Statistics Part 2: Combining Uncertainties

Background

In the last set of notes, I showed the most thorough way to estimate uncertainties: do a full end-to-end simulation of your observation. This is the most accurate method, but can also be very time-consuming.

In this set of notes, I show some shortcuts, which in many situations will save you a lot of time, and allow you to make quick estimates. But be warned – these shortcuts can be dangerous.

Assumptions

The assumptions we have to make to use these shortcuts are:

- All sources of noise follow the Gaussian distribution. If you have outlying points and don't get rid of them effectively, the short cuts discussed here will not work. Poisson distributions will also give you trouble, unless μ is large (> 10), in which case the Poisson distribution looks a lot like a Gaussian!
- When you apply an operation (such as multiplication or taking the log) to a variable with Gaussian noise, the resultant number is assumed also to have Gaussian noise. Can give you grief, as this is not always true! A classic example is converting fluxes into magnitudes, which involves taking a log. This works fine if the noise σ is much less than the mean flux. But if they are comparable, you can get near-zero fluxes, or even negative ones. Taking the log of these will give you near-infinite magnitudes!

If we make these assumptions, we can approximate the noise on any variable as a Gaussian, which is fully specified by its standard deviation s. Which is much simpler than needing the full probability density function for each variable.

The Procedure

1. Choose the statistic *x* you wish to compute – the one that tells you what you are scientifically interested in. It will in general be a function of the quantities *u, v, w* ... that you observe.
2. Work out what the uncertainty is in each of the observed quantities (σ_u, σ_v…)
3. Use the error propagation equation to work out what the uncertainty in your statistic is (σ_x).
4. Use your understanding of the Gaussian distribution to work out whether this predicted uncertainty is acceptable.

Why do you want to do this?

Here are some typical astronomical situations in which you'd want to go through this procedure:

- Writing a telescope proposal. You need to demonstrate what telescope you need to solve your scientific problem, and how much exposure time you require.

- Processing data: you need to understand how your data reduction algorithms affect the final data, so that you can optimise them.
- Testing new models: you may have come up with a new theoretical model. This procedure will tell you whether your model is or is not consistent with existing data.

Estimating your noise

Your first step is to estimate what the uncertainty is in each of the things you measure. As we discussed in the last set of notes, two major sources of noise are:

- The random arrival of photons.
- Electronic noise.

The noise caused by the random arrival of photons can be computed using the Poisson distribution. Remember – the standard deviation of a Poisson distribution of mean μ is $\sqrt{\mu}$. The amount of electronic noise depends on the details of the electronics – you normally will look this up in a manual.

You can (and should) check the amount of predicted noise. The usual way to do this is to observe nothing repeatedly (e.g. a blank bit of sky) and calculate the standard deviation of all these measurements – this is an estimate of the noise, and should agree with what you compute. If it doesn't, work hard to understand why!

Combining noise

Let us say that there is some statistic x we wish to compute. In most cases we do not directly measure x, instead we measure a number of other parameters $u, v, \ldots,$ and work out x using a function $x = f(u, v, \ldots)$. For example, x might be the absolute magnitude of a star, which is a function of u, the measured apparent magnitude of the star, and v, the distance to the star. Let us imagine that our measurement of each of $u, v, \ldots$ is afflicted by noise, and that the noise is Gaussian with standard deviations σ_u, σ_v, … It is fairly straightforward then to deduce the *error propagation equation:*

$$\sigma_x^2 \approx \sigma_u^2\left(\frac{\partial x}{\partial u}\right)^2 + \sigma_v^2\left(\frac{\partial x}{\partial v}\right)^2 + \ldots + 2\sigma_{uv}^2\left(\frac{\partial x}{\partial u}\right)\left(\frac{\partial x}{\partial v}\right) + \ldots$$

where σ_x is the standard deviation of the (assumed Gaussian) noise in the final statistic x. Note that the square of the standard deviation is called the **variance.** This is basically a Taylor series expansion, modified by the fact that standard deviations are computed by squaring values and adding them together.

