Statistics: Part 1

1. Why bother with statistics?

Why is statistics so necessary for observational astronomers? Here are some examples of typical tasks that require some stats knowledge:

- Writing a telescope proposal. You need to know what signal-to-noise ratio you require, and how long you have to expose for to reach it. It's also crucial for many projects to be able to justify why you need to observe a given number of targets. All of these require stats.
- Fitting a model to some data. Let's say you're trying to measure the cosmological constant. You have a variety of models, and you want to see which ones best fit the data.
- Survey modelling. You might be part of a team proposing to build a giant new telescope that will find planets about other stars. What would be the best observing strategy? Should you look at lots of stars, or concentrate on a few and observe them really hard? The choice will affect the whole design of your telescope.
- Proving somebody wrong: let's say a rival group are claiming an exciting result. But your data just don't seem to back up their claim. Statistics allows you to determine how confidently you can stand up and say "you are idiots"!

In these notes, I'll give a very brief and (hopefully) practical introduction to those bits of statistics most vital to observational astronomy.

2. What is a Statistic?

The fundamental purpose of statistics is to baffle students. Whoops – I mean it is to simplify the world around us. There are vastly more numbers out there than anyone's brain can handle: stats is our tool to boil it all down into a few numbers that tell us what we really want to know. A "statistic" is technically just such a number, derived from some vast pile of data, which hopefully tells us something interesting. For example:

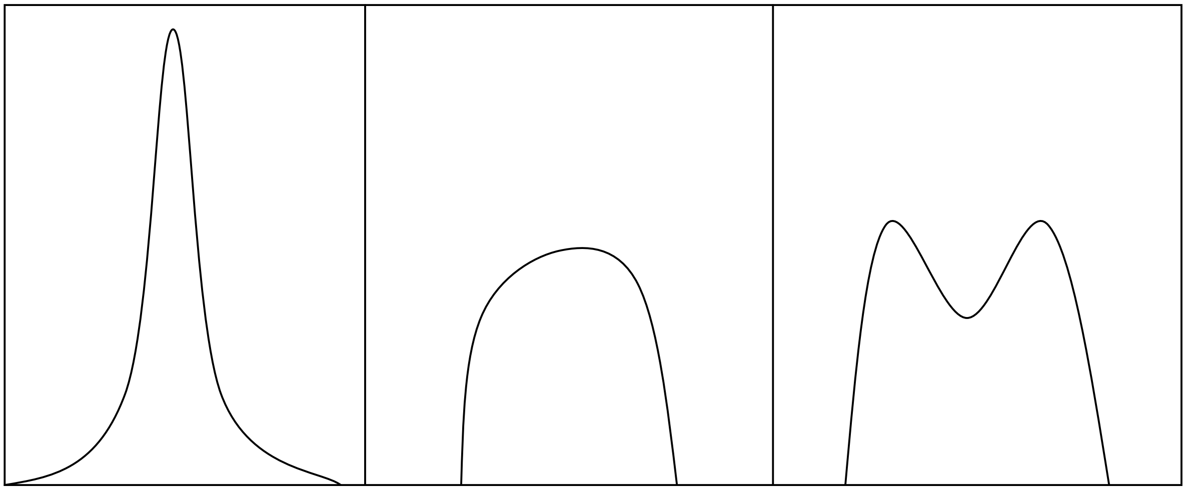
- The 2dF Galaxy Redshift Survey obtained over 200,000 spectra, each containing several thousand numbers. But for many purposes, all you need to know is the redshift of each galaxy. And all some of us want to know is what constraint this survey puts on cosmological parameters. So a handful of values and their error bars are the statistics extracted from more than 10^8 raw measurements.

A very common error is to lose sight of this very basic truth: statistics is the extraction of what you want to know from masses of data. The best way to address most statistical problems is as follows:

1. Think really hard about what it is you are trying to learn.
2. Invent a "statistic" that will tell you this.
3. Measure this "statistic" from your data.

Here's an example. Say you have measured the HI (neutral hydrogen) spectra for lots of galaxies, using the Parkes radio telescope. For each galaxy, you have a plot of intensity vs. velocity.

A bad way to approach this data would be to measure "statistics" from each plot, such as the line width, and then try to do something with them. Here's why:



Which of these lines is the widest? If you measure the width at zero intensity, it is the first one. But a common measure of line width is "full width at half maximum height": by this measure, the last is the widest. Many line profile analysis programs fit Gaussian curves to the lines, and quote the line width and flux as that of the best-fit Gaussian. But clearly none of these lines is a Gaussian. Almost any statistic you apply to line profiles as varied as these will get you into some trouble of this sort.

