the "trendline" in Microsoft Excel) work on the principle of minimizing exactly this quantity. Such routines are called *least squares* fitters. However, least squares fitting is not actually a fair treatment of our data; we have uncertainties (δh_i's) for each data point, and the uncertainties might not be the same as each other. If we're very certain of the average height after 10 days and not at all certain of the 40-day average, is it fair to ask the line to go equally close to both data points? We need to revise our strategy to take uncertainties into account.

If we have several data points with different error bars, the fair strategy is to find (m, b) which make the line *miss each data point by as few error bars as possible*. This criterion is pretty much what motivates a process called ***chi-square fitting***, which finds parameters to minimize the quantity ***chi-square (χ^2)***:

$$\chi^2 = \sum_{i=1}^{N} \frac{(h_i - H(t_i))^2}{(\delta h_i)^2} . \qquad \text{(Eq. 6)}$$

Chi-square fitting can be carried out for any type of function, not just a straight line. Several commercial data analysis packages have built-in chi-square fitting for linear and nonlinear functions. Unfortunately, Microsoft Excel is not among these. Igor and Kaleidagraph, two programs available on the computers in the HMC physics labs, can carry out chi-square fits when proper error bars and instructions are given to them. You will learn to do chi-square fits in Igor as part of Physics 22.

When a chi-square fit is done, the best-fit function (of the type we have specified) is plotted on the graph on top of the experimental data points. The parameter values – in our example, the values of *m* and *b* – are given as well. One wonderful thing about a chi-square fit is that, in addition to values for the parameters, we also get uncertainties for the parameters. In our example, doing a chi-square fit to our data points will allow us to read off a value *m* and an uncertainty δm for the growth per day of bean sprout plants. Of course, the parameter values and uncertainties will only be legitimate if the input data values and uncertainties are legitimate themselves.

The second important output of a chi-square fit is the final value of chi-square itself, as defined in Equation 6. While chi-square fitting finds the best-fit function of the type we have specified, the final value of chi-square can tell us whether or not that type of function is a good choice to describe the data in the first place. After all, perhaps bean sprout growth just *isn't* linear, and the best straight line is still a horrible match to the data!

To determine whether a function fits data well, we calculate something called the ***reduced chi-square (*** $\tilde{\chi}^2$***):***

$$\tilde{\chi}^2 = \frac{\chi^2}{(\#\,datapts - \#\,parameters)} . \qquad \text{(Eq. 7)}$$

The quantity in the denominator of Equation 7 is known as the number of ***degrees of freedom*** in the fit, so reduced chi-square is sometimes referred to as ***chi-square per degree of freedom***. The reason for the denominator is roughly as follows: chi-square is a summed-squared deviation, and we'd like to judge a fit by an average rather than a sum. Thus it's natural to divide by the number of data points. On the other hand, if a function has many free parameters it can have many wiggles, and chances are it can wiggle through all the data points very nicely even if it has no fundamental relationship to the data. Therefore, it's sensible to deduct credit (so to speak) for the number of parameters in the function. In our example of linear bean sprout growth, we have 4 data points and the function *H(t)* has two parameters *m* and *b*. Therefore, in this example we have (4-2)=2 degrees of freedom, and $\tilde{\chi}^2 = \chi^2/2$.

What is a *good* value of reduced chi-square? Well, imagine that our theoretical function really does describe our data. Thus we expect the data values (h_i's) to be pretty close to the function values ($H(t_i)$'s). However, we estimate that our measurement missed the true value of each h_i by about δh_i. **Even if a theory is exactly right, we expect our data to miss it by about one error bar for each point.** Examining Equations 6 and 7, we see that this translates roughly to a reduced chi-square value of 1. Thus a reduced chi-square value of around 1 indicates that the type of function we have used is a good description of our data. Reduced chi-square values much larger than 1 suggest problems with the theory and/or experiment, while reduced chi-square values much smaller than 1 suggest over-inflated error estimates or mistakes in applying the chi-square fit.

Note that least squares fitting (as in Excel) completely misses the important features of chi-square fitting. It does not calculate the true optimal parameters, it fails to produce an estimate of their uncertainty, and it does not give a good measure of whether a function is consistent with data (and error bars). Straightforward least squares fitting is almost never the best basis for a truly quantitative experimental result. Chi-square fitting itself is by no means *always* the best treatment for data and uncertainties, but it is fairly easy to understand and is a good default method for many situations.

Concluding Remarks

There are many aspects of error estimation and analysis that we have not discussed in the preceding pages. We will discuss additional subtleties, rules of thumb, shortcuts, and other techniques as they arise in the experiments of Physics 22. Rigorous justifications of our techniques are reserved for future courses, or for independent reading and consultation with your instructor. However, the principles presented above should provide you with a foundation for the quantitative treatment of uncertainty in experimental science. Error analysis at the highest levels of experimental science continues to be a blend of rigorous statistics, careful observation, and plain old common sense. One of the goals of your work in Physics 22 should be to develop this blend by means of practice. Happy hunting!

Appendix A:
Error Propagation in Calculations: Functions of Several Measured Quantities

Real life is often not as simple as measuring *x* and finding *f(x)*. Many interesting things depend on more than one variable. It can therefore be useful to consider a function *g* of several variables: $g = g(x,y,z,\ldots)$.