The first two terms are straightforward: the noise in x depends on how much noise each of $u, v, \ldots$ have, and how strongly the value of x depends on that of $u, v, \ldots$ (the partial derivatives). The last term may need more explanation. σ_{uv} is the **covariance** of u and v, and is defined by the equation:

$$\sigma_{uv}^2 = \lim_{N\to\infty}\left[\frac{1}{N}\sum\left[(u_i - \bar{u})(v_i - \bar{v})\right]\right],$$ where $\bar{u}$ denotes the mean value of u.

If $u, v, \ldots$ are uncorrelated, then all the covariance terms are zero. If, for example, x is the absolute magnitude of a star, u its apparent magnitude and v its distance, then there is no particular reason why the noise in a given measurement of the distance should correlate with the noise in a given measurement of the apparent brightness. In a case like this, the covariance will sum to zero.

We can now use this equation to work out how to combine uncertainties in some common situations:

Additive Uncertainties

Let's imagine that x is just a weighted sum of u and v: ie.

$$x = au + bv$$

Doing the partial derivatives, we find that

$$\sigma_x^2 = a^2\sigma_u^2 + b^2\sigma_v^2 + 2ab\sigma_{uv}^2$$

If u and v are uncorrelated, the last term disappears, and we find that the noise in x in the square root of the sum of the squares of the uncertainties in au and bv. This is known as **adding in quadrature.**

One special case – let's say you want measure a particular noisy quantity over and over again. For example, you want to obtain a really deep image of some part of the sky. The Hubble Deep Field observations are an example. To detect really faint galaxies, and exposure hundreds of hours in length was needed. But Hubble cannot expose for more than ~ 20-40 minutes at a time without the signal being buried in cosmic rays. So let's say you want to detect a particular galaxy, which is expected to produce a signal s. Each 20min image, however, has noise σ which is much greater than s.

The way you get such a deep image is to add all these short images together, pixel by pixel. Let's say we add n images together. The signal in the final sum image will just be ns. If the noise also increased by a factor n, taking lots of images wouldn't help you. But using the above equation, we find that

$$\sigma_{sum}^2 = \sigma^2 + \sigma^2 + \sigma^2 + ... = n\sigma^2$$

so

$$\sigma_{sum} = \sigma\sqrt{n}$$

Thus the noise increases as the *square root* of the number of exposures: i.e. less rapidly that the signal. Thus the signal-to-noise ratio *increases* as the square root of the number of exposures.

This is the √n rule: you can get a long way with it. Noise goes up as √n. Indeed, it also applies within a given exposure: remember that the standard deviation of a Poisson distribution is proportional to the square root of the mean μ. So as you expose for longer, then mean number of photons you expect is proportional to the exposure time t, and hence the noise is proportional to √t.

Some people get confused by the difference between σ_{sum} and σ. You measure σ by looking at how wide a range of values you measure. σ_{sum} is the standard deviation you get in the sum – i.e. if you obtained n values and computed the average, and did this again and again and again, you'd get lots of different sums: σ_{sum} is a measure of how wide a range of numbers you'd get amidst these sums.

A word of caution: this √n rule can get you a long way, but ultimately it usually breaks down. It relies on the assumption that noise is Gaussian and that all the different sources of noise are uncorrelated with each other. In practice, this is seldom true. Or even if it is, other sources of noise come in to play when you combine lots and lots of data.

For example, let's say you wanted to use the AAT to get an image as deep as the Hubble Deep Field – ie. going to 30th magnitude. The AAT can get an image going to 26th magnitude in about an hour. So if we want to get to 30th magnitude, the signal will be

$2.5112^{(30-26)}$=39 times weaker. So to detect these objects, we will have to decrease our noise by a factor of 39. Using the √n rule, this means that instead of a single 1 hour exposure, we would need to take 39^2 exposures – ie. 1521 exposures, i.e. 6 months worth of clear nights!

But would the result really be an image reaching 30th magnitude? Almost certainly not. Because there are two types of noise – the random noise (which obeys the √n rule) and systematic errors. Such as confusion noise, scattered light, flat fielding errors, weak cosmic rays, unstable biases etc. The systematic errors are relatively small, but because they are systematic, they can be present in the same way in every image (rather than being uncorrelated) and hence they do not obey the √n rule and diminish. Eventually they will dominate, and then taking more exposures won't help you.

