Chapter 5: EXPERIMENTAL UNCERTAINTY

" 'I am no matchmaker, as you well know,' said Lady Russell, 'being much too aware of the uncertainty of all human events and calculations.'"
--- *Persuasion*

5.1 UNCERTAINTY AS A "95% CONFIDENCE RANGE"

We generally assume in physics that any quantity we measure has a "true" value, which is the result that we would get if we had a perfect measuring apparatus. Fifteen minutes in any laboratory, regardless of the sophistication of the equipment, will rapidly disabuse you of the notion that any measurement apparatus is perfect. Real measurement devices suffer from a variety of imperfections that limit our knowledge of the "true" value of any measurement. Devices may be poorly made, out of adjustment, subject to noise or other random effects, or hard to read accurately, and all devices read to only a finite number of digits. These problems mean that the exact value of any measured quantity will always be uncertain.

Uncertainty is therefore an unavoidable part of the measurement process. We will (of course) always seek to reduce measurement uncertainty whenever possible, but ultimately, there will remain some basic uncertainty that cannot be removed. At this point, our task is to estimate thoughtfully the size of the uncertainty and clearly communicate the result.

How can one quantify uncertainty? In this course, we will *define* a value's uncertainty in terms of the range within which we are 95% confident that the "true value" would be found if we could measure it perfectly. This means that we expect that there is only one chance in 20 that the true value does *not* lie within the specified range. This range is called the **95% confidence interval**.

The conventional way of specifying this range is to state the measurement value plus or minus a certain number. For example, we might say that the length of an object is 25.2 cm ± 0.2 cm: the measured value in this case is 25.2 cm, and the **uncertainty** U in this value is *given as* to be ±0.2 cm. The uncertainty thus has a magnitude equal to the difference between the measured value and either extreme edge of the uncertainty range. This statement means that we are 95% confident that the measurement's true value lies within the range 25.0 cm to 25.4 cm.

It may be helpful to point out that there is nothing magic about our choice of the 95% confidence interval for stating the uncertainty. We could just as properly decided to define the uncertainty as the 50% confidence interval (half the measurements lie within the confidence interval) or the 0.1% confidence interval (only one measurement in 1000 lies outside the confidence interval). If you think about the proposed uncertainties just given, though, we think you will agree that the 50% confidence interval excludes too many reasonably likely measurements, while you would need to make 1000 measurements to find the 0.1% confidence interval in the first place, at least using the method described in Section 5.5. We will also see in Section 5.6 that the 95% confidence interval is particularly easy to estimate from the standard deviation, another reason for choosing it for reporting the uncertainty of a measurement.

Now (as you may have already noticed), this definition of uncertainty is rather fuzzy: one person may be more confident about a value's precision than the next. In fact, the definition seems almost devoid of objective meaning except as a description of the experimenter's frame of mind. We will see that the definition is *not* as subjective as it seems, particularly in certain cases to be discussed shortly, where it is possible to make an educated and generally agreed-upon estimate of the uncertainty. But the fact remains that "uncertainty" is *itself* an uncertain concept, and uncertainties should only be taken to be rough estimates, good to one or *at most* two significant digits.

In spite of this problem, knowing the uncertainty of a measured value is *essential* if one is to correctly interpret the meaning of a measured value. For example, imagine that you measure the period of a (very long) pendulum to be 12.3 sec. Imagine that someone's theory predicts that the period should be 11.89275 sec. Is your result consistent with that theory or not? The answer to this question depends *entirely* on the uncertainty of your result. If your result has an uncertainty of ±0.5 s, then the true value of your measured duration could quite easily be the same as the theoretical value. On the other hand, if the uncertainty in your result is ±0.1 s, then it is *not* very likely (less than one chance in 20) that the true value behind your measurement is the same as the predicted value, meaning that the theory is probably wrong. What a measurement *means*, therefore, can depend crucially on its uncertainty!

