Modelling data

Much of the course so far has emphasized how we might go about making simulations of things – i.e. things that are useful for a theoretical physicist who wishes to figure out what might be the expected outcome of some theoretical model. Some of you who plan to do experimental physics or observational astronomy may have been left wondering why computers are really useful for you.

The main point of our lesson this week will be to discuss how to test theoretical models against real data, and how to make estimates of the values of parameters based on theoretical models, when the actual key parameter cannot be measured directly – these are key goals for any experimentalist. First, I'm going to give a bit of a discussion of philosophy about how to work with data. Then, I'm going to give you a recipe for one of the most popular and useful ways to fit models to data, but only after telling you all the reasons not to trust it blindly. At the end of this week's notes, I'll also discuss briefly a few other places where computers are used in experimental physics, but in ways that don't really fit in with this course.

Einstein is quite famous for having said, "No amount of experimentation can ever prove me right; a single experiment can prove me wrong." In the experimental sciences, we can never prove a theory is correct. What we can do, however, is to show that it describes nature very well over a broad range of circumstances – once we do that, it rises from the level of a hypothesis or a model to the level of a theory or a law. We can also prove that one theory provides a better description of nature than another theory does.

Sometimes it can be generations between when an idea is established as a theory which explains a wide variety of phenomena very well until the theory can be established not to describe some experimental data. In these cases, the old theory is often retained as useful and is often still used in cases where it was already established to provide a good description of what Nature does. An example, to which we will come back a bit later, is that of Newtonian gravity versus general relativity as the "correct" theory of gravity. Newton's theory provides calculations which are more precise than any deviation we could hope to measure for a wide range of conditions, and it involves doing calculations simple enough for high school students. Einstein's theory involves using mathematics that very few people learn even in an undergraduate physics degree, and in most cases, results in corrections that don't matter. For certain purposes, almost exclusively in astrophysics, we use general relativity as our theory of gravity. Otherwise, we use Newton's theory.

A general philosophy of how to analyze data

We should never think in science about whether hypotheses are right or wrong. Instead we should think about whether the hypotheses provide a good description of what happens in experiments or a poor description. Hypotheses which provide poor descriptions of nature can be rejected. Hypotheses which provide good descriptions of nature can be retained. This doesn't make them "right" in the sense that something that can be proved mathematically is right. It makes them useful.

A large fraction of the realm of statistical analysis is concerned with hypothesis testing – determining whether we should reject an idea. This is more of an art form than a science – although it's a rather unusual art form that often involves doing calculus.

What we want to do is to find some mathematical technique that helps us to determine whether our model is acceptable or rejected on the basis of the data. There are three important things that we must keep in mind when we are doing this:

1. Sometimes the data are not precise enough to allow us to find the flaws in our model, or the deviations between the model and reality aren't important within the range of parameter space in which we have tested the model.

2. Sometimes the statistical technique we are applying is not quite correct, or the data are not correct.[1]

3. Sometimes the model which is, in essence, correct, deviates from the data more than a model which is totally different from the data. This can happen if there are a lot of fudge factors in one model, and the other model is not fully developed yet.

Let's consider some examples for each of these cases.
Data not precise enough, or tests not made in the right regime

For quite a long time, it appeared that Newton's theory of gravity and Galileo's theory of relativity were just flat out correct descriptions of Nature. We now know that Einstein's theory's of general relativity, and special relativity, are better descriptions of nature. We still use Newtonian gravity and Galilean relativity for a lot of purposes, of course, because the differences between them and Einstein's theories, in "weak" gravitational fields, and at "slow" speeds are very small, and because the calculations involved in using Einstein's theories are much more difficult.