The best thing to do is to think: "what do I really want out of these data"? It might be, for example, that you want to measure the mass of the galaxy, using its rotation velocity. In this case, you should use galaxy models to generate some HI line profiles, and see which measure of their width or profile most closely agrees with the mass.

Here are a couple more examples, from outside astronomy, which may make it clearer why it is really important to think carefully about what statistic you want to measure and why.

- A marketing director from company selling luxury sports cars wants to know where in Australia she should build her first showroom. Her cars sell for several hundred thousand dollars each, so she wants to make sure she puts it in an area with lots of rich people. She checks out the Australian Bureau of Statistics web site, and finds that the ACT has the highest median income of any state or territory. She builds the dealership here, sells no cars, and gets the sack. What went wrong? Well – the median income is the income which splits the population in half – so that half the people have a higher income and half lower. The ACT is a very middle-class place, where most people earn a reasonable income. There are few very poor people, but also few very rich people. Somewhere like NSW, however, has many more poor people. It also has more millionaires. But the number of poor people vastly outnumbers the number of millionaires. So the median income is pulled down, not up. She should have used a different statistic, such as "fraction of the population earning more than a million dollars per year"

- The Government says that "spending on universities has never been higher, please vote for us again". The opposition says "spending on universities has never been lower, throw the bastards out". How can they both be true? They are using slightly different statistics, of course. The government correctly states that the number of dollars spent on universities has increased. But the opposition correctly points out that the fraction of GDP spent on universities has dropped. Which is different, because the GDP of Australia has grown. Which is correct? Well, it depends on why you want to know. If, for example, you want to know whether universities will have to lay off staff, what matters is whether their income has gone up faster than typical salaries. If you want to know whether Australian universities are being funded at an internationally competitive rate, what matters is funding relative to international competitors. And so on…

The bottom line of all this is: don't just blindly apply some well know statistic, such as a mean, standard deviation or correlation coefficient to the data, unless you're quite sure it will tell you what you want to know.

3. Some Common Statistics

Over the years, people have come up with an amazing range of statistics. Most are pretty specialised, but a handful have become widely used.

The most common of all is, of course, the mean, or "average". If you have a lot of numbers, all somewhat different, the mean is one of several ways of summarising a "typical" or central value. If you have n numbers x_i, the mean μ is, of course, defined as:

$$\mu \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$$

The mean is so widely used that it somehow acquires a magical status. But in fact, it's not a particularly good measure for most purposes. Firstly, it is easily biased by even one outlying point. For example, let's say you have taken ten measurements of the brightness of a particular star. The values you measured are 4, 6, 4, 5, -17435, 4,6, 4, 3, 5. Clearly most points are pretty close to 4 or 5. But there is one measurement that looks pretty wrong. Something obviously went wrong with that measurement. If you work out the mean, it is -1739.4, which is not close to *any* of the measurements!

If you are sure that all your data are good, the mean is a sensible statistic to use. But if there is a risk of some iffy numbers, you are better off using a *robust* statistic. Robust statistics are those that are fairly resistant to a few flaky data points. In this case, you'd probably use the median. To compute a median, sort all the numbers into ascending or descending order, and take the middle one (or the average of the middle two if you have an even number). It's harder to compute than the mean, because you have to sort your numbers, but it's resistant to occasional way-off points (in this case the median is 4.5).

Most traditional statistics are like the mean – easy to compute but not robust. There is almost always, however, one or more robust counterparts, which you may be better off using in most practical situations. In radio astronomy, for example, the AIPS package uses traditional statistics, but the MIRIAD package uses robust statistics.

The second most common statistic is probably the standard deviation σ. Given a bunch of numbers of mean μ, it measures how closely the remaining numbers are clustered around the mean. If all the numbers are the same, the standard deviation is zero. It is defined by:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^{\,2} - \mu^2}$$

In other words, you take each point, work out how far it is from the mean, square this distance, average all the squared distances and take the root. It's even less robust than the mean – because you are adding the *squares* of deviations together, it is hideously vulnerable to outlying points. But it is fast to compute – using the second version of the equation above, you can compute it with a single pass through the data.

Some other common statistics:

- The correlation coefficient. Given two sets of numbers, it tells you whether when one goes up, the other does too. Its robust counterpart is Spearman's Rank Correlation Coefficient. For example, does the size of a planetary nebula correlate with the luminosity of the white dwarf in its centre?
- Student's t-test. Given two collections of numbers, asks whether they could be drawn from the same population. For example, you might have the metallicities of stars which do, and which do not have planets. Are they consistent with being the same or not? The robust counterpart is the Kolmogorov-Smirnov (KS) test, which is widely used in astronomy.

