Error Analysis: An Introduction

Welcome to PSI! We sincerely hope that you'll find this portion of the course both enjoyable and informative. Physics, like all sciences, is an experimental discipline. All the elegant theories developed by Newton, Maxwell, Einstein, and others would be essentially irrelevant if not for the fact that experiments have demonstrated that they accurately describe the natural world. In PSI this semester, you'll be introduced to a new perspective on not only physics, but science in general. If you can understand this material, you'll someday be a better scientist, a better doctor, or simply a better-informed citizen of society. You'll be much less easily fooled by misleading statistics cited in the press. And your love life may also improve (though this is not guaranteed).

Early in your education, you may have been taught that the scientific method consists of such steps as formulating a hypothesis, designing and carrying out an experiment, analyzing the results, and drawing a conclusion. In real life, the process is never quite so cut and dried; instead, you will often find yourself going in circles:

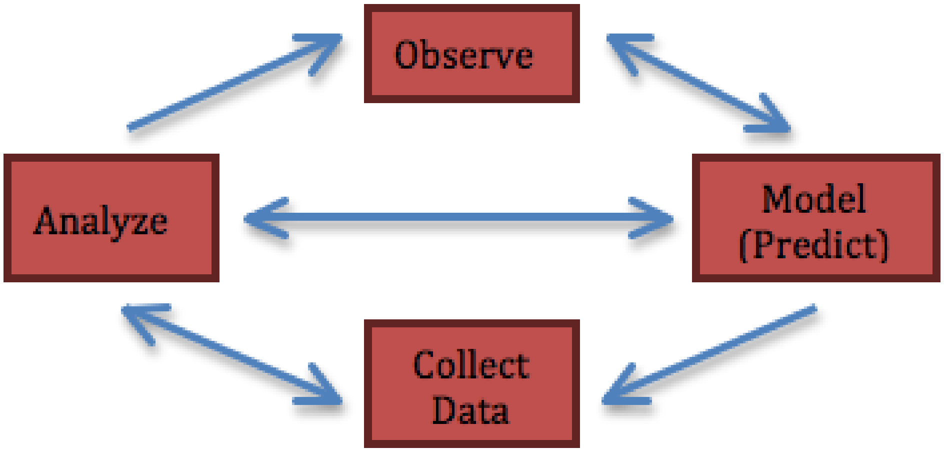

Figure 1: One representation of the scientific process

In your lectures, reading, discussion section, and problem sets, you will get plenty of practice with the kind of problem-solving required for the bubble on the right, labeled "model/predict". This step typically requires you to apply a general model (e.g. Newtonian mechanics) to a particular system to arrive at a relevant equation or set of equations to use as a hypothesis. In PSI, we'll be looking at all of the steps of the process, but for now, let's just focus on one of them, the one represented by the bubble on the left, "Analyze". It is an inescapable truth that any scientific experiment must include, at some point, *a comparison of experimental results to the predictions of some theoretical model.* But the act of comparison is not quite as simple as you might think. What if the model predicts that some speed should be $v = 1.08$ m/s and then you measure a speed of 1 m/s? Does that agree with the model or not?

This question absolutely cannot be answered unless you know the *uncertainty* in the experimental result (and perhaps also in the prediction of the model). If the measured speed is 1.0 ± 0.2 m/s, then it is consistent with the prediction. If the measured speed is 1.00 ± 0.01 m/s, then it is not.

In other words, in experimental science *it is critical to specify not only what we know but how well we know it.* The same experiment, producing the same result of "1 m/s", can be used to either confirm or reject the same hypothesis, depending only on the uncertainty in the result! The purpose of this document is to help you understand some of the framework in which we will discuss experimental uncertainty. That framework is, as you will discover, statistical. So we will begin with a discussion of probability and statistics.

1 Probability Distributions

1.1 Tall Women, Short Women

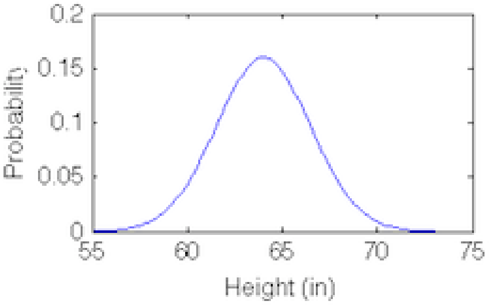

Figure 2: A probability distribution

Suppose you wanted to know the distribution of heights of adult women in the U.S. You could answer this by taking a very large sample of women and actually measuring their heights, but before you start such a significant undertaking, perhaps it would be best to consider what kind of results you *expect* to get. Common sense, or qualitative reasoning, suggests that there is a typical average height, and most women are approximately this height. A few are several inches taller or shorter than average. A very small fraction are much taller (e.g. a foot taller) or much shorter than average. And the further you get away from the average, the fewer women you expect to find that are that tall or that short.

In brief, you might expect the distribution to look something like the bell-shaped curve shown in Figure 2. This graph represents a *probability distribution function.* The x-axis has the different possible values of whatever variable you want to know the distribution of—in this case, height (of adult women in the U.S.), measured in inches. The y-axis shows the relative probability. If the graph is twice as tall at $x = x_1$ as it is at $x = x_2$, that means x_1 is twice as likely a value for x as x_2. A quick glance shows that the ditsribution shown in Figure 2 has, qualitatively, the features we expect: the most likely heights are in the middle, and as you move away from the middle, the probability of finding somebody with that height drops off quickly. A woman is about sixteen times more likely to be 64 inches tall (5'4") than she is to be 58 inches (4'10") or 70 inches (5'10").

1.2 Properties of Probability Distributions

That's well and good for relative probability. But what about absolute probability? What are the chances of a random adult woman being 67 inches tall? This is tricky—height is a continuous variable, so you have to clarify exactly what you mean by "67 inches tall". What about somebody whose height is 67.2 inches? Does she count? What about 67.03? The question really only becomes meaningful when we rephrase it to talk about a *range* of values. For instance, we could ask, "What are the chances of a woman being between 66 and 68 inches tall?" (Or between 66.5 and 67.5, etc.) The answer is that *the probability is equal to the area under the curve* in the graph of the distribution function.

Using the language of integral calculus, we can be more precise about this statement. If the variable whose distribution we are studying is x, and the probability distribution function is $f(x)$, then

$$\boxed{\begin{array}{l}\text{probability that } x \text{ will} \\ \text{fall between } a \text{ and } b\end{array} = \int_a^b f(x)\,dx.} \tag{1}$$

One thing that we can immediately note from this equation is that f must have dimensions reciprocal to that of x itself, because $f(x)\,dx$ is a probability, and therefore dimensionless. (Remember, dx is just a small change in x or a little bit of x, so it has the same dimensions as x.) This clarifies the units on the y-axis of Figure 2, which had been left unlabeled.

Incidentally, any given person will obviously have *some* height, so the total probability that the height will be between $-\infty$ and ∞ should be 1. Graphically, this means that the total area under the curve should be equal to 1. In equation form:

$$\int_{-\infty}^{\infty} f(x)\,dx = 1. \tag{2}$$

Equation (2) is sometimes called the *normalization condition.* If it is not satisfied by a particular $f(x)$, you can multiply $f(x)$ by a constant in order to satisfy equation (2), without changing the shape of the distribution. $f(x)$ is then said to be *normalized.*

1.3 Examples of Other Distributions

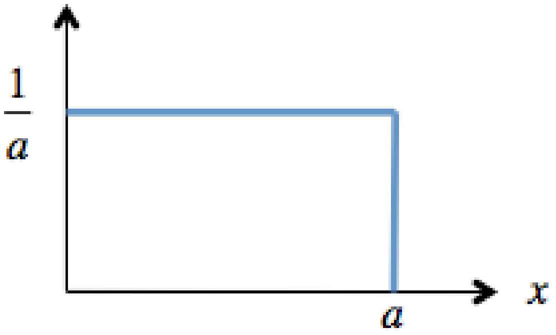

Figure 3: Uniform distribution on the interval from 0 to a

The bell-shaped curve describing the distribution of women's heights is an interesting and highly relevant example in this course, but it's hardly the only thing a distribution could look like. For instance, Figure 3 shows the uniform distribution over the interval from $x = 0$ to $x = a$. In the uniform distribution, any x-value between 0 and a is equally likely, and all other x-values have a probability of zero. (If the distribution is normalized, the height of the distribution must be equal to $1/a$ so that the area—height times width—will be 1.) Admittedly the uniform distribution is pretty boring, but it has practical applications: the reading on the second hand of a clock, for instance, is pretty well described by the uniform distribution over the interval from 0 to 60 s, assuming you don't look at it very often or in a deliberately periodic manner.

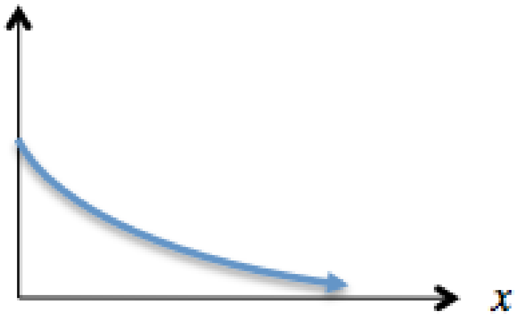

Figure 4: Exponential distribution

Figure 4 shows another example, the exponential distribution. The exponential distribution has the functional form

$$f(x) = Ae^{-\lambda x} \tag{3}$$

where λ is some positive parameter. The exponential distribution is typically defined only for positive x; that is, this distribution characterizes variables which can only be positive. If the distribution is properly normalized, the constant A must take the value $A = \lambda$, as you can verify by plugging equation (3) into the normalization condition (2). In this case, the limits of integration would be 0 to ∞ instead of $-\infty$ to ∞, since the variable x is constrained to be strictly positive. (To look at it another way, you could just define $f(x)$ to be zero for $x < 0$.)

Many variables in nature are exponentially distributed. Examples include:

- the distance between random mutations on a strand of DNA
- the time (or distance) that a molecule in a liquid or gas travels between collisions with other molecules
- the time until a radioactive particle decays
- the energy of particles at thermal equilibrium at a given temperature T. In this special case, the exponential distribution is called the *Boltzmann distribution*, and λ takes the value $1/k_BT$, where k_B is a fundamental constant called Boltzmann's constant.

1.4 Mean and Expected Value

Now that we know how to calculate probabilities based on the distribution function, we can also calculate the average value of x. But first, we have to clarify what we mean by "average". What we want to do is take a *weighted* average of the possible values of x; they should be weighted according to how likely they are to occur. Of course, this is exactly what the distribution function $f(x)$ is supposed to tell us, so we can use $f(x)\,dx$ to weight the average. The result is the expected value of x according to the distribution $f(x)$, which we also call the *mean* of the distribution and denote using angled brackets:

$$\boxed{\langle x\rangle = \int_{-\infty}^{\infty} x f(x)\,dx} \tag{4}$$

Equation (4) defines the average value of x, but you might want to know the average value of something else, some variable that depends on x. But this is easy—$f(x)\,dx$ is the weighting function for taking any average. So, for instance, the average value of x^2 is

$$\langle x^2\rangle = \int_{-\infty}^{\infty} x^2 f(x)\,dx. \tag{5}$$

More generally, the average value of any quantity Q which depends on x is

$$\boxed{\langle Q(x)\rangle = \int_{-\infty}^{\infty} Q(x)\,f(x)\,dx} \tag{6}$$

Equation (6) defines what is ofted called the *expected value* of Q. As we've already seen, the expected value of x itself is just the mean.

1.5 Variance and Standard Deviation

The mean tells you the central value of a distribution, but it doesn't tell you how tightly distribution is bunched around that central value. The two distributions in Figure 5 have the same mean, but one is sharply peaked about the mean, and the other has a much wider spread. Can we characterize this amount of this spread?