Even before Einstein produced general relativity, there were a few measurements that were better explained by it than by Newtonian gravity. The most famous involves the orbit of Mercury – the precession of the periastron of its elliptical orbit. After correcting for the effects on Mercury's orbit from all the other large, known bodies in the solar system, there were still some effects that remained. We know now that general relativity can fix the problem. Before other evidence also showed general relativity was a better theory than Newtonian gravity, the general assumption had been that there was probably an

[1] The data, of course, are always correct, unless we have done something like make a copying error in collecting the data. But often, we don't understand the calibration of our experiment well enough, and the data don't mean what we think they do, and when we convert from the "raw" data, which are things like "a number of photons we have detected with a photon counting device" or "the time series of voltages measured by a voltmeter" to the kinds of things we compare directly with the predictions of a physical theory, we sometimes make mistakes.

unknown planet in the outer solar system or something like that that was causing these effects. Some other tests, notably the detection, during a solar eclipse, of a star that should have been behind the Sun, really helped cement the idea that general relativity worked, because there was no way to explain such a result in Newtonian gravity, and general relativity explained it to high precision.

In any event, if we had had a cleaner solar system, so that there was not a possibility that Mercury's orbit was affected by something we didn't know about, or something like the binary pulsar which came later on, we could have proved that general relativity outperformed Newtonian gravity much earlier on. Similarly, if we had had the opportunity to make measurements in a stronger gravitational field than that of the Sun, we also could have made measurements earlier on, since the deviations from Newtonian gravity would have been smaller. Instead, until the solar eclipse measurements, we didn't have good enough data to tell the difference between Newtonian and Einsteinian gravity for quite a while.

"Systematic errors" or "wrong data"

Many of you may remember a few years ago there was a claim that neutrinos were seem to be travelling faster than the speed of light. Most physicists, at the time, assumed that there was probably something wrong with the experimental data. This is almost always the reaction of most physicists when such an important (or in this case potentially important) discovery is made. It was later found out that there were two pieces of equipment failure in the experiment. The major one was simply a loose cable. As a result, the measurements did not mean what the physicists thought they meant.

Cases where a model is basically right, but not sophisticated enough

We know realize that the Earth goes around the Sun. For a long time, though, there was a philosophical notion that the Sun and the planets orbited the Earth. The astronomers of the time were able to make mathematical formulae that could predict the locations of all the planets to high precision by using "epicycles" – basically making circles within circles. The small circles were orbits within the big circles – sort of the way the moon orbits the Sun, making smaller orbits around the Earth.

When Kepler made the first real quantitative predictions about the orbits of the planets on the basis of a heliocentric solar system, his predictions were not as accurate as the predictions of the geocentric model with epicycles. Those epicycles were basically fixing both the global problem with the geocentric model, and with problems related to the gravitational attractions of the planets on one another and the moon on the Earth.

This is the diciest question for scientists to deal with. One can invoke Occam's razor, which states that the simplest theory that explains all the data is correct. But what happens when one theory is much much simpler and the other explains the theory slightly better? Or when one theory is simpler, but rests on questionable assumptions? In some cases, there are mathematical techniques for dealing with this – *Bayesian analysis*, and the use of something called the f-test. We won't get into these, but Bayesian analysis really just forces us to state a lot of our assumptions very clearly ahead of time, and the f test is really

only useful under cases where one model is exactly the same as another, except for an added level of complication.

Usually the way we should deal with these cases, instead, is by getting some outside information for which the two models make very different predictions. In the case of the solar system, the key idea was put forth by Galileo (and actually before Kepler came along) – that Venus shows phases in the same way that the moon does. Galileo showed that there was no way to have this happen if both Venus and the Sun were orbiting the Earth, but that it would happen naturally in the Copernican/Keplerian view of things. The moral to this story is that we should not become overly reliant on fitting one type of experimental data against models, particularly when the data and the models agree reasonably well. Instead, we should look for the kinds of tests of theories where the competing theories differ as much as possible.[2]

χ^2 fitting: a workhorse for experimental sciences

Most of the time in the experimental sciences, we will be working with data where the measurement errors are well described by a Gaussian function. Another way of saying this is that if one were to make a particular measurement an infinite number of times, then the probability density function that would result would be a Gaussian with mean. If we go back to the formula from a few weeks ago:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-} \frac{(x-\mu)^2}{2\sigma^2}, \tag{1}$$

we can now take a look at the meanings of the parameters for the Gaussian probability density function.

The mean value of the function will be μ and the standard deviation of the function will be σ. Usually, if a value is reported as $x \pm y$, y is the value of σ for the variable x. This is not always the case – in some fields of research slightly different numbers are used. But you can see that it doesn't mean that the full possible range of values of x is from $x - y$ to $x + y$. This is an important point to take home, because this is a point that confuses students a lot of the time.