4. Uncertainties

So let's say you have your data, and you've dreamed up a perfect statistic, which will tell you exactly what you want to know. Your next problem is uncertainties in the data, which will prevent you from measuring the value of this statistic with perfect precision.

There are two broad classes of uncertainty: systematic and random.

4.1 Systematic Uncertainties

What is a systematic uncertainty? This is something that is wrong with your observation, or with the theory that underpins it. Something that you failed to take into account. Here are some examples:

- Edwin Hubble first tried to measure the expansion rate of the universe by looking for the brightest stars in each galaxy. He then assumed that all these stars were the same intrinsic brightness, and hence measured a crude distance. In near-by galaxies he was indeed picking out individual bright stars. But in more distance galaxies, what he thought were individual stars were actually bright compact star clusters. As a result he thought that the distance galaxies were much closer than they really are, and he got the age of the universe badly wrong.

- I was once measuring the spectra of hundreds of quasars. I found that quasars in one part of the sky were systematically much redder in their colours than quasars elsewhere. Luckily, I found out why before publishing. The trouble is that the atmosphere is denser at the bottom than at the top. This means that it acts a bit like a prism, bending light downwards. Because the refractive index of air is greater at blue wavelengths, blue light is bent more than red light. When looking straight up, there is no effect, but the bunch of anomalously red quasars had been observed when they lay quite low in the sky. The blue light from them had been bent so much that it missed the spectrograph!
- About 15 years ago, it was only possible to measure the radial velocity of a star with a precision of ~ 15 m/s. It is now possible to get precisions approaching 1m/s. All this improvement has been due to some very pernickety people tracking down and removing such systematic errors as:
  - Changes in air pressure (and hence air refractive index) shifting spectra.
  - Allowing for the motion of the observatory as the Earth rotates.
  - Allowing for the exact position of the star image in the spectrograph slit.

Systematic errors are both really nasty and really nice. They are nasty because they don't obey the laws of statistics – they don't go away no matter how big your sample. They are nice because, if you are clever enough, you may be able to make them go away, and hence achieve truly fantastic precision.

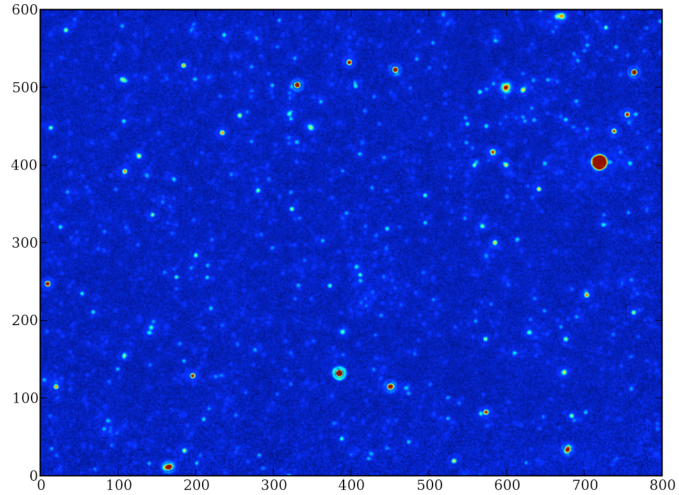
4.2 Random noise

If you've got rid of all systematic errors, what's left to stop you? Truly random noise. This is any process that gives a different number each time you measure it, in a truly random fashion.

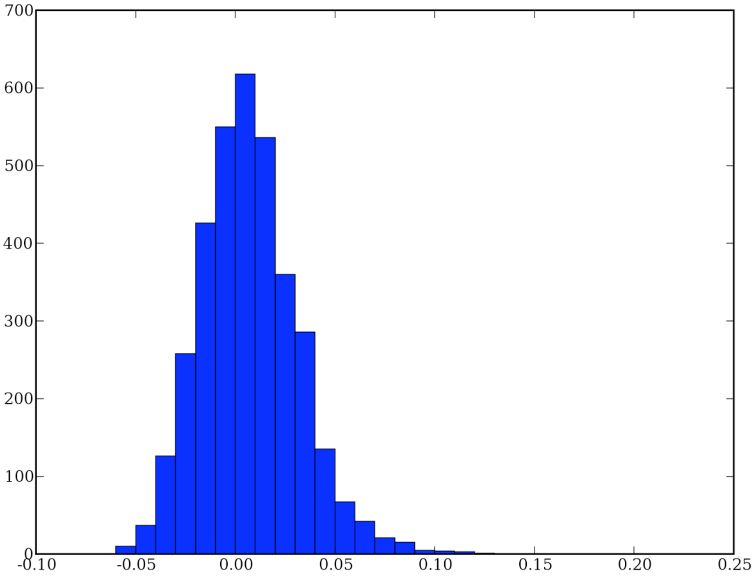
You can characterise the systematic error on a data point by a single number: the difference between the measurement and the true value. But for random noise, you need more than a single number. You need to specify the *probability distribution function.*