The first thing you might do is take the expected value of the distance from x to the mean; in other words, calculate $\langle Q\rangle$ for the quantity $Q(x) = x - \langle x\rangle$. Great idea, but unfortunately, $\langle Q\rangle$ is always equal to 0 for any $f(x)$, because Q is sometimes negative and sometimes positive, just often enough to average out to 0:

$$\begin{aligned}\langle Q\rangle &= \int_{-\infty}^{\infty} (x-\langle x\rangle)\,f(x)\,dx\\ &= \int_{-\infty}^{\infty} x f(x)\,dx \;-\; \langle x\rangle \int_{-\infty}^{\infty} f(x)\,dx\\ &= \langle x\rangle - \langle x\rangle\cdot 1\\ &= 0.\end{aligned} \tag{7}$$

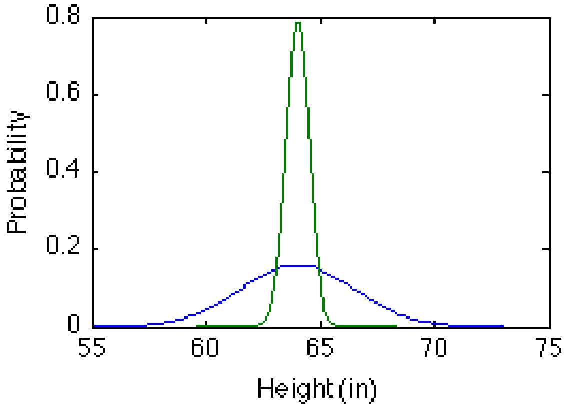

Figure 5: Two distributions with the same mean; one is clustered narrowly about the mean, while the other is widely spread out.

So that's less than useful—but we're on the right track. What if we try $Q = (x - \langle x \rangle)^2$? This Q is always positive, so it won't average out to zero. And it measures what we want: how far away is x from its mean value $\langle x \rangle$? The expected value of this Q is called the *variance* of x:

$$\boxed{\mathrm{Var}(x) \equiv \left\langle (x - \langle x \rangle)^2 \right\rangle = \int_{-\infty}^{\infty} (x - \langle x \rangle)^2 \, f(x) \, dx} \tag{8}$$

The variance is a useful quantity, but it scales as the square of x (i.e. if you multiplied x by a constant a, the variance would increase by a factor of a^2), and has dimensions which are the square of the dimensions of x (so if x were a length, $\mathrm{Var}(x)$ would be a length squared). For that reason, we normally prefer to work with the square root of the variance, a quantity called the *standard deviation* of x:

$$\boxed{\sigma_x \equiv \sqrt{\mathrm{Var}(x)} = \sqrt{\langle (x - \langle x \rangle)^2 \rangle}} \tag{9}$$

The symbol σ is the Greek lowercase letter sigma, and it stands for standard deviation. σ_x is a measure of the spread in the values of x. Roughly speaking, σ_x is the answer to the question, "how wide an interval around the mean do I have to take before I've included most of the total area under the curve $f(x)$?"[1] The standard deviation has the same dimensions and units as the original variable x. For example, if you are studying a distribution of times t according to the distribution function $f(t)$, then σ_t would also be a time, measured in seconds (or whatever units you use for t).

1.6 The Gaussian Distribution

Now that we have seen some examples of distribution functions and studied their characteristics, let's go back and take a closer look at the symmetric, bell-shaped distribution that characterizes the heights of adult women (Figure 2). It's a very important example—in fact, the most important example—of a distribution, because it turns out to be ubiquitous, for reasons we'll get to a little bit later.

For starters, this distribution has a name. Actually, confusingly enough, it has several names. It is sometimes just referred to as a bell curve, and is also commonly called the normal distribution, but in this course we'll always refer to it as the *Gaussian distribution.* (As an aside, there are at least five, perhaps six, technical definitions of "normal", in addition to the common-language definition, in math and physics. To avoid confusion, we'll try not to use too many of them in a context where it might be confusing.) The Gaussian distribution is named after the great mathematician K. F. Gauss and is pronounced (at least by Americans) GOW-see-an.

Mathematically, the Gaussian distribution is described by the equation

$$\boxed{G(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}} \tag{10}$$

where μ and σ (the Greek lowercase letters mu and sigma, respectively) are two parameters that characterize the Gaussian. (Generally, σ is taken to be positive; μ can be either positive or negative.) The function in Figure 2 is a Gaussian with $\mu = 64$ inches and $\sigma = 2.5$ inches.

Definition (10) looks complicated at first, but it's really just a dressed-up version of the function $f(x) = \exp(-x^2)$, a bell curve centered at $x = 0$. There are three simple changes:

1. The exponent is divided by a factor of $2\sigma^2$. This just stretches the distribution horizontally about the peak. For large σ, the distribution is wide; for small σ, it is narrow.

[1] Later on, we'll tighten up this statement by specifying exactly what we mean by "most" of the area.

2. x has been replaced by $x - \mu$. This shifts the peak of the distribution from $x = 0$ to $x = \mu$.
3. There is an overall factor of $1/\sigma\sqrt{2\pi}$ in front. This does not affect the relative probabilities; it is only there so that the function will be correctly normalized.

Let's see if we can gain a little bit of mathematical intuition about these changes. We'll start by taking a look at the behavior of the function

$$e^{-x^2/2\sigma^2}, \tag{11}$$

for different values of the parameter σ. Looking at the graph of this function for $\sigma = 1$ in Figure 6, you can see that it has the generally bell-shaped outline we've talked about. It has a maximum value of 1 at $x = 0$, and is symmetric about that maximum because it has the same value for x as for $-x$. As x gets further from 0, the value of the function decreases quite quickly because $-x^2/2\sigma^2$ becomes a large negative number, and the exponential of a large negative number is almost zero. Figure 6 also shows a graph of the same function for $\sigma = 2$. As you can see, the peak is broader; essentially, σ signifies the width of the function[2].

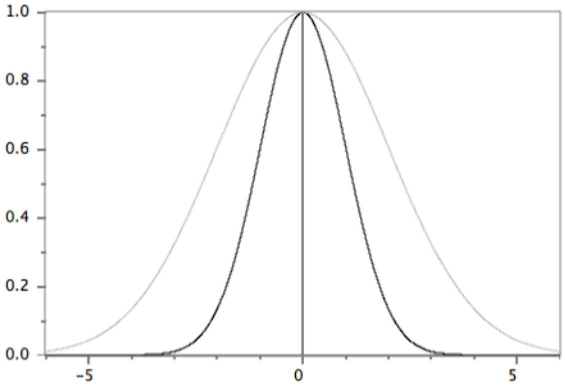

Figure 6: The function (11) for $\sigma = 1$ (black) and $\sigma = 2$ (gray).

Now let's look at the second change. The function in equation (11) is centered at $x = 0$. If we wanted the same function shifted over so that its peak was at some other point $x = \mu$, we can simply replace x in the original function by $x - \mu$: the function

$$e^{-(x-\mu)^2/2\sigma^2} \tag{12}$$

has its maximum at $x = \mu$ and falls off symmetrically on either side of that point, as you can see from Figure 7.

As for the third change: the function (12) is almost, but not quite, a proper probability distribution function. As written, it does not satisfy the normalization condition (2). However, that is easily fixed: we merely evaluate the integral $I = \int_{-\infty}^{\infty} f(x)\,dx$ and then multiply (12) by $1/I$. It turns out that $I = \sigma\sqrt{2\pi}$, which gives us the normalization factor above.

The two parameters μ and σ that characterize a Gaussian turn out to be the mean and standard deviation, respectively, of the distribution. This should make intuitive sense. The Gaussian is symmetric

[2] This is a somewhat loose usage of the word "width"; technically, the function (11) would seem to be infinitely wide, since it remains non-zero for all values of x. But it is perfectly clear what we mean when we say that the $\sigma = 2$ curve is *wider*, so we'll leave it at that.

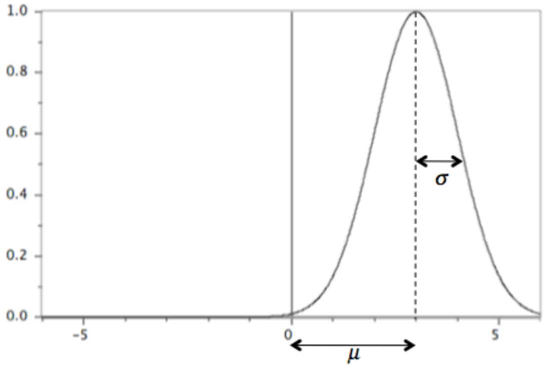

Figure 7: The function (12) for $\mu = 3$, $\sigma = 1$. It is centered on $x = 3$.

about its peak value at $x = \mu$, so that has to be the mean. σ as the standard deviation is less obvious, but in fact the Gaussian has been defined just so that σ would be equal to the standard deviation.

Of course, you don't have to take my word for it: we can explicitly calculate the mean and standard deviation of a probability distribution using equations (4) and (8). When we plug in the Gaussian distribution, we get:

$$\langle x \rangle = \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} = \mu \tag{13}$$

and

$$\sigma_x^2 = \int_{-\infty}^{\infty} (x - \langle x \rangle)^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} = \sigma^2. \tag{14}$$

(The explicit evaluation of these integrals is left as a tedious mathematical exercise.)

2 Sampling

Now that we understand the properties of probability distributions, let's consider the issue of random sampling. To sample from a distribution means to take one or more values which "obey" that distribution. This is not a math or statistics course, so we won't bother rigorously defining what "obey" means, but the general sense is that if you took a very large sample, the distribution of the data points would look arbitrarily close to the underlying probability distribution.

The reason we're interested in distributions and sampling is that we believe, fundamentally, that *any physical measurement is a sampling from some underlying probability distribution.* The problem is that the underlying distribution itself is not known, so we need to figure out a way to characterize it as best we can based on the sample. So it would behoove us to understand the relationship between sampling and distributions.

2.1 Sample Mean and Sample Standard Deviation

Let's start with an example. Suppose you drop a stone out the window and measure the time it takes for it to fall to the ground using a stopwatch. You repeat the experiment a total of five times, and get the results shown in Table 1 (all times reported in hundredths of a second).

Looking at this data, you might ask yourself three questions:

Trial	Time
1	64
2	69
3	67
4	75
5	71

Table 1: Data from the stone-dropping experiment.

1. Why aren't all the values the same?

2. Given this data, what is our best estimate of the time it takes for the stone to drop to the ground?

3. How good is this best estimate?

The first question turns out to be a fairly deep one, so we'll put it off for now. As for the second, perhaps unsurprisingly, the answer turns out to be the *mean* of the measurements. In our case, the mean is given by

$$\frac{64 + 69 + 67 + 75 + 71}{5} = 69.2 \tag{15}$$

(again, in hundredths of a second). More generally, if we take N measurements of a quantity x, all using the same measuring equipment, and obtain the values $x_1, x_2, \ldots, x_N$, then the mean value of the x_i's (denoted $\bar{x}$, using an overbar) is given by the formula

$$\boxed{\bar{x} = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i.} \tag{16}$$

This concept of the mean (or average) of a set of data is probably not new to you, but note that it is *not* the same thing as the mean of a probability distribution, defined in eq. (4). The sample mean is the average of actual data; the distribution mean is the expected average over an infinite number of trials[3]. To try to keep this distinction clear, we'll use the notation $\langle x \rangle$ for distribution mean, but $\bar{x}$ for the mean of a particular sample. In general, angle brackets will always denote an expected value, rather than an average of empirical data. *Warning*: this notational distinction is not standard. Some texts use angle brackets for both EV and sample mean; some texts (including Taylor) use overbar for both.

Of course, we haven't *proven* that the mean is the best guess. Later on, when we discuss the maximum likelihood method, we can justify this claim. For now, though, it seems entirely uncontroversial, so we'll just accept it.