So what is σ? It is the square root of the variance of the data. The variance is what is called the second moment of the data. The nth moment about a particular value c is defined by:

$$\mu_n = (x-c)^n f(x) dx. \tag{2}$$

If we set $c = 0$, and $n = 1$, we get the first moment, which is the mean. If we set $c = \mu$ (so that we are taking the moment about the mean), and set $n = 2$, then we get a quantity called the variance. The square root of the variance tells

[2]This is especially true when the experimental data are known about before the theoretical work is done, which is most of the time. Then, the theoretical work is aimed at explaining the data. This is a perfectly reasonable approach, but it is also something that we need to bear in mind – theorists will often adjust the assumptions they make a little bit to be sure they match the existing data. Coming up with new tests which make radically different predictions pushes the theorists into a corner. Good theorists try to push themselves into a corner by proposing ways that their theories could be proved false.

Number of sigmas	Percentage of distribution
0.67	50
1	68.27
2	95.45
1.645	90
3	99.7
4	99.9936
5	99.999943

Table 1: The table gives the percentage of a Gaussian distribution enclosed within $\pm$ the number of multiple σ away from the mean.

us something about the "typical" uncertainty in the parameter value. If $f(x)$ is a Gaussian, then μ will be the mean and σ^2 will be its variance.

Since you already know how to integrate Gaussians, you can check this, but there are regions called confidence intervals for Gaussians:

For large multiples of σ away from the mean, it can be more convenient to give the probability that an event takes place outside the distribution, rather than within it.

If we have a single variable with a normal distribution, then we can test a measurement by looking at the difference between our data value and the predicted value based on our hypothesis in terms of the difference in units of σ. What we try to do when we compare a model with some data is to estimate something called the "null hypothesis probability". This is the probability that just random chance, based on the measurement errors, would give us as bad a match we get when we do the experiment, even if the model were an exactly correct description of what is happening in the experiment. A large null hypothesis probability means that the hypothesis cannot be rejected *by this experiment alone*. That is, the hypothesis has survived this test. This is, at some level, evidence in favor of the hypothesis, but it is not proof of the theory – as we have stated above[3], no theory can ever be proved correct, but rather theories can merely be demonstrated to provide excellent descriptions of nature over a broad range of conditions.

Suppose we had reason to believe that the Earth's gravitational constant is 9.8 km/sec/sec exactly at sea level, and we were trying to test whether we measured a different value at an elevation of 3000 m. Suppose we measured a value of 9.705 ± 0.019. We then subtract the two values and get: 0.095 ± 0.019. This is 5σ and we can see that 99.999943% of the measurements should be within 5σ of the expected value. Therefore, there is only a 0.000057% chance that such a large deviation would occur due to measurement errors, provided that we have correctly estimated our measurement errors. Under these circumstances, we could very safely say that g has changed between our two measurement locations. This would also provide strong support for the idea that g changes as we increase altitude above sea level, but we would want to take more measurements at

[3]It bears repeating.

different altitudes to be more confident in drawing that conclusion.

This approach is thus great, if we just want to compare one number with another number. That's not what we usually want to do in physics. Usually, we want to compare a set of measurements with a model that attempts to describe what is happening in Nature. Most often, what we will have is data in the form of x and y, with uncertainties on x and on y. Our procedure is greatly simplified if (1) y is a function of x and (2) the errors on x are sufficiently small that we can ignore them. For the purposes of this class, we will proceed as though those conditions are met.[4] We will then have three quantities associated with each measurement we make – the values of x, y and σ_y.

We won't go into a full mathematical derivation, but what we can do is to definte a new quantity called χ^2:

$$\chi^2 = \sum_{i=0}^{n} \left(\frac{f(x_i) - y_i}{\sigma_{y,i}} \right)^2 , \tag{3}$$

where n is the number of data points we have, the subscript i is telling us that we are using the ith value in the sequence, and $f(x)$ is the function we are testing. Essentially, we are summing the squares of the differences between the data and the model. We can perhaps make a leap of intuition without doing all the rigorous math here, and say that this makes sense, since σ is the mean value of the square of the difference between the actual value of the data points and the values you would get by sampling the distribution randomly a large number of times.