The probability distribution tells you what the *probability* is of getting a particular value. Experimentally, you can measure it (approximately) by measuring something over and over again, and plotting a histogram of the results. Formally, the probability distribution function is the limit of this histogram as the number of measurements goes to infinity

Let's look at a real probability distribution. Here is an image taken with the MIPS instrument on the Spitzer Space Telescope.



If we pick a small region of this image that doesn't contain any objects, we will be looking purely at the noise. Here is a histogram of the pixel values in such a region:



So you can see that in a blank bit of sky, the mean value recorded will be pretty close to zero, but that you actually measure quite a range of values. Most lie between -0.05 and 0.05, but there are a few higher pixels. Perhaps because this nominally blank bit of sky isn't really blank but contains a few galaxies, too faint to clearly see in the image, but still contributing to the pixel values.

What causes such random noise? How can something be different every time you measure it? Such as in every different pixel? Usually quantum mechanics is the problem. For example, let's say you are recording the number of photons detected per second from some star. Photons (like any other particle) are described by the wavefunction in

Schrodinger's equation, and the *probability* of detecting a photon in a given time and place is proportional to the square of this wavefunction. So you can never tell whether you actually will or will not detect a photon – all you can calculate is the probability. This is the aspect of quantum mechanics that drove Einstein crazy ("God does not play dice…"). Thermal noise is another culprit. In any electronic circuit, the voltages fluctuate around, partially due to quantum mechanics (you can never tell where an electron really is – only the probability of detecting it at a given location), and partially due to the random thermal jostling motions of atoms, electrons and holes.

Quantum mechanics can also be important on much larger scales. For example, the distribution of galaxy clusters in the universe looks pretty random, and is presumably because of quantum mechanical fluctations at the era of recombination, which eventually became the dark matter seeds about which clusters formed.

4.3 Poisson Distribution

There are many theoretical probability distributions in the stats literature, but two stand out as by far the most widely used, the Poisson Distribution, and the Gaussian Distribution. We'll discuss them in turn.

The Poisson distribution applies when you have a random process that always gives an integer answer. Examples might include:

- The number of cosmic ray protons detected per unit time.
- The number of galaxies in a given volume of the universe.
- The number of photons detected in each pixel of an image.