Back to uncertainty addition: remember that the uncertainty in a statistic is calculated from the sum of the *squares* of the individual uncertainties. This has an important consequence – if there are several sources of noise, the biggest source of uncertainty is much more important than the rest. If, for example, you are trying to work out the age of an elliptical galaxy, based on the strength of some absorption lines in its spectrum, which you compare to a theoretical model. The model depends on the metallicity, which you do not know very accurately. Let us say that the uncertain metallicity introduces a 10% uncertainty in the age, while the noise on your measurement of the emission lines introduces a 7% uncertainty. The total uncertainty is thus $\sqrt{(7^2+10^{2)}} = \sqrt{149} = 12.2\%$ (not 17%). Now let's say you want to improve this measurement, to discriminate between two theories of elliptical galaxy formation. You could improve your measurement of the metallicity, which would drop this uncertainty by 4% (from 10% to 6%), or you could get a better measurement of the absorption lines, which would drop this uncertainty by 4% (from 7% to 3%). Which should you do?

Working on the metallicity drops your final uncertainty from 12.2% to $\sqrt{(7^2+6^2)}$ = 9.2%. Working on the absorption lines, however, only drops it to $\sqrt{(3^2+10^2)} = 10.4\%$. So the motto is – always find out what your biggest source of uncertainty is and work hard on reducing that: the gains from reducing the small contributors to the uncertainty are relatively tiny.

Optimal Weighting.

Let us imagine that we have lots of measurements x_i of a particular statistic, each with an associated uncertainty σ_i. For example, we might have dozens of different measurements of the Hubble Constant made by different research groups, or we might have taken lots of images of a given part of the sky. We want to deduce the best possible value of x, i.e. the value with the smallest uncertainty, by combining all these different x_i values.

You might think that the obvious thing to do is to average (take the mean) of all the different estimates. And you would be right, *provided that all the different measurements had the same uncertainties.* But what if some of the measurements were better than the others? Surely you'd want to give them more prominence? Averaging bad data with good won't produce very pretty results. One approach would be to just take the best measurement and throw the rest away. But surely there is some information in all the worse measurements – they may have larger errors, but if the uncertainties are not THAT much larger, they should still be able to contribute.

The answer is to do a *weighted* average. So instead of the standard mean:

$$\bar{x} = \frac{1}{n}\sum x_i$$

we compute a weighted mean:

$$\bar{x} = \sum w_i x_i$$

where w_i is the weighting of data point number i. To be a true average, we require that $\sum w_i = 1$.

The better data points should be given larger weights, and the worse ones smaller weights. But how exactly should these weights be chosen? This is a fairly straightforward calculation using the error addition equation: it turns out that the optimum strategy is *inverse variance weighting* – ie. the weight applied to a given point should be inversely proportional to its variance. The variance, remember, is the standard deviation squared. So the optimum weights are given by

$$w_i \propto \frac{1}{\sigma_i^2}$$

with the constant of proportionality set by the requirement that the sum of the weights be 1.

This is very widely used in astronomy, and can make a big difference. But, as usual, be careful. This assumes Gaussian independent errors. While, in principle, adding crap data to good, albeit with a low weight, will improve your result, it is often worth drawing the line at the most crap data – you are probably adding in some low-level systematic errors which will more than outweigh the benefits.

Multiplying Uncertainties

If x is the weighted sum of u and v,

$$x = auv$$

then using the error propagation equation, we find that

$$\left(\frac{\sigma_x}{x}\right)^2 = \left(\frac{\sigma_u}{u}\right)^2 + \left(\frac{\sigma_v}{v}\right)^2 + \frac{2\sigma_{uv}^2}{uv}$$

As usual, the last term disappears if the uncertainties in u and v are independent, in which case the equation simply says that the percentage error in x is equal to the sum of the percentage errors in u and v.