5.2 SYSTEMATIC ERRORS

Why aren't measurements perfect? The causes of measurement errors can be divided into three broad classes: *systematic problems*, *limited precision*, and *random effects*. The focus of this chapter will be on the last of these, but the first two causes need to be discussed briefly.

Systematic errors occur when a piece of equipment is improperly constructed, calibrated, or used, or when some physical process is going on in the experiment that you haven't thought of in your experimental design. As a somewhat contrived example of problematic equipment, suppose that you measured lengths with a meter stick that you failed to notice had been cut off at the 5 cm mark. This would mean that all of your measured values would be 5 cm too long. Systematic errors resulting from equipment problems are relatively easy to identify once you have some reason to suspect they exist; you compare your equipment to two other, similar, pieces of equipment and see if they all give the same result for the measurement. If your device doesn't agree with the other two, probably your device has a problem.

As an (also somewhat contrived) example of a systematic error due to unexpected physics, suppose you were trying to measure the acceleration of gravity by timing the motion of a falling object. But – and this is the contrived part, at least after the invention of Newtonian physics – suppose you didn't know about air resistance, so you used for your object a wadded-up ball of paper. You would find a much smaller value for g than the accepted value!

You don't normally include systematic errors in the uncertainty of a measurement; if you know that a systematic problem exists, you should *fix the problem*. In the meter-stick example above, you would use a complete meter stick, or add 5 cm to all your measurements. Systematic errors arising from unanticipated physics are harder to find, although they're the source of many Nobel Prizes in physics. Unfortunately, no well-defined procedures exist for finding systematic

errors: the best that you can do is to be clever in anticipating problems and alert to trends in your data that suggest their presence. You might come to suspect the short meter stick, for example, if you noticed that your data would agree with theoretical predictions if all of your length measurements were about 5 cm shorter. In the second example, you would come to suspect the presence of air resistance if you let the ball fall through several different heights and observed that your calculated acceleration got systematically smaller with increasing height. To notice this, though, you do have to think of carrying out those measurements at different heights.

In some cases, it *is* appropriate to estimate the magnitude of possible systematic errors and include them in the uncertainty of a measurement result. For example, it is possible to read most automobile speedometers to a precision of about 1 mph. But it is well known (to some people, at least) that variations in the manufacture and calibration of speedometers mean that the reading of a typical speedometer may be off by as much as 5%. Therefore, the uncertainty of a speedometer reading of about 60 mph would have to be taken to be ± 5% of 60 mph, or about ± 3 mph. The uncertainty in this case is called a **calibration uncertainty**. We will deal with calibration uncertainties only rarely in this course.

5.3 LIMITED PRECISION

No measurement device can read a value to infinite precision. Dials and linear scales, such as meter sticks, thermometers, gauges, speedometers, and the like, can *at best* be read to *one tenth of the smallest division* on the scale. For example, the smallest divisions on a typical metric ruler are 1 mm apart: the *minimum* uncertainty for any measurement made with such a ruler is therefore about ± 0.1 mm. This statement is *not* an arbitrary definition or convention: rather, it is a rule based on experience. If you try using a ruler to make as precise a measurement as you can, you should be able to see that it is really quite difficult to do better than the stated limit.

For measuring devices having a digital readout, the minimum uncertainty is ± *1 in the last digit*. For example, imagine that you measure a time interval with a stopwatch, and find the result to be 2.02 s. The measurement's "true value" in this case *could* be anywhere from 2.010...01 s to 2.0299...99 s. We cannot narrow this range without knowing details about the design of the stopwatch: does it round up to 2.02 s just after the true elapsed time exceeds 2.01 s, or does it round to the nearest hundredth of a second, or does it not register 2.02 s until at least 2.02 s have passed, or what? So in this case, we must take the uncertainty to be at least ± 0.01 s.

In both of the cases described above, these rules are meant represent the *minimum* possible uncertainties for a measured value. Other effects might conspire to make measurements *more* uncertain than the limits given (as we shall see), but there is nothing that one can do to make the uncertainties smaller, short of buying a new device with a finer scale or more digits.