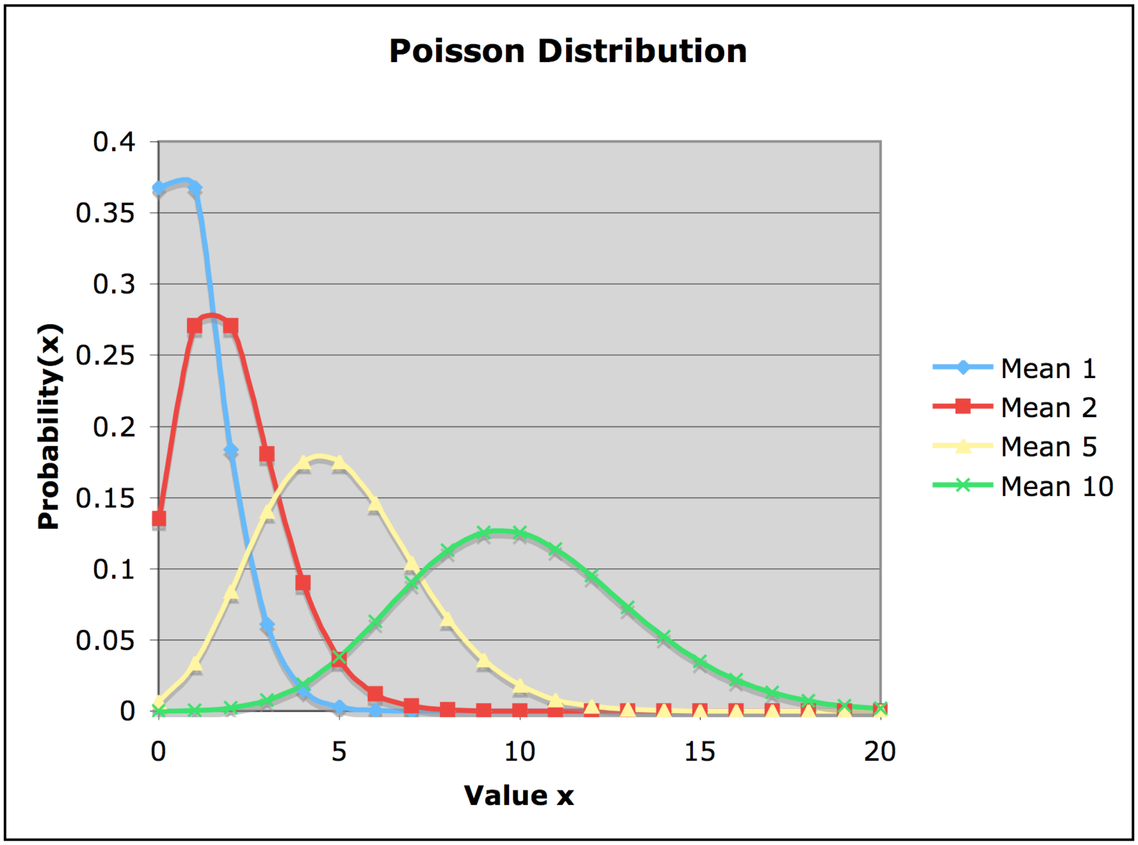
The Poisson distribution assumes that there is a particular probability of finding a given object (such as a photon being detected during a particular time interval or a galaxy being found in a given volume). This probability must be constant, and must not depend on how many photons or galaxies have already been found. So it's fairly accurate for photons, but less so for galaxies, as they cluster, so where you find one, you are more likely to find more.

The probability P of finding x objects is given by:

$$P(x:\mu) = \frac{\mu^x}{x!}e^{-\mu}$$

where μ is the average number of such objects that you would find, and x! is the factorial of x (i.e. x.(x-1).(x-2)…..1).

Here is what a Poisson distribution looks like, for a range of values of μ:



4.4 Gaussian Distribution

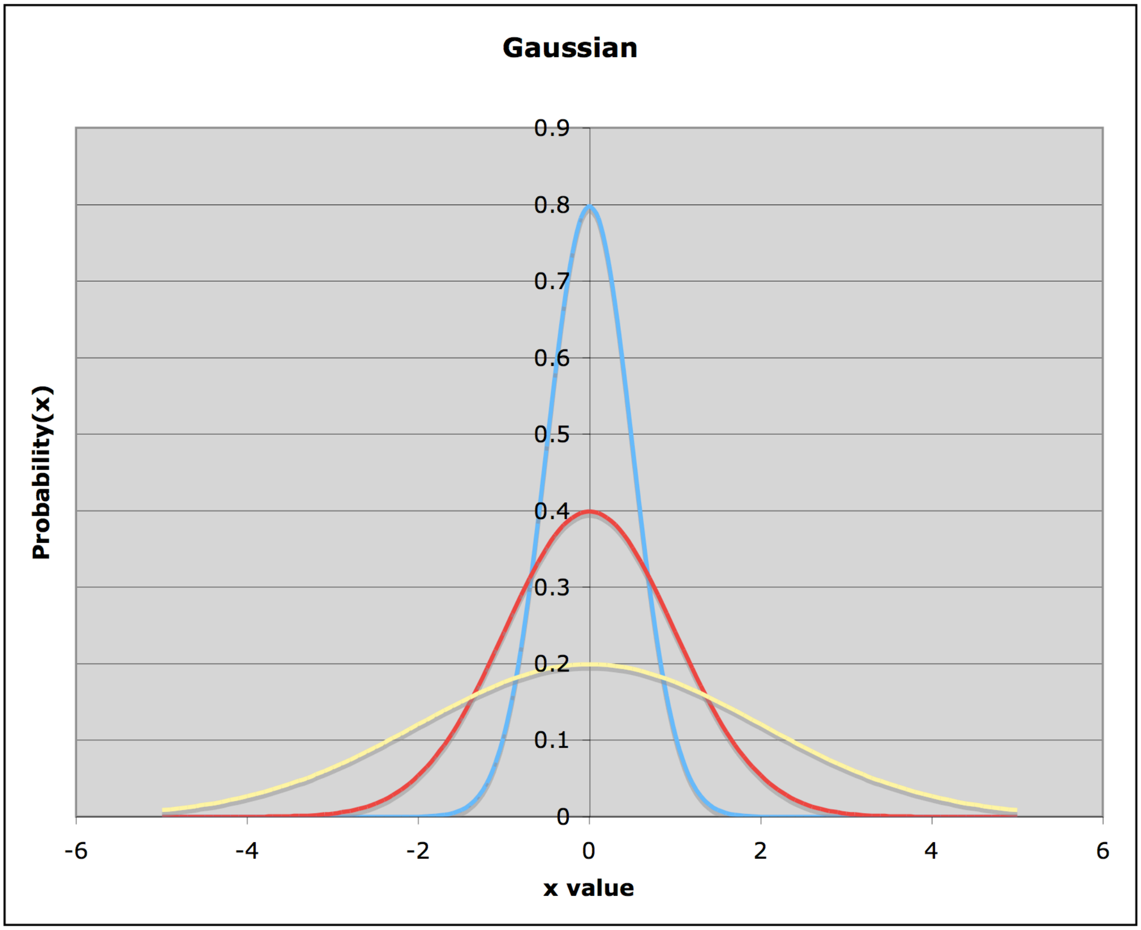
The Gaussian distribution (also known as the Normal distribution and as the "Bell Curve") is **by far** the most commonly used distribution in all of statistics. Indeed it is widely used in places where it shouldn't be.