Next, we have to figure out how to interpret that. We need to know how many "degrees of freedom" we have for the model. A degree of freedom is essentially a "test" of the model The number of degrees of freedom is typically represented by ν. If we have a model which is completely constrained ahead of time, then the number of degrees of freedom is the number of data points. However, in most cases, our model will have parameters that we are trying to estimate at the same time that we are trying to test whether the model is correct. Every time we have a parameter that we allow to vary in order to determine whether the data can be explained by the model, we use up one of the degrees of freedom provided by the data. There are mathematically rigorous ways to think about this, which are a bit too complicated for this course, but a quick way to think about this is that you can always fit an n element data set with a polynomial of order $n-1$ if you can vary all n coefficients (including the constant) – but that doesn't mean anything, really. On the other hand, if you can fit $n+50$ data points with a polynomial of order n, that really probably does mean that the data are well described by a polynomial.

OK – so now we know how to compute χ^2 and we know how to compute the number of degrees of freedom. As you have probably guessed, if χ^2/ν equals 1, we have a good fit. The model is a statistically acceptable description of the data. But how do we figure out what the situation is when χ^2 is *not* 1? When

[4] It is often not the case, and then more complicated procedures must be used for testing hypothesis.

is a fit rejected? We can compute the probability of getting a value of χ^2 at least as large as the measured one by chance. The formula:

$$F(\chi^2, \nu) = \frac{\gamma\left(\frac{k}{2}, \frac{\chi^2}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \tag{4}$$

gives the total probability that one would obtain a value of χ^2 less than or equal to the measured value, so the probability of getting a worse answer by chance is $1-F$. The function γ is something called the lower incomplete gamma function and the function Γ is the actual gamma function. If you ever want to write a computer program to compute the these things both of these functions are in the GNU Scientific Library and most commerical mathematical packages.

The one last thing we need to worry about is how to get the best values of the parameters for the model we have in mind. What you want to do is to find the minimum value of χ^2 for a given model by changing around the parameters. A variety of algorithms do this. The Levenberg-Marquardt algorithm is probably still the most commonly used one, but there are more computationally intensive algorithms that can sometimes do a better job. The chief worry with any fitting algorithm is that it may find what is called a "local" minimum of χ^2 – that is a set of parameters which does not give the actual minimum for χ^2, but for which *small* changes in the parameter values result in a worse fit.

Fitting functions with GNUPLOT: some recipes and some examples

Now, with the philosophy of hypothesis testing, and the theoretical foundations for using χ^2 fitting laid out, we can move on to how we go about actually fitting functions to data within `gnuplot`.

I have generated a file, "output_values_linear" with fake data with errors. The data are produced from a straight line with $y = mx + b$, with $m = 1$, and $b = 2$. I then randomly assigned 30 values of x, and 30 values of σ_y to the data points. I then computed what y should be, and then changed y by taking a random deviate from a Gaussian distribution with $\sigma = \sigma_y$. Basically, this data set should be consistent with a straight line with slope 1 and intercept 2. We probably will get values very close to those values, but not exactly equal to them. Let's give it a try.

First, here are the numbers:

Now, let's first plot the data in `gnuplot`:

```
gnuplot> plot "./output_values_linear" u 1:2:3 with yerrorbars
```

and we see that it looks pretty much like a straight line, so it is reasonable to try a straight line model.

Now, we need to define the model we want to use:

```
gnuplot> f(x) = m*x+b
gnuplot> m=1
gnuplot> b=2
gnuplot> fit f(x) "./output_values_linear" u 1:2:3 via m,b
```

We then get some output: $m = 0.997465, b = 2.03362$ with 28 degrees of freedom and χ^2/ν of 0.779. We have a good fit, and we recovered what we thought we were going to recover.

Now, let's see what happens if we try to fit an exponential to the data.
`gnuplot> f(x) = m*exp(-x/b)`
`gnuplot> fit f(x) "./output_values_linear" u 1:2:3 via m,b`
Here, we get a value of χ^2/ν of more than 75 – basically the exponential is such a horrible fit that b runs up to a very high value, so that effectively the exponential becomes a constant that is the mean value of y.