5.4 RANDOM EFFECTS

This chapter is mainly focused on the analysis of *random effects*. It is commonly the case that repeated measurements of the same quantity do not yield the *same* values, but rather a *spread* of values. For example, you might determine the speed of sound by standing at a fairly large distance (like the width of Marston Quad) away from an object that simultaneously emits a

flash of light and a loud sound when someone flips a switch. After measuring the distance to the source, you would measure the time between seeing the flash and hearing the sound. If you measure this interval five times, you are almost certain to get five different results (for example, 0.54 s, 0.52 s, 0.55 s, 0.49 s, and 0.53 s).

Why are these results different? In this case, the problem is that it is difficult for you to start and stop the stopwatch at *exactly* the right instant: no matter how hard you try to be exact, sometimes you will press the stopwatch button a bit too early and sometimes a bit too late. These unavoidable and essentially random measurement errors cause the results of successive measurements of the same quantity to vary.

Random perturbing effects, which are sometimes human and sometimes physical in origin, are a feature of almost all measurement processes. Sometimes a measuring device is too crude to register such effects: for example, a stopwatch accurate to only one decimal place might read 0.6 s for each of the measurements in the case described above. But laboratory instruments are often chosen to be just sufficiently sensitive to register random effects. Granted, you'd like as precise an instrument as possible, but there is no point in buying an instrument much more precise than the limit imposed by unavoidable random effects. For example, a hand-held timer like a stopwatch that reads to a hundredth of a second is better than one that registers to only a tenth, because (as we'll see in a couple of weeks) it's possible to pound down your random experimental uncertainty down to a few hundredths of a second by making enough measurements. But there would be no scientific point, although there might be a marketing advantage, in making a stopwatch that reads to a thousandth of a second, because the added precision of the watch would be swamped by the scatter in the measurements resulting from its operation by a human being.

The point is that random effects will be an important factor of many of the measurements that you will encounter in any scientific experiment. Now, it should be clear that such effects increase the uncertainty in a measurement. In the stopwatch case, for example, that different trials lead to results differing by several hundredths of a second implies that the uncertainty in any given measurement value is *larger* than the basic ± 0.01 s uncertainty imposed by the digital readout.

5.5 THE DISTRIBUTION OF VALUES DUE TO RANDOM EFFECTS

The first step towards describing the magnitude of the uncertainty due to random effects is to understand more precisely what these effects do to a set of measurement values. Consider, for example, a simple experiment where 25 different people measure the length of a soft-drink can with a ruler. Assume that the "true" length of the can is 8.51293... cm. None of the 25 people will measure the object to have exactly this length, of course, because people will view the ruler and can from slightly different angles, make different judgments about the exact reading, and so on.

Nevertheless, we would expect the measurements to cluster around the value 8.51 cm, with most of them agreeing with that result within a few hundredths of a centimeter or so (that is, tenths of a millimeter). Results different from the true value by roughly 0.1 cm will be less common, but not really rare. Results that differ much more dramatically from the true value will be rarer, and the more that a measurement value differs from the true value, the less likely it is to occur.

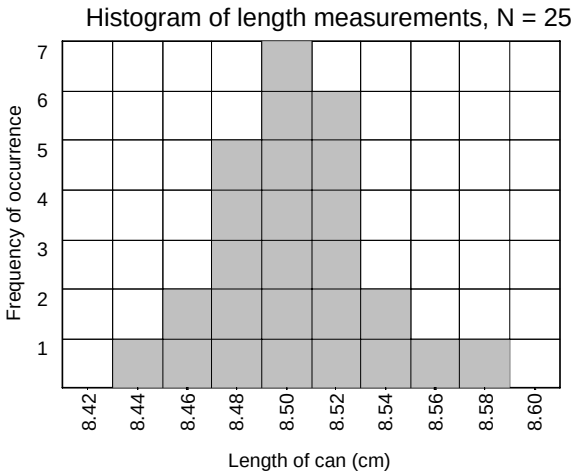


Fig. 5.1: Distribution of 25 measurements of the length of a can

If one were to plot the number of measurements obtained versus the measurement value, we might obtain a graph looking something like the graph shown in Figure 5.1. Note that the range of measurement values has been divided into "bins", each 0.02 cm wide, so, for example, measurement values of 8.50 cm and 8.51 cm would both be counted as being in the central bin. (The purpose of grouping values into bins like this is to show more clearly the characteristic shape of the distribution: a graph where each bin was only 0.01 cm wide would be flatter and less revealing.) A graph of this type is called a **histogram**.