The Gaussian distribution is a continuous one – so it's used for variables which can take any value, not just integers (unlike the Poisson distribution). So it might be used for things like voltages in a circuit. The probability distribution function P of a variable x is given by:

$$p = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

where μ is the mean of the distribution, and σ is its standard deviation. Exp indicates exponent – i.e. raising e to this power.

Note: because this is a continuous distribution function, the probability of x having exactly a given value is infinitesimally small. Instead, we have to ask what the probability of x lying in the range x_1 to x_2. This is given by integrating the above equation between these limits. The integral of this curve from minus infinity to infinity is 1.



Gaussian curves of standard deviation 0.5 (blue), 1(red) and 2 (yellow).

Unfortunately, this curve cannot be integrated analytically. But many numerical routines and tables exist to integrate it over various ranges. Around 34% of values lie more than one standard deviation from the mean, about 5% lie more than two standard deviations from the mean, but less than 1% lie more than three standard deviations out.

What is special about this functional form, rather than any other curve that goes up and then comes down again? The answer lies in a bit of maths called the 'Central Limit Theorem". Basically, whenever you combine lots of different random things, you end up with something like a Gaussian! As long, that is, as they are all independent of each other, and have not-too-pathological probability density functions themselves.

Check out the following web pages for some examples of how this works:

http://en.wikipedia.org/wiki/Illustration_of_the_central_limit_theorem
http://en.wikipedia.org/wiki/Concrete_illustration_of_the_central_limit_theorem

The Gaussian distribution is also the limiting form of the Poisson distribution, for large values of μ.

This remarkable result is why the Gaussian distribution is so widely used. It is assumed to apply to the IQ of students, to the variability of share prices, to the scatter of galaxies around the Fundamental Plane, to electronic noise in radio receivers, and almost everywhere else.

Indeed, it is often applied to situations where the uncertainties are really systematic, and not random. For example, you often hear that Type-1a supernovae are standard candles. This is approximately true, but for any individual supernova, there is an uncertainty of ~7% in its distance. So some are further than you think while others are nearer. Why? Probably a whole range of physical reasons. Hopefully enough different and independent reasons that the Central Limit Theorem applies, and we can model this scatter as a Gaussian. And if you plot this scatter, it does look somewhat Gaussian.

A word of caution, however. While a Gaussian does provide a reasonable fit in a wide range of circumstances, it should not be trusted too far out in the wings of the curve. Almost invariably you find more extreme points than a Gaussian curve predicts. The Central limit theorem only strictly applies in the limit as the number of independent variables becomes very large, which is never the case in reality..

4. End-to-End Monte-Carlo Simulations.

We now have everything we need to know in order to simulate an observation!

The most accurate way to estimate what exposure time you'll need, and to optimise your observing strategy, it to do an *end-to-end Monte-Carlo* simulation. It is also a lot of work, so in practice shortcuts are often used. We'll discuss some of these in part-2 of the stats notes.

The basic idea is to try and simulate all the steps you will go through in obtaining your real observations. But using fake data, and random number generators to produce fake noise. Then analyse this fake data in the same way that you ultimately plan to analyse the real data and see if you get the right answer.

Here is an example. You'll do another example as the assignment this week.