What about the third question? That is, how well do we know the drop time (if we take the mean value as our best estimate)? What we'd really like is some quantity that represents the amount of spread in the data. If all of the experimental values are tightly clustered around the mean, it seems reasonable to suggest that our knowledge of the "true" drop time is quite good; however, if the data points are generally very far away from the mean, then we can't be nearly as confident that our series of measurements has given us a great deal of precise information about the true drop time.

One way to look at how far the data points are from the mean is, of course, to simply take the difference between each data point and the mean, namely, $d_i = x_i - \bar{x}$ for each i. These are called the *deviations* from the mean. If the deviations are all very small, then we can say with confidence that our measurements

[3] We'll come back to this idea in section 2.3

were very precise; if some of them are large, we cannot be as confident. Here are the deviations for our stone drop experiment in table form (again, all times in hundredths of a second):

Trial	Time	Deviation
1	64	-5.2
2	69	-0.2
3	67	-2.2
4	75	5.8
5	71	1.8

Table 2: Data from the stone-dropping experiment, now shown with deviations from the mean.

Of course, instead of a column of numbers, we'd really like just *one* number to express the spread in the data. As we did when calculating the spread in a distribution, we'll take the *squares* of each of the deviations and then average them. Then we'll take the square root again to recover a number with the same units and scale as the original quantities. The number achieved in this way is called the *standard deviation* (SD) of the sample:

$$\boxed{\mathrm{SD} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(d_i)^2} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \langle x\rangle)^2}.} \tag{17}$$

You won't generally be called upon to calculate the numerical value of the SD for a data set by hand, but it helps to know what the definition is.[4] For our sample data set, we can calculate it:

Trial	Time	Deviation	Deviation squared
1	64	-5.2	27.04
2	69	-0.2	0.04
3	67	-2.2	4.84
4	75	5.8	33.64
5	71	1.8	3.24

Table 3: Data from the stone-dropping experiment, now shown with deviations and deviations squared.

The last column adds up to 68.8, so we divide by 4 and take the square root to get SD = 4.15 hundredths of a second. And indeed if you eyeball the original data set, it makes sense that a single number which represents the spread in the data should be on the order of 4 (as opposed to on the order of 1 or on the order of 20).

Again, there is a difference between the standard deviation of a sample (SD) and the standard deviation of the underlying distribution (σ_x). However, it will be harder to keep the notations distinct, because σ gets used for a lot of things (as we'll see).

So is the SD the answer to the question "how good is our best estimate"? Well, no, not exactly. But it's closely related—stay tuned and all will be revealed in section 3.6.

[4] You might be wondering: "What's with the $N-1$ in the denominator? If we're taking the mean of all N of the deviations, shouldn't it just be N?" The answer is basically that if you take N measurements, there are really only $N-1$ meaningful deviations. Think about the case of very small N: the formula is undefined for $N=1$, but that's as it should be. With a single measurement, you don't yet have any information about the spread. When you take two measurements, there's only one deviation (the other has to be equal and opposite); and so on. So that explains the $N-1$. In practice, we will usually be working with large values of N, and only need to know the SD to one significant figure, so it will not matter whether we use N or $N-1$. Some texts use the term *sample standard deviation* for definition (17) and *population standard deviation* for the analogous definition with N in the denominator.

2.2 Histograms

Suppose you are feeling extra-careful one day and make not 5, but 50 repeated measurements of some quantity x. (Perhaps x is the distance traveled by something, in centimeters, but the specifics are not important.) Your 50 measurements are as follows:

23	29	23	28	19	25	26	27	29	25
32	25	22	22	28	25	31	23	29	32
25	26	30	28	25	29	30	26	31	31
33	28	27	29	24	33	25	26	27	27
31	28	31	28	21	34	26	29	28	30

Table 4: Measured values of x.

You can calculate the mean (27.4) and SD (3.3) of this set of 50 values, but with such a large sample, these two numbers can't quite capture all of the relevant information in the data set. Here's where a graph can come to your rescue. Graphs and figures are often able to express quite a lot of information in a way that lots of words, or a table of numbers, can't readily convey. The particular type of graph that is most useful in this context is a *histogram.* A histogram is a special type of bar graph that shows how often each value of the measurement was observed to occur:

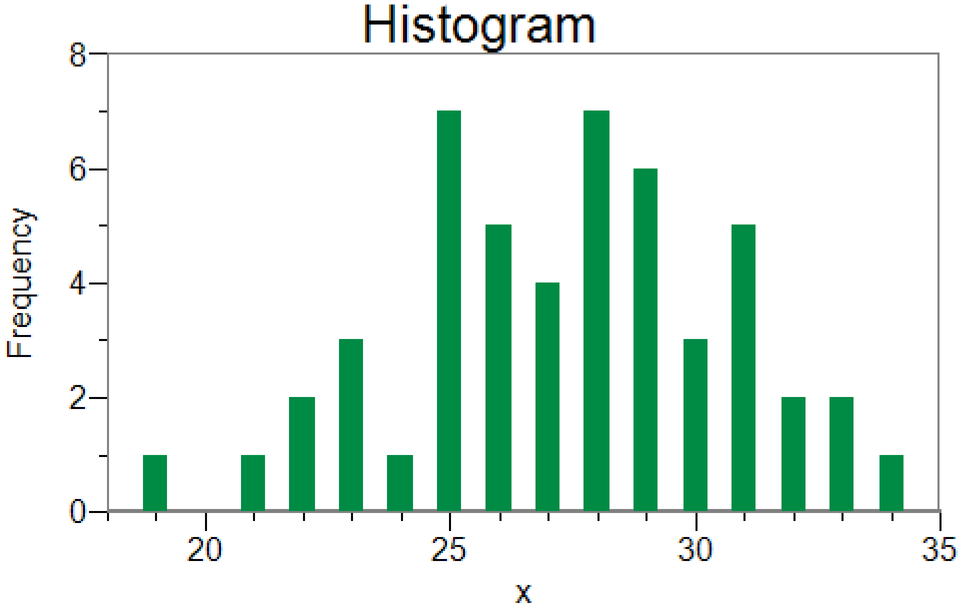

Figure 8: Histogram showing the frequencies of each measured value of x.

As you can see from Figure 8, the possible x-values are laid out on the horizontal axis and the observed frequency of each value is charted on the vertical axis. The result is very pretty. A histogram is a great way of showing immediately all of the important features of a set of data: where the mean value is, how clustered the values are around the mean, the overall shape of the distribution, and whether there are any outliers (observed values unusually far from the mean).

However, when making physical measurements you will often have data that isn't quite so tidy. For example, your data will often look like this:

Now your histogram isn't so pretty, as you can see from the left-hand graph in Figure 9.

22.9	29.1	23.3	28.3	19.1	25.5	25.8	27.4	28.7	25.1
32.1	25.4	22.3	21.7	28.4	25.0	30.8	23.2	28.9	31.7
25.5	26.2	30.0	27.8	24.8	28.8	30.2	25.9	31.0	30.7
32.6	28.4	27.2	29.1	23.7	33.0	25.4	25.9	27.1	26.8
31.0	27.8	30.6	27.6	21.0	34.3	25.9	28.6	28.4	30.0

Table 5: Measured values of x, with tenths digit.

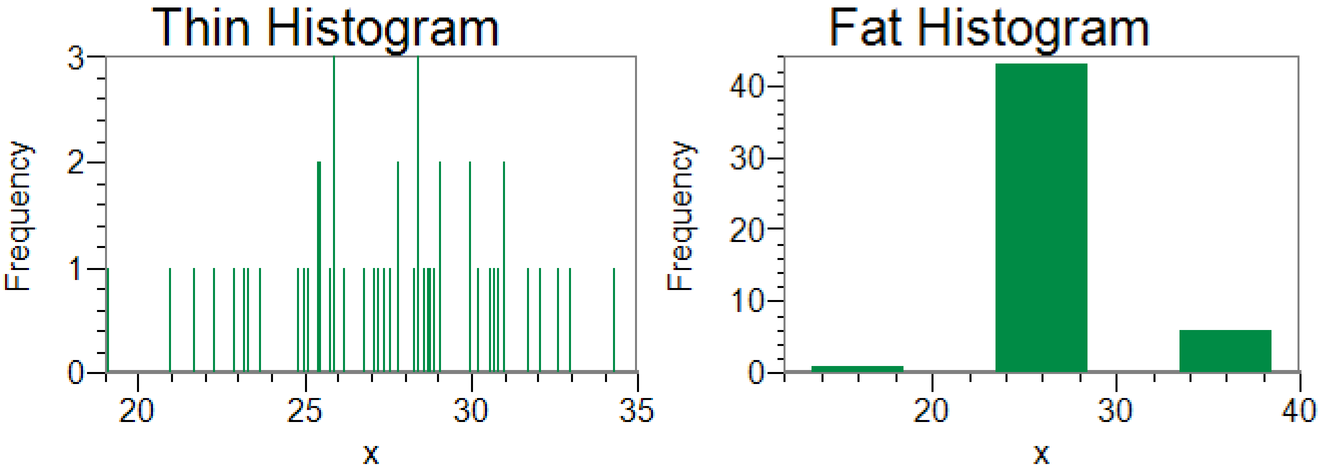

Figure 9: Two flawed histograms: the one on the left has bins that are too narrow, and the one on the right has bins that are too wide.

There are too many bars, almost all of which have a frequency of 0 or 1. The way to fix this is to group the data points into *bins*, which are intervals of constant width. Not all of the points in the same bin will be exactly the same value, but they'll be pretty close. Then we just count the number of measurements which fall into each bin and make our usual histogram. For example, if we use a bin width of 1.0, we'll recover the histogram from Figure 8.

Choosing the correct bin width for a histogram is a tricky procedure. You've seen what happens when you don't bin at all; essentially the same thing happens when you have bins but they are too narrow. At the right of Figure 9 you can see what happens when you make the bins too big: there are very few bins (in the stupidest case, only one bin, which contains the entire data set). Since the point of a histogram is to show both the *range* and *frequency* of the measured values, you should aim for somewhere in between these two extremes.[5] Also important to note is the fact that the appropriate bin width will depend on your sample size: the more data points you have, the smaller you can make the bins without making the frequencies too small. When you get down to fewer than about five or ten points, any histogram at all is of questionable value.

2.3 Distribution as the Limit of a Histogram

If you were *really* bored one day, you might decide to take more than 50 measurements. Perhaps more than 100... more than 1000... In fact, if you kept repeating the experiment over and over, you'd find that your histogram approached a definite shape. Figure 10 shows what happens if we increase the number of

[5] We might note that although both are undesirable, at least the graph on the left of Figure 9 has all of the information from the original data set. The graph on the right has obscured almost all of the information by taking unlike data points and lumping them together.

data points to 5000, and cut the bin width to 0.5. The distribution looks quite smooth and regular at this point.

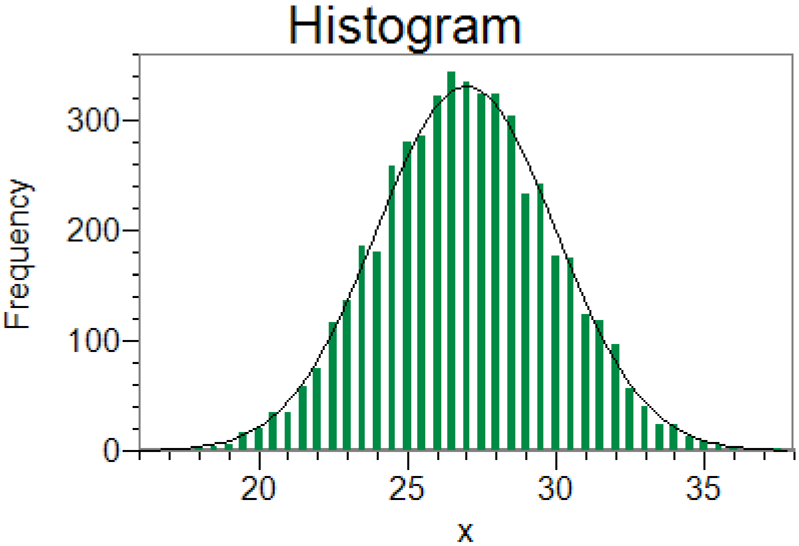

Figure 10: Histogram of 5000 trials of the same measurement of x.