Finally, you should always take a look at the residuals any time you fit a function to some data. `gnuplot> plot "./output_values_linear" u 1:(2-f(1)):3 with yerrorbars` will give you a plot of the data minus the model. Sometimes when you get a reasonably good fit, you will still see a trend in the residuals that tells you that there is something interesting going on that you might have missed otherwise. Also, when you get a bad fit, you can often isolate what is going wrong and figure out how to make a slightly more complicated model that will then become a good fit. I saw a nice example of this from a carefully-done experiment made by my lab instructor when I was a freshman in college. He had measured current versus resistance with a variable resistor and an old fashioned ammeter that had a needle on it. He found that a straight line was a very good fit to the data, but then there was a small additional smoothly varying residual. This was due to the weight of the needle on the ammeter.

So, this is just a start to how to analyze data, but if you master the ideas in this set of notes, and you can do the exercises, and you remember when you are violating the key assumptions that go into χ^2 fitting and then either take the results with a grain of salt, or use a more sophisticated approach, you should be able to do a lot of useful work.

Other uses of computers in experimental/observational sciences

1. Computers are often used in the data collection process in the experimental sciences. For example, computers can be used to control experiments, or to collect readout data from an instrument (e.g. an oscilloscope or a voltmeter or a charge coupled device). This type of work often involves specialized techniques for programming, since it gets in to the details of the hardware, and often involves specialized software for using, since the user just wants the answer, and doesn't want to have to understand the details of the computer hardware, unless the computer hardware can lead to distortions of the measurements. It is thus a very important area, but not really appropriate for a first course on programming.

2. Computers are often used to do simple processing of large data sets. We discussed the case of the Fourier transform already, which is often used for processing time series data (and sometimes, two dimensional Fourier transforms are used to process images). A variety of other types of time series and image processing are done in physics, astronomy, geology, engineering, and increasingly in certain areas of biology and chemistry. The data from the Large Hadron Collider, and other particle accelerators, are compared with computer simulations of what different kinds of particle interactions should look like, and this is how it is inferred what types of particle interactions took place in the collider.

3. Computers, and the internet can be used to distribute data to colleagues, and to the general public. This is rather straightforward in the case of simply sending data to your colleagues. One of the more interesting applications in the past couple of decades has been that of "citizen science" projects.

   These projects have taken a couple of forms. One is the "volunteer computing" approach. Certain types of tasks require a large number of computations, which can be split up easily. At the same time, large numbers of computer owners use their computers only a small fraction of the time. Some enterprising scientists have created software which people can download that allows their computers to do work on projects for the outside scientists. A few physics-related projects which have taken advantage of this include LHC@home, which does simulations needed for making the most of the Large Hadron Collider, Einstein@home which searches radio telescope data for evidence of gravitational radiation, and SETI@home, which involves searching through data from radio telescopes for signals of extraterrestrial intelligence. Interestingly, there are also many biology projects using volunteer computing, but they are almost all for theoretical calculations of protein structures, rather than for analysis of experimental data. If you are interested in contributing computer CPU cycles, check out boinc.berkeley.edu.

   The other way that citizen science can work is when there are types of data that need to have a large number of humans look at them. There are still types of pattern recognition that humans do much better than machines. Many of these are on the borderline between psychology and artifical intelligence – e.g. the Royal Society's Laughter project in which you listen to some laughs and try to determine if they are real or fake. Others involve physics, like the Galaxy Zoo project in which you look at images of galaxies from large sky surveys and tell which type of galaxy it is, which direction it's rotating, and also flag up if anything looks unusual. Large numbers of non-experts can be trusted to get the right answers, as a group, an awful lot of the time, and a small fraction of the data can be examined by experts in order to determine how often things go wrong by trusting non-experts. This is a type of project that requires the internet. If people needed to make printouts and pay postage in order to send the data back and forth, the cost of such projects would be prohibitive. The Galaxy Zoo project has resulted in a few discoveries of things that were made only because a human looked at the images, but which didn't require real experts to be the first ones to notice something unusual. I am not enough of an expert on the marine biology experiments using the same techniques to identify types of fish in videos from underwater experiments, but I would guess that they are just as effective.