Imagine that you are observing a red dwarf star, which regularly flares in brightness. The flares only last 1 second, but double the brightness of the star while they last. The star has a V-band magnitude of 11.3. You are observing it in the V-band, using a photon counting device that reads out every ten seconds. You have been allocated three nights of observing time. The star will be high enough to observe for 24 hours between these three nights, so if the weather is good, you should be able to measure its brightness 8640 times.

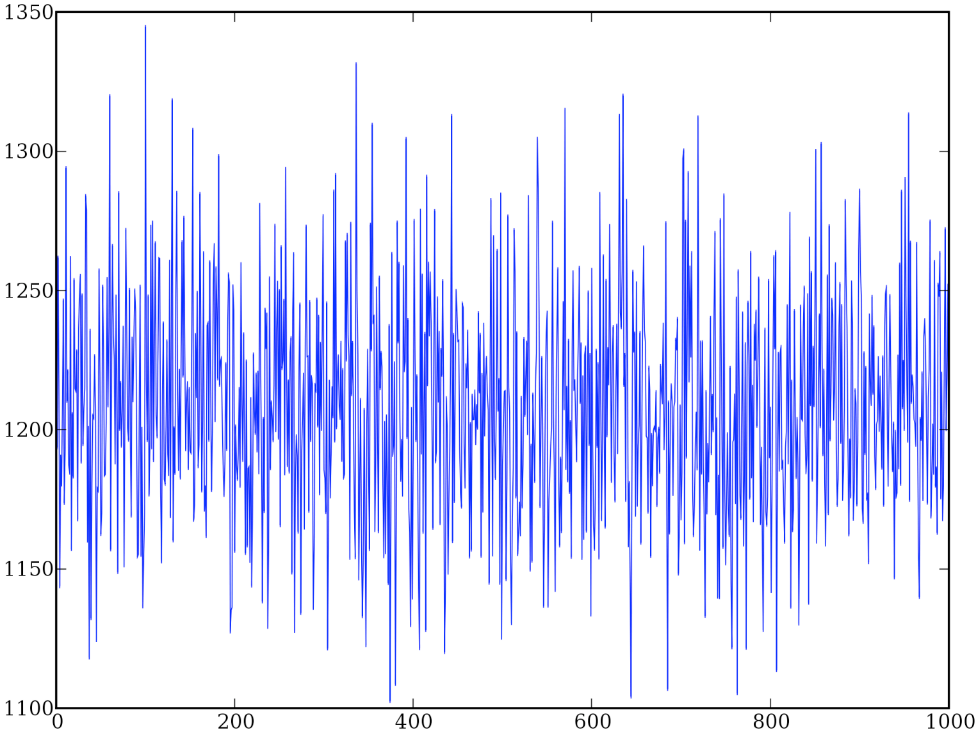
The first thing you'll need to work out is how many photons per second you would detect from this star, when it is not flaring. You are using a telescope with a 1m diameter mirror. On the telescope web page, you read that a 10th V magnitude star gives 400 counts per second (detected photons per second). You also learn that the read-out electronics introduces Gaussian noise into the output, with a standard deviation of 20 counts.

Your target star is 1.3 mag fainter than 10th mag, so should give $2.5112^{1.3}$ fewer counts per second: i.e. 120.8 counts per second, or 1208 counts per ten second exposure.

You write a computer program to simulate this. A copy can be found on the web page. You generate an array, and fill it up with random numbers. As we are talking about a discrete variable (the number of photons) these should be drawn from a Poisson distribution (using numarray's wonderful random number generation facilities). This gives a realistic simulation of how many photons you'd get in every ten-second read-out.

Now let's add an artificial flare. The brightness increases by 100% but only for one second. So over a whole 10 sec integration, the increase is only 10%. So lets increase the 100th array element by 1%. Set this to be a random number drawn from a Poisson distribution with mean 1208x1.01=1220.08.

Now we need to add the Gaussian electronic noise. Create another array, and fill it with Gaussian random numbers with mean 0 and standard deviation 20. Add it to the first array, and we have our simulated data. Here is what they look like (the first 1000 observations):



Note the spike at the 100th measurement. That's our flare: what we want to measure. It is the highest point in the first 1000 measurements, but only just, which is a bit worrying.