As the number of measurements keeps growing, the distribution approaches a definite, continuous curve. This curve is called the *limiting distribution*. The limiting distribution for the x measurement has been drawn onto Figure 10 in black.

There are a couple of niggling problems that you can run into when taking the "limit" of a histogram as we keep taking more measurements. First, the vertical axis represents total counts, so its scale will keep changing as we increase the sample size (i.e. if you double the sample size, chances are you will double the height of every bin). This problem can be fixed by rescaling: instead of plotting raw count totals on the vertical axis, we divide by the number of measurements. The height of each bin is now interpreted as the *fraction* of measurements which fall into this particular bin, rather than the raw number of counts.

The second problem is more subtle. A big reason to take more measurements in the first place is so that you can make the bins narrower and thus get a "higher definition" histogram. But if you change the bin width, again the heights change; the narrower the bins, the fewer counts for each bin. For example, the fraction of measured lengths which are 27 cm might be 20% if you round lengths to the nearest cm. But if you round (or measure) to the nearest 0.1 cm, then the fraction which would now be classified as 27.0 might be only 2%. (Imagine chopping the 1-cm bin into 10 bins of 0.1 cm each; it is reasonable to suppose that 27.0, 27.1, 27.2, etc. are all roughly equally likely, so each one would get a tenth of the 20% we started with, or 2% of the total measurements.) If you specify to the nearest 0.01 cm, the fraction which are 27.00 would be only 0.2%, even if 27.00 is the single most likely outcome. Eventually, we'd like to let the bin widths become infinitesimally small, but then their heights would all go to zero.

The solution is this: the height of the bin alone cannot be interpreted as a meaningful fraction as the bin width goes to zero. Instead, it's the *area* of the bin (i.e., the height multiplied by the bin width) which represents the fraction of the total measurements which fall into that range. Interpreted in this way, the height can remain constant as we take the limit where the number of measurements increases without bound and the bin width simultaneously goes to zero. In this limit, the shape of the histogram approaches a definite curve, and this curve is called the *limiting distribution*.

Let's make this a bit more formal. If you remember your calculus class, you've probably noticed that the process we went through to get the limiting distribution—approximating the area under a curve by a progression of gradually skinnier rectangles—is exactly the process used to define a definite integral using a Riemann sum. Let's call the function representing the limiting distribution $f(x)$. Then if our bin width is Δx, the area of the bin starting at x is equal to its height $f(x)$ times its width Δx, and it represents the fraction of measurements falling between x and $x + \Delta x$. Taking the limit as $\Delta x \to 0$, we can replace Δx by the differential dx, giving us a fraction $f(x)\,dx$ falling between x and $x + dx$. Then if we want to know the total fraction in a larger interval, we just add the areas of successive bins until we have covered the entire interval. In the limit as $dx \to 0$, we get a definite integral:

$$\boxed{\begin{array}{l}\text{fraction of measurements}\\ \text{that fall between } a \text{ and } b\end{array} = \int_a^b f(x)\,dx.} \tag{18}$$

This is very nearly the same as eq. (1). At the time, we were talking about probability; Here we are talking about a fraction of a very large (hypothetical) data set, which turns out to be a very natural way of thinking about what probability means in the first place. If we knew $f(x)$ for all x, we would know essentially the results of an infinite number of previous measurements of x. Is that enough information to *predict* the outcome of the next measurement? Suitably interpreted, the answer is yes. However, we can't predict the actual *value* of the next measurement; what we can predict is the *probability* that it will fall between any two given values. So we can reinterpret the quantity "the fraction of past measurements that fall between a and b" as "the probability that the *next* measurement will fall between a and b." In other words, a limiting distribution is the same thing as a probability distribution.

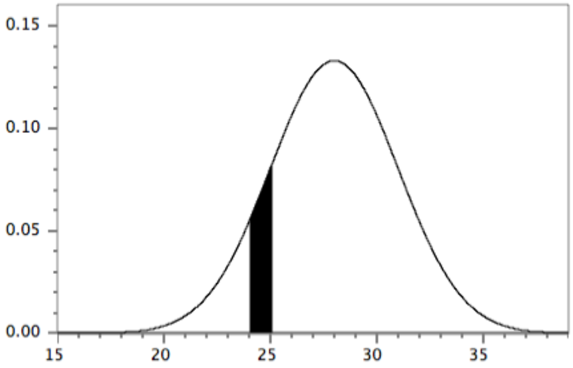

Figure 11: Limiting distribution of measurements of x.

Figure 11 shows the limiting distribution of measurements of x, along with a shaded region representing measurements between 24 and 25. The area of that shaded region (about 0.09) is the fraction of measurements that fall into that interval. So the graph tells us that about 9% of measurements collected were between 24.0 and 25.0. This is also what you would get if you took the definite integral of the distribution function (a Gaussian, in case you hadn't already guessed) between the limits of 24 and 25.

3 Error and Uncertainty

Now that we know quite a bit about probability distributions in general and the Gaussian distribution in particular, you might naturally be wondering why you should care. The answer is that probability distributions are intimitely related to the concept of *error*, and that the Gaussian is a special distribution which crops up with surprising frequency in experimental science. Let's investigate these ideas in a bit more detail.

3.1 What is Error?

First, what do we even mean when we talk about *error* in a lab setting? Let's begin to answer that by stating what we *don't* mean by "error": that the experimenter made a mistake. While that's always a possibility, the technical use of error in this course really refers to the inevitable uncertainty that accompanies any measurement. You cannot eliminate error simply by being careful. That's a key idea, so we'll put it in a big highlighted box:

Error refers to the inevitable uncertainty that accompanies any measurement.

Why is this uncertainty inevitable? The basic reason is because any real, physical measuring device has limitations on its precision. In a theoretical context, we can say things like "the stick is 75 centimeters long" and understand that to mean that the stick has a length of *exactly* 75 cm—no more, no less. But in an experimental setting, the point is to make measurements of physical quantities. We can measure the length of the stick with a ruler, but maybe the ruler is only marked to the nearest centimeter. "But," you object, "we can just get a better ruler." Fine—we'll get a ruler marked to the nearest millimeter. But that merely lessens the problem; it does not eliminate it. We can improve the measuring device as much as we want (there exist laser interferometers which can measure lengths to incredible precisions), but no matter what we do, we can never achieve infinite precision in a real experimental setting.

So instead, we do the next-best thing: figure out how much precision we *can* achieve, and keep track of it. If you understand errors properly, you can determine how much precision you need and how to obtain. That way, you can avoid the embarrassment of being error-ignorant. Imagine a researcher writing a grant proposal for a super-expensive laser interferometer to measure the length of a stick! Depending on what experiment is being done with the stick, it may be entirely sufficient to know its length to the nearest millimeter, or centimeter, or even "well, it's a bit less than a meter."

3.2 Truth? What is Truth?

Throughout the last section, we have been nonchalantly referring to a quantity we call "the length of the stick." But the entire point of the section was that there is no way to actually pin down *the* length of the stick; some amount of uncertainty in the length is absolutely inevitable. You can only make closer and closer approximations as you improve the quality of your measurement apparatus. So does it even make sense to talk about the stick as having a single, well-defined length?

The answer is a qualified "yes." In general, we make the assumption that any quantity we are measuring does have a "true" value, and that our measurements represent guesses as to what this true value might be. This view is largely validated by the results of countless real experiments, which have shown that if performed carefully, the results of a physical experiment are at least *repeatable*—that is, they give consistent results if performed again, by other experimenters in other locations.[6] So even though it is an assumption, it's quite probably a very good one, and one that we'll make henceforth without batting an eyelash. The stick does have a true length, and we can perform experiments to try to measure it.

[6] One caveat to this assumption is that if your measuring apparatus is really precise, you may run into problems of definition. For instance, if you did spring for that fancy-schmancy laser interferometer to measure the length of the stick, you'd find that upon closer examination, the ends of the stick are bumpy, rather than sharply cut; and the length of the stick changes slightly with temperature, and whether you hold it vertically or horizontally; and so on. It gets harder and harder to specify exactly what we mean by "the length of the stick" under these conditions. In these cases, the precision of the measurement can be limited by the definition of the quantity being measured, rather than by the resolution of the measuring apparatus.

Another important exception will arise when you study quantum mechanics, where you'll discover that it is technically impossible to make any measurement of a quantity in a physical system without fundamentally altering the system itself. But we won't have to worry about that for quite some time.

While we're on the subject of truth, we should point out that *an experiment is correct if it has been performed and analyzed correctly.* Even though we assume that there is a "true" value to a quantity we are measuring, the error in an experimentally measured quantity is *never* found by comparing the measured value to some number found in a book or other reference. If we tell you to measure the gravitational acceleration at the Earth's surface, and your result does not agree with the 9.8 m/s^2 that you are expecting to get, this discrepancy does *not* mean that your result is wrong. (Of course, it *could* be wrong—maybe you made a mistake somewhere. But you should not assume it is wrong merely because it is unexpected.)

3.3 Systematic and Random Error

There are two main classifications of error in measurement that we have to distinguish between: *systematic error* and *random error.* Figure 12 shows the difference between these two types of error.

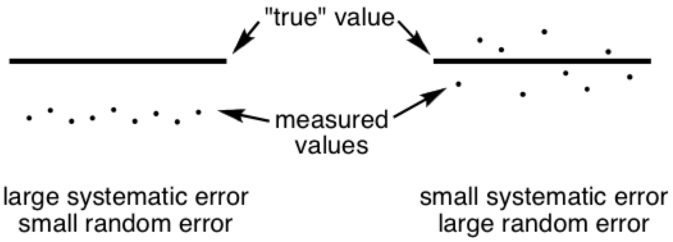

Figure 12: The difference between systematic and random error.

The set of values on the left exhibit a large systematic error: they are clustered around a point which is not the true value of the parameter being measured. (In particular, they are all too low.) The data graphed on the right do not have this problem; on average, they are neither too high nor too low. However, they exhibit more random error: they vary more from measurement to measurement than the data on the left.

The terms *precision* and *accuracy* are used to describe these two different properties. A set of measurements with very little systematic error is said to be accurate: its data points should, on average, be equal to the true value. A set of points with very little random error is said to be precise. We've already used the term *precision* quite a bit to describe the limitations of a measuring apparatus.

Incidentally, a set of measurements can be precise, but not accurate: an example is the data set on the left in Figure 12. However, it is usually incorrect to describe a set of measurements as accurate but not precise. If a data set has a great deal of scatter, but the average happens to be very close to the true value, that is generally an accident, rather than a repeatable experimental occurrence. By and large, the amount of scatter in a data set limits the extent to which we can claim that the result is accurate, as we discussed in Section 2.1.

In principle, you can eliminate or reduce systematic error in your measurements by being careful and using well-calibrated measuring devices. For instance, you may have a systematic error in measuring lengths if you measure them with a tape measure that has been stretched; getting a better tape measure would fix this problem. Or you might have a systematic error in measuring times with a stopwatch because there is a delay between when the event occurs and when you press the button to stop the timer. Accounting for, or eliminating, this delay will reduce the systematic error.

However, there is no way to eliminate random error from your measurements. Random errors are a fact of life, for the reasons we have discussed earlier. For the rest of this document, we will be discussing random error, not systematic error. While a great deal of effort goes into finding and eliminating systematic error, it mostly boils down to two things: calibration of instruments, and common sense. Other than that, there is not much that can be said about systematic errors *in general*. A good experimenter has to consider systematic errors in each experiment on a case-to-case basis.