You can derive similar equations for different functions – e.g. when your statistic is the log, exponential or sine of the measured quantities: see a stats textbook or derive them yourself from the error propagation equation. But in practice, adding and multiplying are the most useful equations.

What does your final uncertainty mean?

OK – let's say you've done all this maths, and computed the uncertainty in your statistic. What does this mean?

This all depends on what you wanted the statistic for in the first place. As always, you have to be clear about this. Let's look at a couple of examples:

- You want to measure the luminosity function of galaxies down to a particular magnitude limit. How long do you need to expose for on a given telescope? In this case, your statistic might be the signal-to-noise ratio achieved for a galaxy at your magnitude limit. What value must this have? It must be large enough that you are detecting most of the galaxies with this brightness, and not detecting large numbers of spurious objects due only to fluctuations in the noise.
- You want to see whether dark energy could actually be made of vegemite. The vegemite model predicts that supernovae at redshift 0.7 have a particular brightness. Your statistic is the mean brightness of observed supernovae at this redshift. You want to compare this measured mean value against the prediction of your model and see whether your model could be true. So you will look at the difference between the observed mean brightness and your model and compare it to the uncertainty: are they close enough together that the difference could just be a statistical fluke, or so far apart that your model is consigned to the dustbin of history?

A note on terminology: when people quote uncertainties in the literature, they generally write something like q = 0.7 ± 0.2. This means that the uncertainty is 0.2. But what does this mean? The most common usage is implicitly assume that the probability distribution is Gaussian, and to quote the standard deviation as the error: so in this case, if you measured q repeatedly, you would find a Gaussian distribution of mean 0.7 and standard deviation 0.2. But this can be confusing to novices – they might assume that q must always lie between 0.5 and 0.9. Whereas in reality you would expect a number drawn from a Gaussian distribution to lie beyond one standard deviation from the mean roughly 32% of the time (first set of notes). So people sometimes quote not the standard deviation, but the range within which 95% or 99% of measurements are expected to lie. Which can cause grief if you misinterpret what they mean.

So – let's say you predict that a given statistic x has the value x=1.0± 0.2. What does this mean? What it means is that if you measure x over and over again, you will get different answers each time. And these answers will follow a Gaussian distribution of mean μ=1.0 standard deviation σ=0.2.

You can work out how after a given value will be obtained by integrating the Gaussian numerically over various ranges. You expect x to lie within ± one standard deviation of the mean 68.3% of the time, and to lie within ± two standard deviations 95.4% of the time. The fraction at other values is listed below:

Number of standard deviations from the mean	Fraction of the time that a measurement will lie within this many standard deviations of the mean
0.5	38.3%
1.0	68.3%
1.5	86.6%
2.0	95.4%
2.5	98.76%
3.0	99.63%

3.5	99.95%
4.0	99.994%
4.5	99.9993%
5.0	99.99994%

How do you use these values? Some examples will hopefully help.

Example 1: an image.

Imagine that you want to take an image of some part of the sky. You are taking it with a CCD camera, which has a known rms read-out noise of 5.3 counts (rms = root mean squared deviation = standard deviation). It has 1024x1024 pixels. Based on your knowledge of the sky brightness, you expect about 10 counts (detected photons) per pixel per second, and you plan to expose for 100 sec.

Thus, in the absence of any objects, you would expect to get 1000 counts per pixel. As the arrival of photons is a random, quantum mechanical process, you won't get exactly 1000 in each pixel. Instead, you will get a Poisson distribution of mean 1000. The standard deviation of a Poisson distribution is the square root of the mean – i.e. 31.6 counts. The Poisson distribution looks more and more like a Gaussian as the mean increases – for a mean of 1000 this is a pretty good approximation.

So if you have a bunch of pixels containing no objects, the mean number of counts will be 1000, and the standard deviation will be the quadrature sum of 31.6 (the photon Poisson noise) and 5.3 (the read-out noise) – ie. 32.06 counts. Thus the read-out noise, being much smaller than the photon noise, is essentially irrelevant (a common situation in broad-band optical/IR imaging).