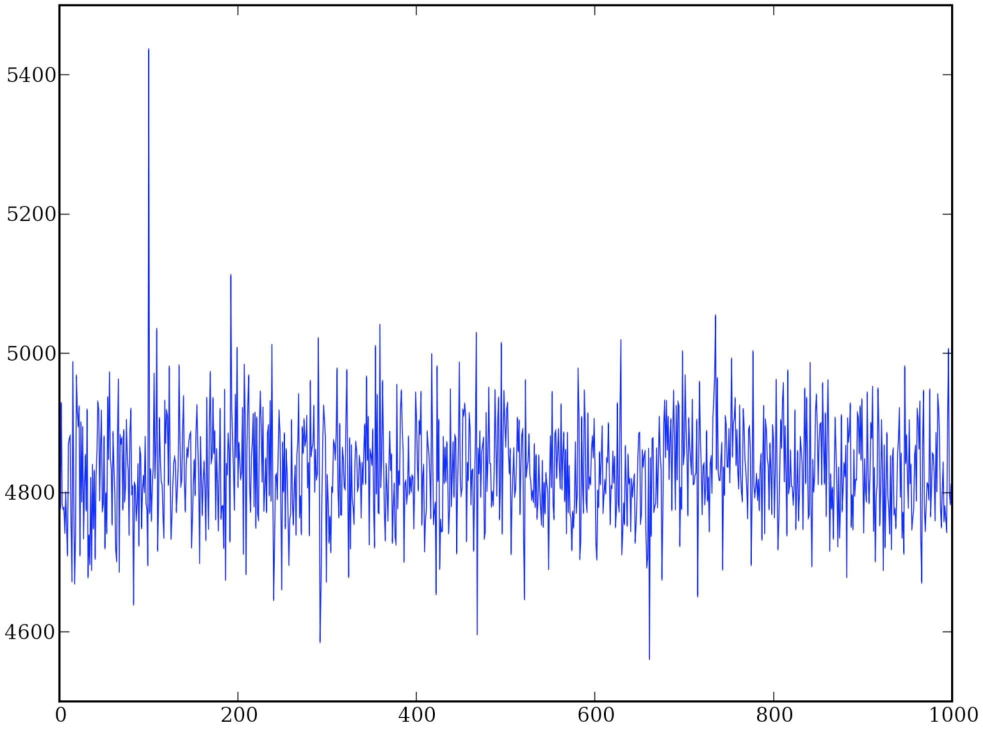
Now we need to decide what statistic we will use to find flares. We could look for any element in this array which has 1208x1.1 = 1328 counts or more. Let's try this. Observation 100 passes the test. But so does observation 336. So we've found the (simulated) real flare, but the noise is so bad we've mistakenly identified some noise as a second flare. And that's just in the first 1000 observations – we got 13 more spurious flares in the remaining unplotted observations.

And there is worse: we can run the program over and over again. It will generate a different set of random numbers each time. And we find that we always get around 14 spurious detections. But worse, we miss detecting the "real" flare half the time. That is because while its mean value is 1328 counts, the random arrival of its photons (as modelled by the Poisson distribution) means that half the time the actual detected value is lower than this. We could rectify this by changing our statistic: for example claiming to have detected a flare whenever an observation has more than (say) 1300 counts. This means that we will detect a larger fraction of the real objects. But we will also detect more bogus flares.

So right now, this doesn't seem like a feasible observation. We are finding lots of flares that aren't there, and missing several that are.

This could still be useful if, for example, we triggered follow-up observations on a bigger telescope whenever we thought we saw a flare. More than 90% of these follow-up observations would be wasted, but ~10% would be useful, which is still a better use of the big telescope time than using it all three nights.

But let us say we used a bigger telescope for all three nights. If it is a 4m telescope, all other things being equal, we would get 4^2 times as many photons per ten second exposure (as the collecting area of the telescope is this much bigger). So a mean of 4832 photons. Here is what our simulated data would look like now:



That's much more like it. The flare at observation 100 is now very clearly above the noise. We could set our detection threshold (the value measured in an observation at which we claim a flare) at somewhere like 5250, which should allow us to detect pretty much all the real flares, without getting any spurious ones. To check this, we run our simulation program ten times. All ten times it found the real flare, and only the real flare.

A similar approach can be used to simulate almost any observation. If, for example, you are trying to find faint galaxies, you could generate a fake image array. You could add fake galaxies to it. And then apply your favourite photometry program and try and see how many of these simulated galaxies you actually could recover. You could simulate spectroscopy: taking a model spectrum, adding appropriate noise and seeing if you could do your science with the resultant noisy spectrum.

That's the end of Part 1 of the Stats notes. In the next part, we'll look at some shortcuts which allow you to estimate what you can observe crudely but much faster. We'll cover how you add errors together, and model fitting.