Random errors, on the other hand, have a number of properties which are universal in character, and worthy of our attention. That's why we've already spent so much time talking about repeated measurements, histograms, and distribution functions, particularly the Gaussian. Let's see what useful conclusions we can draw from those discussions.

3.4 Quantitative Description of Random Error (or, How I Learned To Stop Worrying And Love The Gaussian)

Because measurements always involve some amount of random error, no measurement is exactly predictable. As we discussed in Section 2.3, the best we can do is predict not the particular outcome of a measurement, but the probability distribution of possible outcomes (and even the probability distribution is only known to us after we perform a very large number of measurements to get the limiting distribution of the histogram).

So in characterizing the random error associated with a measurement, we really need to make a statement about probability. If we make a statement like "the speed v is 1.52 m/s," what we mean is that if somebody were to repeat the experiment and measure v again, they should get something *close* to 1.52 m/s *most of the time*. So far, so good—but we want something more quantitative. How close is "close"? How most is "most"?

To answer these questions, we can use our knowledge of the probability distribution for v. If we provide the functional form of the distribution function $f(v)$, or just a detailed graph plotting $f(v)$ vs v, we can answer any quantitative question we want. How often will the measurement of v fall between 1.48 and 1.55 m/s? Simple—just evaluate the integral $\int_{1.48}^{1.55} f(v)\,dv$, or calculate the area under the $f(v)$ curve between 1.48 and 1.55 m/s. How often will we get a measurement larger that 1.60 m/s? Answer: $\int_{1.60}^{\infty} f(v)\,dv$. What's the largest speed which is still smaller than 90% of the expected measurements? This one is trickier, but it can still be answered using the distribution function.

It seems like this is exactly what we need to describe the random error: the probability distribution function. However, there are two disadvantages of this description:

1. It can be a bit unwieldy to have to include a distribution function every time we want to specify the error of a measurement.
2. We would need to perform an infinite number of repeated trials of the experiment before we know what the distribution function is in the first place.

#1 is really just an inconvenience; #2 is a deal-breaker. Fortunately, we can overcome both of these hurdles due to one ridiculously useful empirical fact:

Most measurements have a probability distribution which is approximately Gaussian.

This is a remarkable statement, but it turns out to be true. (No wonder we kept referring to that familiar-looking bell-shaped curve as being "ubiquitous"!) The reason why is related to a statistical result known as the *Central Limit Theorem*, which states that under normal conditions, the sum of several independent random variables—even variables that do not themselves obey a Gaussian distribution—produces a Gaussian distribution. Convergence to a Gaussian is very quick near the mean (you only need to add 5 or so random variables before it starts to look very much like a bell curve), but much slower on the tails of the distribution. Luckily, we are usually more interested in the behavior near the mean.

What does the Central Limit Theorem have to do with random error? Think of all of the things that go into the "noise" or randomness of a physical measurement: perhaps the reaction time of the experimenter, parallax issues (not viewing a needle scale exactly dead-on), air currents in the room, thermal fluctuations, etc. It is easy to imagine that you'd have 10 or more perturbative sources, each of which has a small random effect on the outcome of the measurement. The Central Limit Theorem says that the cumulative effect of all of these random sources is a Gaussian random variable, even if the individual sources are not Gaussian at all.

Perhaps it is clearer now why we spent so much time learning about the properties of the Gaussian back in Section 1.6. We can now take full advantage of that knowledge: rather than worry about the particulars of the distribution function, we can simply assume it is Gaussian[7] and then proceed from there. A Gaussian is completely described by only two parameters (its mean μ and standard deviation σ), so we need only specify those two parameters for our distribution.

Well, this is a vastly simpler problem to solve: what Gaussian (in terms of mean and standard deviation) is the best fit for our data set? If the data set happens to be large enough to construct a meaningful histogram, you should definitely do that, and then try to fit a Gaussian to the histogram (either eyeballing it, or by more advanced numerical techniques). This method has a lot going for it: you can see the overall shape of the distribution right away (including whether it looks Gaussian to begin with); the mean will probably jump right out at you from the peak location, if there is one; and you can see if there are any notable statistical outliers.

But let's say you can't make a histogram because you only have a handful of points in your data set. It turns out[8] that the best choice you can make for the Gaussian that describes the underlying limiting distribution of your data is the one with μ equal to your sample mean, and σ equal to your sample standard deviation. So now life is easy—just calculate the mean and standard deviation of your data set, and you can happily describe the distribution function using the corresponding Gaussian.

3.5 Confidence Intervals

Now that we can use Gaussians to describe the random error of a set of measurements, what has this bought us, really? Quite a lot, actually—Gaussians have nice well-defined properties like being symmetric and peaked in the middle and all that. However, there are two incredibly useful facts about Gaussians which we haven't yet mentioned:

1. There is about a 68% chance that a single measurement will fall within one standard deviation of the mean, i.e. between $\mu - \sigma$ and $\mu + \sigma$.
2. There is about a 95% chance that a single measurement will fall within two standard deviations of the mean (between $\mu - 2\sigma$ and $\mu + 2\sigma$).

The numbers may seem arbitrary to you, but this "rule of 68-95" is a good one to remember. This rule is illustrated in Figure 13. Mathematically, the equivalent statements are:

$$\int_{\mu-\sigma}^{\mu+\sigma} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}\, dx = 0.68 \tag{19}$$

$$\int_{\mu-2\sigma}^{\mu+2\sigma} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}\, dx = 0.95 \tag{20}$$

[7] Real experimental scientists don't do this, actually. They spend a lot of time worrying about possible "nongaussianity" (no joke, they really use that word) of their data. Of course, if you have the time, the best practice is to repeat the measurement enough times to be able to tell if it fits a Gaussian. But for the most part, we won't worry too much about nongaussianity. We'll be sure to warn you if we ask you to work with nongaussian data.

[8] Again, we can justify these claims later when we study maximum likelihood.

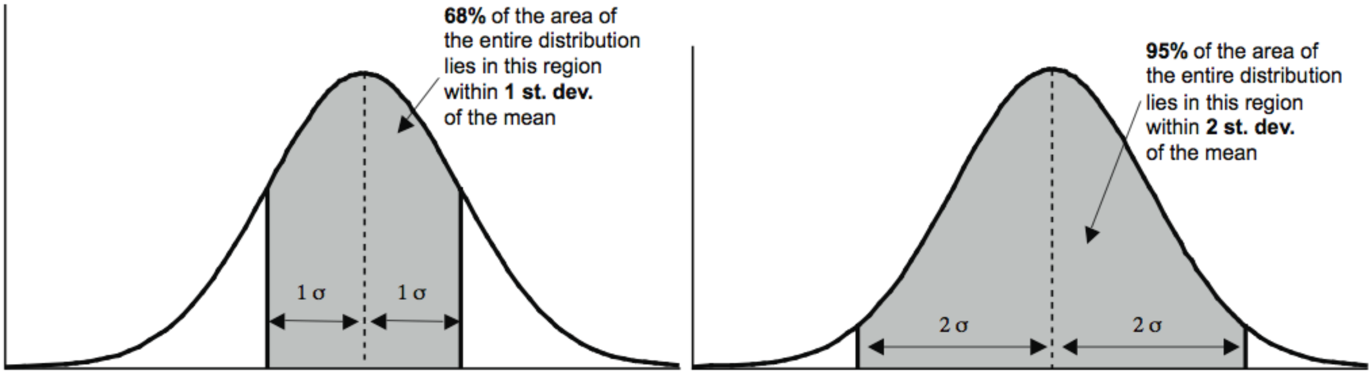

Figure 13: 68% of the area under the Gaussian curve falls within one σ of the mean (left); 95% falls within two σ of the mean (right).

Of course, we don't have to stop there; 99.7% of the area falls within 3σ of the mean, and so forth. (You could look up all these figures in a table of numerical integrals of the Gaussian distribution function.) But for practical purposes, 68% and 95% will suffice.[9] The key point is that these figures establish *confidence intervals*. If a set of measurements obeys a Gaussian distribution with mean 100 and standard deviation 1, then we can be 68% confident that any given measurement from that set lies between 99 and 101, and 95% confident that it lies between 98 and 102. In particular, we can predict that the *next* measurement made in the same way will also fall into those ranges with the same probabilities.

This is exactly what we wanted: a concise but specific quantitative description of the random error associated with a measurement. It works so well that we'll use it to define what we mean by *uncertainty* going forward. Since one standard deviation turns out to be such a convenient quantity to work with, we'll use that:

The *uncertainty* in a reported value is defined as the interval in which we are 68% confident the true value lie.

From this definition, for a single measurement taken from a data set, the uncertainty is equal to the standard deviation of the data set. Rather than specifying the endpoints of the interval, we often express uncertainties as the central value "plus or minus" a certain amount. So in our example, rather than specifying that the measurement is between 99 and 101 about 68% of the time, we can say that the measurement is 100 with an uncertainty of 1, or write it even more concisely as 100 ± 1.

A bit of notation: if a measured quantity is represented by x, then the uncertainty in x is sometimes written as δx. δ is the Greek lowercase letter delta, and in this context it means "uncertainty of". So if you measure some length to be $L = 100 \pm 1$ cm, then $L = 100$ cm and $\delta L = 1$ cm.

3.6 Standard Deviation and Standard Error

Even with this definition of uncertainty, we can still run into some confusion when it comes to repeated measurements. Consider the example we saw back in Section 2.3, when we took the set of 50 measurements of a length (Figure 8) and expanded it to a set including 5000 data points (Figure 10). In our original data

[9] Actually, there is a good reason we don't bother with 4σ and beyond, and rarely with 3σ: Treating almost every distribution that comes up as a Gaussian is an approximation. For most distributions, it's very good approximation near the center of the distribution, and not so good way off on the tails. This agrees with common sense—the Gaussian is infinitely wide in the sense that it never goes to zero, but of course there are physical measurements where you will obviously never get a negative value, for example.

set, the mean was about 27 cm and the standard deviation was about 3 cm; in the expanded data set, the mean was again about 27 cm and the standard deviation was still 3 cm. Both data sets support the same conclusion: if you did the experiment exactly one more time and measured the length, you'd have a 68% chance of measuring between 24 and 30 cm (and a 95% chance of measuring between 21 and 33 cm, and so on). Another way of saying this is that a single measurement of the length gives 27 cm, with an uncertainty of 3 cm associated with it.

But we didn't do the experiment one time; we did it many times, and chose the mean of all of the measurements as our best estimate of the true length. You might think that the mean, as a judicious combination of all of our measurements, would have a smaller uncertainty than any single individual measurement—and you'd be right. It turns out (and we can show why, later on) that the uncertainty in the final answer $x_{\text{best}} = \bar{x}$ for a set consisting of N repeated measurements is given by

$$\boxed{\delta\bar{x} = \frac{\text{SD}}{\sqrt{N}}} \tag{21}$$

where SD is the standard deviation of the N measurements. This quantity, $\text{SD}/\sqrt{N}$, is called the *standard error of the mean* or just the standard error (SE). (Confusingly enough, some texts also call it the "standard deviation of the mean," but we won't use that term.)

It is very important to understand the distinction between standard deviation and standard error. As you take more and more measurements of the same quantity, you would not expect the standard deviation SD to change appreciably, in either direction. On the other hand, the standard error SE would slowly decrease as you increased N. Taking the mean of a larger sample of measurements should make the final answer more reliable, and that $\sqrt{N}$ in the denominator indicates that it does.