For simplicity sake, let us assume that the pixels are pretty big, so that all the light from any object will fall in a single pixel. To find all the objects in our image, we must set a threshold value *t*. Any pixel that exceeds *t* will be listed as a detected object. What value of *t* should we choose?

We now need to think about what we are trying to achieve scientifically. Normally our scientific goal is to avoid abject humiliation when we publish a paper based on this image. To do this, we need to make sure that when we list all the objects we saw in our image, the objects in this list should be real, and not just empty pixels where the noise has resulted in an unusually high value. So that if someone else tries to get follow-up observations of one of our objects, they will find that it is really there, and they haven't wasted their time looking at a blank bit of sky.

This means that we should set out threshold *t* high enough that we get few if any spurious detections.

On the other hand, we want *t* to be pretty small, or we will be throwing away faint objects.

So what value of *t* to use? This depends on how sure you want to be that there are no spurious detections. Let's say you want to be really sure that every object you claim to have detected is really there. In this case, you want the expected number of spurious detections to be much less than 1. You have 1024x1024 ~ 1 million pixels, so you want the probability of a pixel which contains only empty sky having a measured value greater than *t* to be less than one in a million. Looking at the above table, you see that 7×10^{-6} of

the time you get results more than 4.5 standard deviations out, while for 5.0 standard deviations, this probability drops to 6×10^{-7}. These are the probabilities of getting a value either this much higher than, or this much lower than the mean. A pixel that is anomalously low will not generate a spurious detection – it's only the high ones that are an issue, so we can divide these probabilities in half.

So to have an expected number of spurious detections less than one, you'd need to set the threshold to be five standard deviations above the mean: 4.5 standard deviations isn't quite good enough. Thus your threshold should be 1000.0 + 5.0x32.06 = 1160.03. This is called "setting a five standard deviation threshold", or a "five sigma threshold".

Alternatively, you might be prepared to tolerate a few spurious points. For example, you might be trying to work out how many galaxies there are down to some magnitude limit. You expect to find 7000 galaxies in this image. So it wouldn't really matter if, say, 100 were spurious. You could work out the expected mean number of spurious galaxies and subtract it from the number you actually detected – this won't be perfectly accurate, but an error in the 100 (typically the square root of 100 – i.e. 10) won't make that much difference to the 7000 galaxy counts.

In this case, you need the expected number of spurious detections to be 100 – i.e. the probability of any one of the million pixels having a value above the threshold must be 10^{-4}. Looking at the table, this occurs at 4.0 standard deviations above the mean, so our detection threshold need only be 1000.0 + 4.0x32.06 = 1128.24. We can thus detect galaxies that are 20% (0.2 mag) fainter. A four sigma detection threshold.

If we desperately wanted to detect these 20% fainter galaxies, but were more fastidious and hence unwilling to tolerate that any of our detections might be spurious, we would need to increase our exposure time. The signal from a given galaxy is proportional to the exposure time, while the noise (being dominated by the Poisson photon noise) is proportional to the square root of the exposure time. Thus the signal-to-noise ratio of a given galaxy (the number of standard deviations it is brighter than the mean) is inversely proportional to the square root of the exposure time. Thus to detect galaxies 20% fainter with five sigma confidence, we'd need to increase the exposure time by a factor of $1.2^2=1.44$.

Example 2: Testing a theory.

Let us imagine that a particular theorist (Prof Zog) has run a supercomputer simulation of the formation of Milky-Way-like galaxies, using cold dark matter (CDM). He finds that such galaxies should have a mean of 630 dwarf galaxies in orbit around them. However, the exact number depends on the merging history of the galaxy: he ran his simulations 100 times and found dwarf galaxies ranging from 112 to as high as 954. He found a standard deviation of 243 dwarf galaxies around the mean of 630.

You have just completed a survey of 10 nearby Milky-Way-like galaxies, counting the number of dwarf galaxies. You found the following numbers: 523, 12, 144, 15, 3, 44, 320, 2, 0, 97.

What you want to know is, can you publish a paper saying that Prof Zog's simulation is inconsistent with the data? Or will all the CDM mafia pillory you if you do this?