To find the value of *g*, we industriously go into the lab and measure all the independent variables $x \pm \delta x$, $y \pm \delta y$, $z \pm \delta z$, etc. We can calculate a value for *g*, but what is the uncertainty δg? From our single-variable rule there is a contribution $\delta x\left|\frac{\partial g}{\partial x}\right|$ due to the uncertainty in *x*. (We change the *d*'s to ∂'s to denote the *partial* derivative with respect to x, since *g* is a function of multiple variables.) But there is also a contribution $\delta y\left|\frac{\partial g}{\partial y}\right|$, and a contribution $\delta z\left|\frac{\partial g}{\partial z}\right|$, and so on. How can we combine all these? The key lies in the realization that each uncertainty contribution is unrelated to the others; they are all independent, and not required to be in the same direction or opposite directions. We have already learned how to combine unrelated (independent) errors – they add in quadrature! Thus we can write down the *full* one and only rule of error propagation:

$$\delta g(x,y,z,\ldots) = \sqrt{\left(\delta x\left(\frac{\partial g}{\partial x}\right)\right)^2 + \left(\delta y\left(\frac{\partial g}{\partial y}\right)\right)^2 + \left(\delta z\left(\frac{\partial g}{\partial z}\right)\right)^2 + \ldots}$$

as long as *x, y, z, …* are variables independent of each other. (Eq. 8)

Let's take an example in which we wish to calculate the area of a rectangle. We measure the length $\ell = 2.0 \pm 0.1cm$ and the width $w = 1.2 \pm 0.1cm$. The area is then $A = \ell w = 2.4cm^2$, but it has an uncertainty

$$\begin{aligned}\delta A &= \sqrt{\left(\delta\ell\left(\frac{\partial A}{\partial \ell}\right)\right)^2 + \left(\delta w\left(\frac{\partial A}{\partial w}\right)\right)^2} \\ &= \sqrt{(\delta\ell(w))^2 + (\delta w(\ell))^2} \\ &= \sqrt{(0.1cm(1.2cm))^2 + (0.1cm(2.0cm))^2} \\ &= \sqrt{0.0144cm^4 + 0.04cm^4} = 0.2cm^2.\end{aligned}$$

Notice an interesting implication of this calculation: even though ℓ and w have the same individual uncertainty, they have unequal contributions to the uncertainty of A.

Appendix B:
A Reality Check for Error Propagation: Fractional Uncertainty

In the previous section, we presented and at least partially justified Equation 8 for the uncertainty in a function of several variables, based on the uncertainties in each of the measured variables. Armed with Equation 8, you need no other error propagation formulas – but for complicated functions it can be challenging or at least time-consuming to compute the final uncertainty, and it is useful to have some way of anticipating and/or reality-checking the answers you get. For this purpose, one of the most powerful tools for quickly checking error bar results is the concept of *fractional uncertainty*.

The fractional uncertainty in x is simply the name we give to the quantity ($\delta x/x$). Thinking in terms of fractional uncertainties is very useful, because the fractional uncertainty in many common functions ($\delta f/f$ or $\delta g/g$) is similar to the fractional uncertainty of the variables. While following Equation 8 is imperative for getting quantitatively correct uncertainties, considering fractional uncertainty is a much simpler way to see roughly what values those uncertainties should have.

To illustrate the usefulness of fractional uncertainty, consider propagating errors (using Equation 8) in several simple and commonly-encountered functions. First, we consider a product of two variables, possibly with a constant coefficient c:

$$g(x,y) \equiv cxy$$

$$\delta g = \sqrt{(\delta x)^2 (cy)^2 + (\delta y)^2 (cx)^2}$$

$$\frac{\delta g}{g} = \frac{\delta g}{cxy} = \sqrt{\frac{(\delta x)^2 c^2 y^2}{c^2 x^2 y^2} + \frac{(\delta y)^2 c^2 x^2}{c^2 x^2 y^2}} = \sqrt{\left(\frac{\delta x}{x}\right)^2 + \left(\frac{\delta y}{y}\right)^2}.$$

In this case the fractional uncertainty in g due to each variable, x or y, is actually equal to the fractional uncertainty in that variable itself. Or consider a slightly less straightforward relationship:

$$f(x,y) \equiv x^2 y$$

$$\delta f = \sqrt{(\delta x)^2 * (2xy)^2 + (\delta y)^2 * (x^2)^2}$$

$$\frac{\delta f}{f} = \frac{\delta f}{(x^2 y)} = \sqrt{\frac{(\delta x)^2 4x^2}{x^4 y^2} + \frac{(\delta y)^2 x^4}{x^4 y^2}} = \sqrt{\left(2\frac{\delta x}{x}\right)^2 + \left(\frac{\delta y}{y}\right)^2}.$$

Here the fractional uncertainty in f due to x is not quite equal to the fractional uncertainty in x. However, it is still comparable, and the relationship between them is a simple one. Many functions we encounter in nature are products and low-order polynomials of this sort; for them, comparing fractional uncertainties in functions and their variables can be a good way to arrive quickly at a roughly correct error propagation result.
However, keep in mind that for some functions Equation 8 does indeed lead to fractional uncertainties in functions which are not at all similar to the fractional uncertainties in the variables. For $x = 0.05 \pm 0.01$ radians, what are the value and the uncertainty of $\cos(x)$? For $x = 3.5 \pm 0.1$ cm and $y = 3.4 \pm 0.1$ cm, what are the value and uncertainty of $x - y$? As a final cautionary note, for $x = 10 \pm 1$ find the value and uncertainty in the function e^x.