However, the standard error doesn't change very *quickly* as we keep increasing N. Even when we went from $N = 50$ to $N = 5000$, we only improved the uncertainty of the mean by a factor of 10. That's a pretty small return on an extra 4950 repeated measurements. (And remember that we are also for the moment neglecting systematic errors, which are *not* reduced by increasing N.) So in practice, it's good to repeat your measurements several times, but if you really want to significantly increase your precision, you would do better to improve your technique or your measuring apparatus than to mindlessly rely on more repetitions.

3.7 Rounding

If you'll indulge me, an old joke I like:

> A little boy and his mother walk into the natural history museum and see a huge dinosaur skeleton in the lobby.
>
> "Wow!" says the boy. "I wonder how old those bones are?"
>
> A nearby security guard answers, "That skeleton is 90,000,006 years old!"
>
> "90 million *and six*? Are you sure?" asks the boy's mother.
>
> "'Course I am," replies the security guard. "I asked the curator how old it was when I started working here, and he said 90 million years. And I've been working here for six years."

Okay, it's a silly joke, but it gives us pause for thought. What's wrong with the security guard's answer? The uncertainty in the age of the skeleton is on the order of millions or tens of millions of years. So it's preposterous to make any claims as to the ones digit; the extra six years don't matter in the least. Even if we don't necessarily quote uncertainties when tossing numbers around in everyday conversation, it's still absurd to make such a huge error in the number of significant figures.

Arguably, it's even more absurd to make such an error when you are, in fact, quoting the uncertainty along with the value. For example, on in the past, students have made claims such as:

"We measured a value of 1.4969461 ± 0.27777777."

If the uncertainty in the value was about 0.28, why would you claim to know the result to seven decimal places? Just because that's what your calculator said when you plugged in the numbers? Don't believe everything that your calculator tells you! Keep in mind what the numbers mean—and what they don't mean. In this case, no matter what your calculator says, you know full well that you have no knowledge of even the hundredths digit of the answer, and only partial knowledge of the tenths digit.

Here's a rule of thumb you can rely on: round the uncertainty to *one* significant figure. Then round the answer to match the decimal place of the uncertainty. In our example above, we would just say that the uncertainty is 0.3, and the actual value would be rounded to the nearest tenth in accordance with the uncertainty. So the final result would be:

"We measured a value of 1.5 ± 0.3."

If the uncertainty were 0.0027777, we'd call it 0.003 and the result would be 1.497 ± 0.003. The small uncertainty gives us confidence that the thousandths digit is meaningful.

One exception to the rule of thumb: If rounding the uncertainty to one significant figure would cause that figure to be a 1, then you keep the next digit as well. So for instance, if you have $5.83333333 \pm 0.14285714$, then you would round the uncertainty to 0.14 (instead of just 0.1) and report 5.83 ± 0.14 (instead of 5.8 ± 0.1). But for the most part, knowing the uncertainty to one sig fig is just fine.

3.8 Reading Error

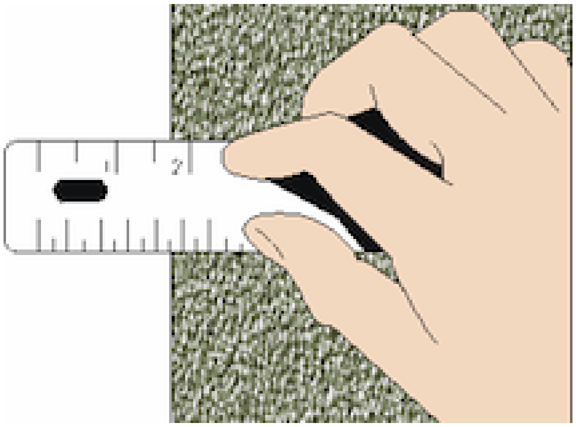

Figure 14: Reading the position from a ruler

We've seen that the best way to determine the uncertainty of a measurement is to repeat it. But repeating a measurement is not the *only* way to determine the uncertainty. In many cases, you can get a very reasonable estimate of the uncertainty with just a single measurement, just by using common sense. For example, suppose you're using this ruler to measure the position of the edge of the surface shown in Figure 14 (figure borrowed from Harrison).

It looks like it's around 1.75 inches. But how do we get the uncertainty? On a basic level, we can just estimate the minimum and maximum values the reading could have without us noticing the difference.

Here's how I would do it: the upper bound seems to be 1.75"; the lower bound is trickier, but I'd guess it's about 1.70". (The two ruler marks on either side are 1.50" and 2.00"; it's pretty clearly more than 40% but less than 50% of the way from 1.50 to 2.00.) So I might say that the position is 1.72 ± 0.03 inches; the uncertainty of 0.03" is called the reading error of the measurement.

Now, another measurer might well disagree with my estimate of the reading error, but probably not by a lot. Determination of reading error can be a little crude, but the important thing is to get a reasonable estimate of the uncertainty without having to do the measurement many times. In the case of measuring a static length with a ruler, it seems a little ridiculululous to imagine repeating the measurement. After all, we only need the uncertainty to one sig fig anyway.

Not all reading errors are created equal. For example, using the same ruler to measure the position of a rough edge (rather than a smooth one) would give a larger reading error. Alternatively, using a ruler with more closely spaced markings might well give a smaller reading error. Ultimately it always comes down to your judgment and common sense.

Once you have estimated the reading error on a measurement, you can treat it just like a statistical uncertainty, as if the reading error were a standard deviation of repeated measurements. That may be a stretch, but estimating reading error is a quick and dirty approach to uncertainty that produces generally reasonable results. The key thing to remember is that most of the time, we don't need to know the uncertainty very precisely; a reasonable estimate will often do just fine.

3.9 Fractional Uncertainty

If you have some measured quantity x with an uncertainty of δx, there are a few different ways to talk about the uncertainty. The one we've been using is just to use the absolute uncertainty: if $\delta x = 1$ inch, then x is known to plus or minus an inch. But this might be considered to be a very large or very small uncertainty depending on the value of x itself. If $x = 2$ in, you haven't made a very precise measurement; if $x = 1$ mi, you've made an incredibly precise measurement. So another way to discuss uncertainty is to use the *fractional uncertainty*, defined as the uncertainty divided by (the best-guess value of) the quantity x itself:

$$\boxed{\text{fractional uncertainty} \;=\frac{\delta x}{|x|}} \tag{22}$$

The fractional uncertainty is sometimes called the *relative uncertainty*; δx itself is called the *absolute uncertainty* in cases where a distinction needs to be made with fractional uncertainty. Note that fractional uncertainty is always dimensionless, whereas absolute uncertainty has the same dimensions and units as the quantity being measured. For example, if $L = 100 \pm 2$ cm, then the absolute uncertainty δL is 2 cm, but the fractional uncertainty $\delta L/L$ is 0.02, with no units or dimensions.

Most of the time, the fractional uncertainty of a quantity that you measure (and calculate) is small compared to 1. Sometimes it's more convenient to express it as a percent uncertainty (e.g. 2% instead of 0.02).

3.10 Significant Figures

The concept of fractional uncertainty is closely tied to the idea of significant figures. When a quantity has a fractional uncertainty of greater than about 0.1 (10%), it should probably only be reported to one sig fig, because only one digit of the answer carries significance. When the uncertainty is on the order of 1%-10%, there are two significant figures, and so on.

Note that the term *significant figure* has a slightly different meaning in a scientific context than in a mathematical one. A quantity that has been measured to be $x = 32 \pm 3$ can be said to have two significant figures, even though a mathematician would reserve such terminology for a number that is known to be between 31.5 and 32.5 (i.e. a number that is equal to 32 after rounding to two digits). We're not mathematicians, so we'll use "significant" to imply that we know something about a figure, not that we are 100% sure what its value is.

4 Fitting

One of the most common analysis techniques in science is to compare experimental data to a model by analyzing the relationship between two variables. (In a high-school science class, you may have seen these referred to as the indepedent and dependent variables, but we'll just call them x and y.) Suppose your model predicts that there is some specific functional relationship $y = f(x)$ between the two variables; very commonly, y will be a linear function of x. If you then make some measurements of y_i for different values of x_i, how can you use your data to test the model?

4.1 Rejection of a Model

To answer this question, let's start with a related question (although on the surface, it may not be evident how this is related): you take N measurements of some physical quantity x. You hypothesize that the data sample is drawn from an underlying Gaussian distribution with mean μ and standard deviation σ. How can you use the data to test your hypothesis?

The statistical test used for this purpose is called the *chi-squared test*. For a single measurement x, hypothesized to be sampled from a Gaussian with mean μ and SD σ, χ^2 (chi squared) is defined as

$$\chi^2 = \frac{(x-\mu)^2}{\sigma^2}. \tag{23}$$

Essentially, χ^2 is a measure of how far away x is from its expected value μ, normalized to the scale of the SD. When x is exactly what you expect, $\chi^2 = 0$; when it's 1σ away, $\chi^2 = 1$; if x is 2σ away, $\chi^2 = 4$, and so on.

This definition is easily generalized to a set of many measurements. If you take a sample of size N, we just add up the individual χ^2 values to get the total χ^2:

$$\boxed{\chi^2 \equiv \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{\sigma^2}.} \tag{24}$$

If the N measurements were all truly drawn from the hypothesized Gaussian, you would expect that most of the x-values are within 1 SD of the mean, so that each individual χ^2 is on the order of 1, and therefore the total χ^2 is on the order of N. That's the gist of the chi-squared test. If the total χ^2 is significantly greater than N, you have good reason to believe that the x-values were *not*, in fact, sampled from your hypothetical Gaussian. There are more rigorous numerical tests that you can perform on the specific χ^2 values for a given sample size N, but we won't be too concerned about those just yet.

What if χ^2 is much *less* than N? That should also be cause for concern. If your data is truly sampled from an underlying Gaussian with standard deviation σ, it is very unlikely that each point would be much less than 1σ from the mean, on average. The most likely explanation is that your hypothesized value for the standard deviation σ is much too large.

Getting back to our original question, how do we use the chi-squared test to assess the relationship between $y = f(x)$ predicted by our model? The basic idea is that for each measured x_i, you can calculate the corresponding $y_{\text{model},i}$ that the model would predict for it, and then compare those to the actually measured y_i's using the chi-squared test. (For now, let's assume that there are uncertainties σ_{y_i} associated with each measured y_i, but no uncertainties in the x_i.) Then χ^2 would be equal to

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - f(x_i))^2}{\sigma_{y_i}^2}. \tag{25}$$

If the total χ^2 is on the order of N or a litlte less, the data looks like a reasonable fit for the model. If $\chi^2 \gg N$, the data should cause you to reject your model.

4.2 Principle of Maximum Likelihood

It will often be the case that the functional relationship $y = f(x)$ predicted by a model will have one or more adjustable parameters. The most common example is that the model predicts a linear relationship, $f(x) = Ax + B$, but the particular values of the slope A and intercept B may not be predicted by the model. As a simple example, if your model is constant velocity motion in one dimension, then the model predicts that x is a linear function of t.

In such a case, you can't calculate χ^2 for the data vis-à-vis the model because there *isn't* a model, per se; instead what you have is a whole family of models, one model for each value of A and B. To assess this family of models, what we do is first pick out the *best* model from the family, and then apply the χ^2 test to it.

How do we know what the best model is? We follow something called the *principle of maximum likelihood*, which is best illustrated by example. Let's consider a simple one: Recall our claim, back in section 2.1, that the best estimate of a distribution mean that you can make from a data sample is the sample mean $\bar{x} = \sum x_i/N$. It seems intuitive, but why is this true?