Your data have an average (mean) of 116, and a standard deviation of 165. Is this consistent with Prof Zog's simulation?

You can never prove a theory correct - all you can do is prove a rival theory wrong. In this case, we are trying to discredit Prof Zog's theory. So let's assume this theory was correct, and work out the consequences. If these consequences are inconsistent with our observations, we can say "Zog is an idiot" and get away with it…

So let us say that Zog's theory was correct. There should thus be 630±243 dwarf galaxies around each of our target galaxies. If we assume a Gaussian distribution, we can then ask, for each of our galaxies in turn, what are the odds of seeing this, were the theory correct? Our first galaxy had 523 dwarfs. This is only (630-523)/243=0.44 standard deviations away from the prediction. Using the table, you'd expect over 60% of observed galaxies to be at least this different from the mean. So this could well be a fluke and doesn't disprove anything. But what about the second galaxy, with only 12 dwarfs? This is now (630-12)/243 = 2.5 standard deviations away. Only 1.3% of galaxies are expected to be at least this far from the mean.

Could you publish given just this one galaxy with 12 dwarfs? It's a bit marginal. Just from random fluctuations, you'd expect more than 1% of galaxies to be this different from the predictions. So you can say "this is fairly unlikely, given Zog's theory". But if there were lots of papers trying to disprove lots of theories using 2.5 sigma discrepancies, more than 1% of them would be wrong, and thus several theories would have been unjustly pilloried.

And this is assuming nice Gaussian distributions: in the real world, unusual evens are typically more common than this nice Gaussian integral tells us. So normally any paper based on 2 sigma statistics is regarded with caution.

In this case, however, we have lots more galaxies. None of them are more than 3 standard deviations below the prediction, but they are **all** low, which must be telling us something. If Zog's theory was correct, what are the odds of all ten galaxies giving such low values?

One way to do this would be to work out the probability of seeing a result this low for each galaxy in turn, and then simply multiplying the probabilities together. You don't have to multiply too many ~1% numbers together until you get a very small probability.

Another quicker way is to look at the mean number of dwarf galaxies you measured. This was 116. What is the uncertainty in this? Well, the standard deviation between the measured galaxies was 165. The standard deviation in the sum is thus $165 \times \sqrt{n}$, where n=10. The mean is simply the sum divided by n, so the standard deviation in the *mean* is $165/\sqrt{10}=52.2$. So if you measured lots of different sets of ten galaxies, and computed the mean for each, you estimate that these means would form a Gaussian distribution of standard deviation ~52.

Similarly, Zog's theory predicts that if we measured 10 galaxies, and averaged the number of dwarfs found, the mean would be 630 and the standard deviation *of the mean* would be $243/\sqrt{10}=76.84$.

So is our observed mean of 116±52 consistent with the predicted value of 630±77? Another way of looking at this is to say: the difference d between these means is d=630-116=514. The uncertainty in d can be estimated using the error addition equation: it is the quadrature sum of the uncertainties in the two individual means – i.e. 93. So d=514±93.

If our data are consistent with Zog's model, then we would expect d=0. If this were the case, and our uncertainty estimate is accurate, then our observed value lies 514/93=5.5 standard deviations away from the prediction. From the table, less then one

in a million observations should be thus discrepant. Thus if all the astronomy papers ever published used 5.5 sigma results, on average none of them would be in error. So this is pretty conclusive: Zog's theory is cactus.

Conclusions

The material covered in these notes is **very** widely used. People often talk colloquially about "that's only a 2 sigma result- I don't believe it", or "This catalogue is split into two parts, the 5 sigma detections and the 4-5 sigma ones which should be used with caution" This approach is an approximation, only valid if everything is nice and Gaussian and uncorrelated, which is seldom perfectly the case in the real world.

So be careful! Look at histograms of your data to see if they really are Gaussian. Err on the side of caution in choosing at what "sigma" value to publish. And if things are really crucial, do a full end-to-end Monte-Carlo simulation as discussed in the previous set of notes.

But for many applications, this quick and dirty approach is perfectly adequate, and much quicker than the alternative!