This graph roughly sketches what is often called a "bell-shaped curve." If we were to plot 100 or 1000 measurements on this graph instead of just 25, the curve would be more smooth, symmetrical, and bell-like. Measurement values subject to random effects are almost always distributed in such a pattern. In fact, it is possible to show that a bell-shaped distribution of values having specific and well-defined characteristics is the mathematical consequence of perturbing effects that are truly random in nature and continuously variable in size. We call the specific bell-shaped distribution of values caused by such random influences a **normal** or **gaussian** distribution.

Simply by looking at this graph, we can make a rough estimate of the uncertainty of any individual measurement value. The definition of "uncertainty" that we have adopted implies that the uncertainty range should enclose the true value (in this case 8.5129... cm) about 19 out of 20 times. In the case shown above, a range of ±0.06 cm attached to any of the measurements would include the true value, except for the *one* case in the rightmost bin. One out of 25 is roughly equal to one out of 20, so ±0.06 cm would a reasonably good estimate of the uncertainty of a given measurement in this (hypothetical) case.

As mentioned in Section 5.1, we could have chosen to define experimental uncertainty with a 50% confidence interval or a 0.1% confidence interval. It's pretty clear that we don't have enough measurements to find a 0.1% confidence interval! Finding the 50% confidence interval for the data above should be easy: we just have to throw out the 12 or so measurements that lie farthest from the mean. Doing this for the given set of measurements does raise a

somewhat troubling point, though. The 12 measurements to eliminate that seem most obvious are probably the eight measurements in the bins labeled 8.44, 8.46, and 8.48, and the four measurements in the bins labeled 8.54, 8.56, and 8.58. But this tosses out measurements rather unsymmetrically, which doesn't seem like a good idea. This problem could be caused by our choice of bin size, but smaller bins would spread out the histogram and make the distribution less clear. So the 95% confidence interval seems to be a good compromise between keeping enough data points for a symmetric distribution of data, and having to make huge numbers of measurements.

5.6 THE MEANING OF THE STANDARD DEVIATION

In chapter 4 of this manual, we defined the **standard deviation** s of a set of N measurements $x_1, x_2, x_3, \ldots x_N$ with mean $\bar{x}$ to be given by the expression

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2 + \ldots + (x_N - \bar{x})^2}{N - 1}} \qquad (5.1)$$

Let the symbol x_i stand for an arbitrary "*i*th" measurement in our set. In chapter 4, we said that the standard deviation was a measure of how much any one data point was likely to differ from the mean of the set of measurements, so it's a good candidate for describing the experimental uncertainty of a set of measurements. In this chapter, though, we're introducing a different definition, one that looks at all the measurements and characterizes the uncertainty by the range of all but possibly a few of the measurements. You probably won't be surprised to find out that the two ways of estimating the uncertainty are related to each other.

Recall that we have defined the uncertainty U of any measurement x_i to be the value such that we are 95% confident that the "true value" of the measured quantity lies within the range $x_i \pm U$. If we have happened to take a large number of measurements of this quantity, our otherwise somewhat subjective "95% confidence" can be given a directly quantitative meaning: the measurement's true value (which should correspond to the value at the peak of the bell curve) should lie within the range $x_i \pm U$ for 95% of the measurements x_i. Given a set of measurement values, then, we can use this criterion to determine the value of U. The only problem is that we need hundreds (if not thousands) of measurements to accurately estimate U this way; to accurately determine U, N must be large enough that the number of measurements in the 5% that *exclude* the true value is more than just a handful.