Let's begin by supposing that we happen to already *know* that the true underlying distribution is a Gaussian with mean μ and SD σ. Then the probability of observing a particular value x_i of the variable is

$$Prob\,(x\,\text{between}\,x_i\,\text{and}\,x_i + dx) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\,dx. \tag{26}$$

Here we don't really care about the interval width dx or the factor $\sigma\sqrt{2\pi}$. (We'll see in a minute *why* we don't care.) So we can shorten this to

$$Prob(x_i) \propto \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right). \tag{27}$$

That's the probability of finding a particular single value x_i; for the entire data set of N independent measurements, the probability of observing them all, given the assumed underlying Gaussian, is the *product* of the individual probabilities. So the probability of observing all of the known values $x_1, x_2, \ldots, x_N$ is

$$Prob(x_1, x_2, \ldots, x_N) \propto \exp\left(-\frac{\sum(x_i-\mu)^2}{2\sigma^2}\right). \tag{28}$$

Using the definition of χ^2 that we learned in the previous section, we can simplify this to

$$Prob(x_1, x_2, \ldots, x_N) \propto \exp\left(-\frac{1}{2}\chi^2\right). \tag{29}$$

This gives the probability of measuring the observed data, given some μ and σ. But if you think about it, this is a strange thing to be calculating. We already *know* the data points x_i were, in fact, observed. So why should we calculate the probability of something happening that we know already happened?

The key is to recall that we started by *assuming* that we knew the underlying μ (and σ). In fact, for any given model (in other words, for any given μ and σ) that you choose, eq. (29) can alternately be interpreted as the *likelihood* of the model, given the data. That makes more sense than thinking about it as the probability of the data, assuming the model.

With this definition of likelihood, we can now state the *principle of maximum likelihood*:

> The best model is the one that, if it is indeed true, is most likely to have produced the actual data. (30)

The application of the principle of maximum likelihood to determine the best values of unknown parameters is called *maximum-likelihood estimation* (MLE).

Getting back to the task at hand: to find the best values of μ and σ, we want the ones that will maximize the likelihood, which means we want to *minimize* the χ^2 in the exponent. To do this, we just differentiate it with respect to μ and set the derivative to 0, giving

$$0 = \sum_{i=1}^{N} (x_i - \mu) = \sum_{i=1}^{N} x_i - N\mu, \tag{31}$$

or in other words, $\mu = \bar{x}$. This is not surprising, of course—we've been using this as our best guess at μ all along—but it's nice to know that there is a sound reason for it.

The proof that the sample SD is the best guess for the distribution σ is more complicated, because the total likelihood is a more complicated function of σ (and there are a few other subtleties that come in). But the idea is the same, and indeed it is true that the sample SD is the result of applying MLE to the problem of guessing the underlying σ for a given set of repeated measurements.

4.3 MLE and Best-fit Methods

Now we can return to the problem we were considering before, where instead of repeated measurements of the same value, we were trying to use data over a range of x- and y-values to test a model relationship $y = f(x)$. Using MLE to approach the problem, we can write down the likelihood of finding the actual data (x_i, y_i) given the hypothesized model $y = f(x)$.

Let's remind ourselves about the straightforward procedure: Assume, for now, that the x_i values have negligible uncertainty. Based on each of those x_i values, the model predicts a corresponding $y_{\mathrm{model},i} = f(x_i)$. If the actually measured y_i does not exactly agree, attribute it to random error: the y-value is a sampling from an underlying Gaussian with mean $y_{\mathrm{model},i}$ and standard deviation σ_{y_i}. Then the likelihood of the model given the entire data set (x_i, y_i) is

$$\begin{aligned} \text{Likelihood} &= \exp\left(- \sum_{i=1}^{N} \frac{(y_i - f(x_i))^2}{2\sigma_{y_i}^2} \right) \\ &= \exp\left(-\chi^2/2\right), \end{aligned} \tag{32}$$

using the definition of χ^2. So for the question of picking the best model from a family of models, the maximum likelihood corresponds to the *minimum* value of χ^2.

It is important to note that this interpretation can be used in addition to, rather than in place of, the χ^2 test of whether a model should be rejected outright. To summarize, the procedure goes like this: take the family of models and parameterize it using one or parameters $A, B, \ldots$, so that each choice of parameters corresponds to a specific model. Calculate which choice of coefficients minimizes χ^2. That's the best-fit model from the family. In addition, look at the actual minimum value of χ^2. If the best-fit model has a χ^2 significantly larger than N, then the *entire family of models* is conclusively rejected by the data. In other words: when the best fit is a bad fit, every fit is a bad fit.

If, on the other hand, the best-fit χ^2 is N or slightly below, then you have a model that is consistent with the data. Let's say you calculate best-fit values of some parameters A and B (the slope and intercept of a best-fit line, say). One important question is: what is the *uncertainty* of the values A and B calculated from this procedure?

To answer this question, let's return to the interpretation of $-\chi^2/2$ as the exponential in a Gaussian probability distribution. Recall that in a simple Gaussian distribution, 68% of the probability (or area under the curve) lies within 1σ of the mean. For a single-variable Gaussian distribution, the exponent

$$-\frac{(x-\mu)^2}{2\sigma^2} \tag{33}$$

ranges from $-1/2$ to 0 to $-1/2$ as x goes from $\mu-\sigma$ to μ to $\mu+\sigma$. Similarly, 95% of the probability lies between $x=\mu-2\sigma$ and $x=\mu+2\sigma$, as the exponent goes from -2 to 0 to -2.

Back to χ^2, then: if the likelihood of a model given some data is $\exp\left(-\chi^2/2\right)$, then the 68% confidence interval corresponds to the range of parameters around the best-fit values that produce a χ^2 within 1 of the minimum (so that the exponent will be within 1/2 of the minimum). Likewise, the 95% confidence interval includes the region of parameter space that produces a χ^2 within 4 of the minimum.

This relationship may be seen more clearly by looking at a graph. For a functional form $f(x)$ which is linear in its parameters (such as $f(x)=Ax+B$, which is linear in both A and B), χ^2 is a quadratic function of the parameters.[10] So plotting χ^2 vs. any parameter near the best-fit value will look like this:

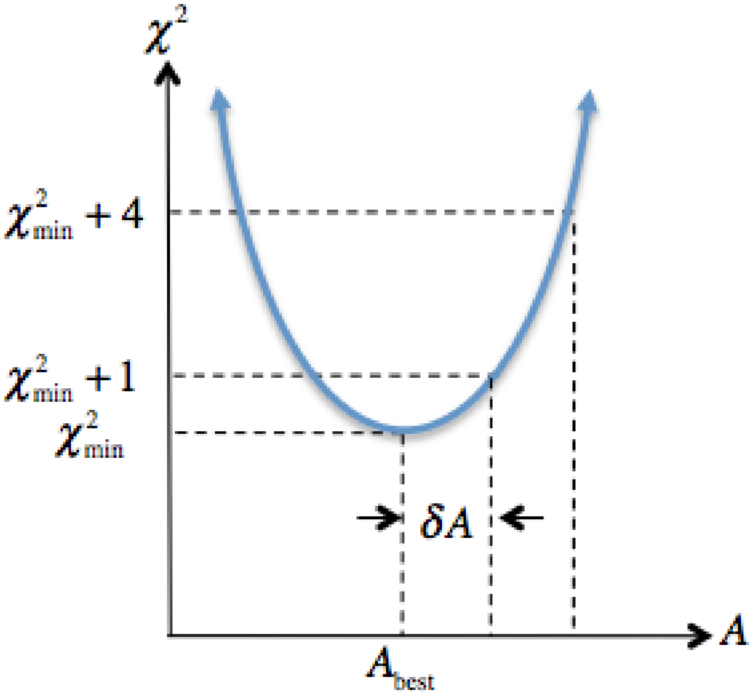

Figure 15: χ^2 vs. some parameter A near the best fit

If you stray far enough from the minimum to reach $\chi^2=\chi^2_{\min}+1$, then you are "one sigma" away from the best fit, or in other words, the parameter A has changed by δA from its best value. When you get to $\chi^2=\chi^2_{\min}+4$, you are $2(\delta A)$ away. Likewise, 9 over the minimum χ^2 corresponds to $A\pm 3(\delta A)$.

Putting it all together, we can use χ^2 to determine the best-fit value of a parameter (or parameters) thus:

1. Calculate the parameter values that produce the minimum value of χ^2.
2. If this minimum value is significantly larger than the total number of data points in your fit, then the fit is bad. You can rule out all models in the family of your model.

[10] If you write out the expression for χ^2 in terms of A, you'll see that although it's complicated, the only operations you have to do are squaring terms which are linear in A, and adding.

3. Otherwise, report the parameter values corresponding to χ^2_{min} as the best-fit values. For uncertainties, report the range of parameters around the best fit corresponding to a χ no greater than $\chi^2_{\text{min}} + 1$.

For a one-parameter family of models (e.g. $y = Ax$, for various different values of the slope A), step 3 above can easily be done by examining the parabola in figure (15). For a two-parameter family, χ^2 will be a function of A and B. It will still depend quadratically on those parameters, but the graph of the function won't be a parabola—it'll be a three-dimensional paraboloid.

5 Error Propagation

Suppose you measure some quantities $a, b, c, \ldots$ with uncertainties $\delta a, \delta b, \delta c, \ldots$. Now you want to calculate some other quantity Q which depends on a and b and so forth. What is the uncertainty in Q? The answer can get a little complicated, but it should be no surprise that the uncertainties $\delta a, \delta b$, etc. "propagate" to the uncertainty of Q. Here are some rules which you will occasionally need; all of them assume that the quantities a, b, etc. have errors which are *uncorrelated* and *random.* (These rules can all be derived from the Gaussian equation for normally-distributed errors, but you are not expected to be able to derive them, merely to be able to use them.)

5.1 Addition or Subtraction

If Q is some combination of sums and differences, i.e.

$$Q = a + b + \cdots + c - (x + y + \cdots + z), \tag{34}$$

then

$$\boxed{\delta Q = \sqrt{(\delta a)^2 + (\delta b)^2 + \cdots + (\delta c)^2 + (\delta x)^2 + (\delta y)^2 + \cdots + (\delta z)^2}.} \tag{35}$$

In words, this means that the uncertainties add *in quadrature* (that's the fancy math word for the square root of the sum of squares). In particular, if $Q = a + b$ or $a - b$, then

$$\boxed{\delta Q = \sqrt{(\delta a)^2 + (\delta b)^2}.} \tag{36}$$

Example: suppose you measure the height H of a door and get 2.00 ± 0.03 m. This means that $H = 2.00$ m and $\delta H = 0.03$ m. The door has a knob which is a height $h = 0.88 \pm 0.04$ m from the bottom of the door. Then the distance from the doorknob to the top of the door is $Q = H - h = 1.12$ m. What is the uncertainty in Q? Using equation (36),

$$\begin{aligned}
\delta Q &= \sqrt{(\delta H)^2 + (\delta h)^2} &(37)\\
&= \sqrt{(0.03 \text{ m})^2 + (0.04 \text{ m})^2} &(38)\\
&= \sqrt{0.0009 \text{ m}^2 + 0.0016 \text{ m}^2} &(39)\\
&= \sqrt{0.0025 \text{ m}^2} = 0.05 \text{ m}. &(40)
\end{aligned}$$

So $Q = 1.12 \pm 0.05$ m.

You might reasonably wonder, "Why isn't δQ just equal to $\delta a + \delta b$? After all, if I add $a \pm \delta a$ to $b \pm \delta b$, the answer is definitely at least $a + b - (\delta a + \delta b)$ and at most $a + b + (\delta a + \delta b)$, right?" The answer has to do with the probabalistic nature of the uncertainties δa and δb; remember, they represent 68% confidence intervals. So 32% of the time, the true value of a is outside of the range bounded by $a \pm \delta a$; likewise for b.

But how often is the sum outside of the range bounded by $a + b \pm (\delta a + \delta b)$? A little bit less often than that. In order for it to be higher than $a + b + (\delta a + \delta b)$, for instance, you'd need either both a and b to be on the high end of the expected range (or one of them to be very high), which is less likely than one being high and the other being low. (Remember that this formula assumes that the uncertainties in a and b are *uncorrelated* with each other.) So $\delta a + \delta b$ is actually a slight *overestimate* of the uncertainty in $a + b$. If you were to go through the math in detail, you'd arrive at the conclusion that the expected uncertainty is given by equation (36), rather than by the simpler expression $\delta a + \delta b$.