Fortunately, mathematicians have shown that it is possible to accurately estimate the value of U that *would* have this property for a very large set of measurements from a much smaller set. The uncertainty U of any given *single* measurement can be estimated using the standard deviation of a small set of similar measurements as follows:

$$U \approx ts \quad \text{(uncertainty of a } \textit{single} \text{ measurement)} \qquad (5.2)$$

where t is the so-called **Student *t*-factor**, a number that depends somewhat on N, the number of measurements in the set used to calculate s. A table of t-values as a function of N is given below, and is also reproduced in the inside front cover of this manual.

TABLE OF STUDENT *t*-VALUES

N	t-value	N	t-value
2	12.7	10	2.26
3	4.3	12	2.2
4	3.2	15	2.15
5	2.8	20	2.09
6	2.6	30	2.05
7	2.5	50	2.01
8	2.4	100	1.98
9	2.3	∞	1.97

And this is the other reason to define the experimental uncertainty as the 95% confidence interval, and not the 50% confidence interval or the 0.1% confidence interval. The *t*-value for a reasonably but not onerously large number of measurements (10 or more) is pretty close to 2. Two is easy to remember, and so is 95%. (Two is almost too easy to remember: see the next paragraph.) The appropriate factor for the 50% confidence interval would be about 0.675; that for the 0.1% confidence interval would be 3.29. So once again, the 95% confidence interval is chosen for utility, not because there's anything magic about 95%.

While uncertainties are generally accurate only to one significant digit, this table states values to two or three significant digits to show clearly the difference between adjacent values. Note that for $N > 30$, the *t*-value is within a few percent of being 2.0: for this reason, some books will tell you that the 95% confidence range for a given measurement x_i is simply $x_i \pm 2s$. However, this is *not* a good estimate of that range for the small values of N that we will commonly encounter. In using the table, you should also keep in mind that it is really only valid for measurements that are *randomly* distributed. Specifically, if the uncertainty of your measurement is limited by the precision of the apparatus rather than random effects, you should *not* use equation 5.2: you should instead use one of the strategies outlined in section 5.3.

Please note that equation 5.2 estimates the uncertainty of any given *single* measurement in the set. As we'll see later, though, if we have already bothered to take a set of measurements required to determine *s*, we might as well compute the *mean* of the set, which is a better estimate of the measurement's true value (that is, it has a smaller uncertainty) than any arbitrarily chosen single measurement is.

Note in addition that the table is telling you indirectly something about the number of measurements you need to get a good estimate of the uncertainty. In particular, two measurements are not enough! Three measurements are a bare minimum, and five is a good compromise between getting a good estimate of the uncertainty and spending lots of time on a single measurement.

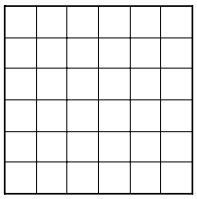
EXERCISES

Exercise 5.1
A standard household thermometer has one mark for every two °F. What is the *minimum* uncertainty that you should assign to the temperature that you read from such a thermometer? What do you think is the *best* uncertainty to assign to this reading, do you think larger than this? Explain your reasoning. (Be aware that there is no absolute right answer to this last question.)

Exercise 5.2
According to your digital bedside clock, it took you exactly 12 minutes to dress for class some morning. What uncertainty should you assign to this result? Explain your reasoning.

Exercise 5.3
Imagine that you are one of ten different people who measure the time of flight of a thrown baseball. Assume that these ten measured speeds are as listed below and to the left. On the grid below and to the right, plot a histogram of this data. (Choose a "bin" size that displays this data as a pseudo bell-curve rather than scattered data or only one or two columns.)

2.53 s 2.58 s
2.67 s 2.63 s
2.59 s 2.60 s
2.62 s 2.56 s
2.66 s 2.61 s



Exercise 5.4
Compute the standard deviation of the data from the previous problem, and estimate the uncertainty of your particular measurement.