There is good news, though. The more complicated expression in equation (36) has a very nice feature: it puts more weight on the larger uncertainty. In particular, when one of the uncertainties is significantly greater than the other, the more certain quantity contributes essentially nothing to the uncertainty of the sum. For instance, if $\delta a = 5$ cm and $\delta b = 1$ cm, then equation (36) gives

$$\begin{aligned}\delta(a+b) &= \sqrt{(\delta a)^2 + (\delta b)^2} && (41)\\ &= \sqrt{(5\text{ cm})^2 + (1\text{ cm})^2} && (42)\\ &= \sqrt{25\text{ cm}^2 + 1\text{ cm}^2} && (43)\\ &= 5.1\text{ cm}. && (44)\end{aligned}$$

Since we generally round uncertainties to one significant figure anyway, 5.1 isn't noticeably different from 5. So the 1 cm uncertainty in b didn't end up mattering in our final answer. As a general rule of thumb, when you are adding two uncertain quantities and one uncertainty is more than twice as big as the other, you can just use the larger uncertainty as the uncertainty of the sum, and neglect the smaller uncertainty entirely. (However, if you are adding more than two quantities together, you probably shouldn't neglect the smaller uncertainties unless they are at most 1/3 as big as the largest uncertainty.)

As a special case of this, if you add a quantity with an uncertainty to an *exact* number, the uncertainty in the sum is just equal to the uncertainty in the original uncertain quantity.

5.2 Multiplication or Division

If Q is a product or quotient, such as

$$Q = \frac{ab\cdots c}{xy\cdots z}, \tag{45}$$

then

$$\boxed{\frac{\delta Q}{|Q|} = \sqrt{\left(\frac{\delta a}{a}\right)^2 + \left(\frac{\delta b}{b}\right)^2 + \cdots + \left(\frac{\delta c}{c}\right)^2 + \left(\frac{\delta x}{x}\right)^2 + \left(\frac{\delta y}{y}\right)^2 + \cdots + \left(\frac{\delta z}{z}\right)^2}.} \tag{46}$$

What this means is that the *fractional uncertainties* add in quadrature to give the fractional uncertainty of the product. (If you want the absolute uncertainty at the end, don't forget to multiply by Q itself.) In practice, it is usually simplest to convert all of the uncertainties into *percentages* before applying the formula.

One caveat here: in order for this equation to hold, the fractional uncertainties on a, b, etc. *must be small compared with 1.* Uncertainties of up to 10% or 15% are fine, but not 50%.

Example: a bird flies a distance $d = 120 \pm 3$ m during a time $t = 20.0 \pm 1.2$ s. The average speed of the bird is $v = d/t = 6$ m/s. What is the uncertainty of v?

$$\begin{aligned}
\frac{\delta v}{v} &= \sqrt{\left(\frac{\delta d}{d}\right)^2 + \left(\frac{\delta t}{t}\right)^2} && (47)\\
&= \sqrt{\left(\frac{3\text{ m}}{120\text{ m}}\right)^2 + \left(\frac{1.2\text{ s}}{20.0\text{ s}}\right)^2} && (48)\\
&= \sqrt{(2.5\%)^2 + (6\%)^2} && (49)\\
&= \sqrt{0.000625 + 0.0036} = 6.5\%. && (50)
\end{aligned}$$

So

$$\begin{aligned}
\delta v &= v(6.5\%) && (51)\\
&= (6\text{ m/s})(6.5\%) && (52)\\
&= 0.39\text{ m/s}. && (53)
\end{aligned}$$

So the speed of the bird is $v = 6.0 \pm 0.4$ m/s. Note that as we saw with addition, the formula becomes much simpler if one of the fractional uncertainties is significantly larger than the other. At the point when we noticed that t was 6% uncertain and d was only 2.5% uncertain, we could have just used 6% for the final uncertainty and gotten the same final result (0.36 m/s, which also rounds to 0.4).

The special case of multiplication or division by an exact number is easy to handle: since the exact number has 0% uncertainty, the final product or quotient has the same percent uncertainty as the original number. For example, if you measure the diameter of a sphere to be $d = 1.00 \pm 0.08$ cm, then the fractional uncertainty in d is 8%. Now suppose you want to know the uncertainty in the radius. The radius is just $r = d/2 = 0.50$ cm. Then the fractional uncertainty in r is also 8%. 8% of 0.50 is 0.04, so $r = 0.50 \pm 0.04$ cm.

5.3 Raising to a Power

If n is an exact number and $Q = x^n$, then

$$\delta Q = |n| x^{n-1} \delta x, \qquad (54)$$

or equivalently,

$$\boxed{\frac{\delta Q}{|Q|} = |n| \frac{\delta x}{|x|}} \qquad (55)$$

The second form is probably easier to remember: the fractional (or percent) uncertainty gets multiplied by $|n|$ when you raise x to the nth power.

There is a very important special case here, namely $n = -1$. In this case the rule says that the percent uncertainty is unchanged if you take the reciprocal of a quantity. (This, incidentally, is why multiplication and division are treated exactly the same way in section 5.2 above.)

Example: the period of an oscillation is measured to be $T = 0.20 \pm 0.01$ s. Thus the frequency is $f = 1/T = 5$ Hz. What is the uncertainty in f? Answer: the percent uncertainty in T was $0.01/0.20 = 5\%$. Thus the percent uncertainty in f is also 5%, which means that $\delta f = 0.25$ Hz. So $f = 5.0 \pm 0.3$ Hz (after rounding).

5.4 More Complicated Expressions

Occasionally you may run into more complicated formulas and need to propagate uncertainties through them. Generally, the above rules, when used in combination, will be sufficient to solve most error propagation problems, as long as you remember the rule that *the errors being propagated must be uncorrelated.* Practically speaking, this means that you have to write your equation so that the same variable does not appear more than once.

This is best illustrated by an example. Suppose you have a variable x with uncertainty δx. You want to calculate the uncertainty propagated to Q, which is given by $Q = x^3$. You might think, "well, Q is just x times x times x, so I can use the formula for multiplication of three quantities, equation (46)." Let's see: $\delta Q/Q = \sqrt{3}\delta x/x$, so $\delta Q = \sqrt{3}x^2\delta x$. But this is the wrong answer—what happened?

What happened was that equation (46) does not apply if the three quantities have correlated uncertainties. (Obviously, any variable is correlated with itself.) Instead, you have to use equation (55), and you would get the proper answer ($\delta Q = 3x^2\delta x$).

The same sort of thing can happen in subtler ways with expressions like $Q = \frac{x}{x+y}$. You can't use the usual rule when both the numerator and denominator contain x, so you have to first rewrite Q as $\frac{1}{1+y/x}$, and then apply several rules: first the quotient rule to get the uncertainty in y/x, then the addition rule to get the uncertainty in $1 + y/x$ (fortunately this one is trivial), and then the power rule to get the uncertainty in $(1 + y/x)^{-1}$. The process is perhaps easiest to understand by working backwards:

$$\begin{aligned}
\delta Q &= Q\,|-1|\,\frac{\delta(1+y/x)}{1+y/x} && (56)\\
&= Q\frac{\delta(y/x)}{1+y/x} && (57)\\
&= Q\frac{(y/x)\sqrt{(\delta y/y)^2+(\delta x/x)^2}}{1+y/x} && (58)\\
&= \frac{xy}{(x+y)^2}\sqrt{(\delta y/y)^2+(\delta x/x)^2}. && (59)
\end{aligned}$$

At this point you might be wondering, "Isn't there some kind of general expression that can be used to propagate uncertainties?" There's good news and bad news on that front. The good news is that the answer is yes. The bad news is that there are actually several more general expressions. Let's start with one: suppose you have a single uncertain variable x and some other variable Q that can be written as a function of x: $Q = f(x)$. (If Q depends on any other quantities, they must be exact quantities with no uncertainty.) Then the uncertainty in Q is given by:

$$\delta Q = \left|\frac{df}{dx}\right|\delta x. \qquad (60)$$

As an example, if you measure some angle θ with uncertainty $\delta\theta$, you could use equation (60) to determine the uncertainty in $\sin\theta$ or $\cos\theta$.

Now, equation (60) itself could be generalized to functions of several independent uncertain variables (and then further generalized still, to cover cases where the variables are not independent), but in this course, we won't need to use such cases. If you are interested in reading more about it, look up the references cited at the end of this document, or just do a google search for error propagation.

5.5 Comparing Two Quantities

One of the most important applications of error propagation is comparing two quantities with uncertainty. For example, suppose Anya and Buffy both measure the speed of a moving ball. Anya measures 3.6 ± 0.2 m/s and Buffy gets 3.3 ± 0.3 m/s. Do the two measurements agree? If the two values were slightly closer together, or if the two uncertainties were slightly larger, the answer would be a pretty clear yes. If the uncertainties were smaller, or the values further apart, it would be a pretty clear no. But as it is, the answer isn't clear at all.

Here's where error propagation comes to the rescue. Let's call Anya's result $A \pm \delta A$, and Buffy's result $B \pm \delta B$. To see if they agree, we compute the *difference* $D = A - B$. Using error propagation, we can figure out the uncertainty δD. Then the question of whether A agrees with B, with uncertainties on both sides, has been simplified to the question of whether D agrees with zero, with uncertainty on only one side.

Let's work out the result of our example. Using the rule for a sum (or difference), we get

$$\begin{aligned} \delta D &= \sqrt{(\delta A)^2 + (\delta B)^2} & (61) \\ &= \sqrt{(0.2 \text{ m/s})^2 + 0.3 \text{ m/s})^2} = 0.36 \text{ m/s}. & (62) \end{aligned}$$

Since $D = 0.3 \pm 0.4$ m/s, we see that zero is comfortably within the uncertainty range of D, so the two measurements agree.

This raises an interesting question: have we now shown that Anya's measurement and Buffy's measurement are equal? The answer is no—we've merely shown that they *could* be equal. There is no experiment you can perform to prove that two quantities are equal; the best you can do is to measure them both so precisely that you can put a very tight bound on the amount by which they might differ. (For example, if we had found that $A - B = 0.003 \pm 0.004$ m/s, you might be more convinced that $A = B$.)

However, it *is* possible to show that two quantities are *not* equal, at least to a high degree of confidence. If you follow the above procedure and find that D is 3 times as big as δD, that puts serious doubt into the hypothesis that the two quantities are equal, because $D = 0$ is three standard deviations away from your observed result. That could just be due to chance, but the odds are on your side: you can be 99.7% confident that the two quantities were actually different.

This is often how discoveries are made in science: you start by making a prediction of what you'll see *if whatever you're trying to discover isn't actually there.* This prediction is called the *null hypothesis.* Then you compare the null hypothesis to your experimental result. If the two differ by three standard deviations, you can be 99.7% confident that the null hypothesis was wrong. But if the null hypothesis was wrong, then you've made a discovery! (Of course, if it is an important discovery, the scientific community will want to repeat or extend your experiment, to increase the level of confidence in the result beyond a simple 3σ measurement.)

References

[1] Harrison, David M. *Error Analysis in Experimental Physical Science.* Dept. of Physics, University of Toronto, 2004. `http://www.upscale.utoronto.ca/PVB/Harrison/ErrorAnalysis/`.

[2] McCarty, Logan. *An Introduction to Measurement and Uncertainty.* Physical Sciences 2 course homepage. Dept. of Physics, Harvard University, Sept 2006. `http://isites.harvard.edu/fs/docs/icb.topic109487.files/PS2-Error-Uncertainty.pdf`.

[3] Taylor, John R. *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. 2nd ed. Sausalito, CA: University Science Books, 1997.