Poisson Likelihood and Chi Squared

In the several lectures following this one we will go over some aspects of Bayesian statistics. To set that up we will do two things in this lecture: first, we will discuss the Poisson distribution and Poisson likelihood, given their importance in astronomy, and second, we will talk about the calculation and use of χ^2, which is often used in astronomy without a proper consideration of whether it applies in a given circumstance.

The Poisson Distribution

In the Poisson distribution, if the expected number of discrete "events" is m (which can be any positive real number), then the probability of observing a non-negative integer number d of events is

$$\text{Prob}(d|m) = \frac{m^d}{d!}e^{-m} \ , \tag{1}$$

where $\text{Prob}(d|m)$ means "the probability of d given m". Here "events" are things that come in discrete packages, such as photons arriving at a detector, radioactive decays, calls to a radio show, or a host of other examples. The Poisson distribution is the correct distribution if the following conditions hold (see https://en.wikipedia.org/wiki/Poisson_distribution for more details):

1. Events can be counted as integers: there can be $0, 1, 2, \ldots$ events, but not 0.7, 12.2, or other non-integer numbers of events. Thus measurements of continuous quantities (say, the maximum temperature in a given location on a sequence of days) could not be treated using the Poisson distribution, but numbers of radioactive decays could if they satisfy the next criterion.
2. The events are *independent* of each other. Thus, for example, having a photon arrive in one time interval cannot make it more or less likely that another photon arrives in the next time interval if the Poisson distribution is to be correct. This is often the case, but not always; for example, given that a detector takes a finite amount of time to read out a photon, if the photons arrive so frequently that the time to the next photon is comparable to or less than the readout time, the readouts will *not* be independent of each other even if the actual photon arrival times really are independent.

Sometimes, as in the Wikipedia page, there is another assumption listed: that the rate of events is constant. This, however, is for a slightly different application, which is the question of the distribution of event totals in many consecutive intervals of the same length. Here we ask only about the probability distribution of the event totals in a given segment of data.

Figure 1 plots the probabilities for a Poisson distribution for different expectation values (also known as arithmetic means) $\langle x \rangle$ (for a Poisson distribution, $\langle x \rangle = m$), and against

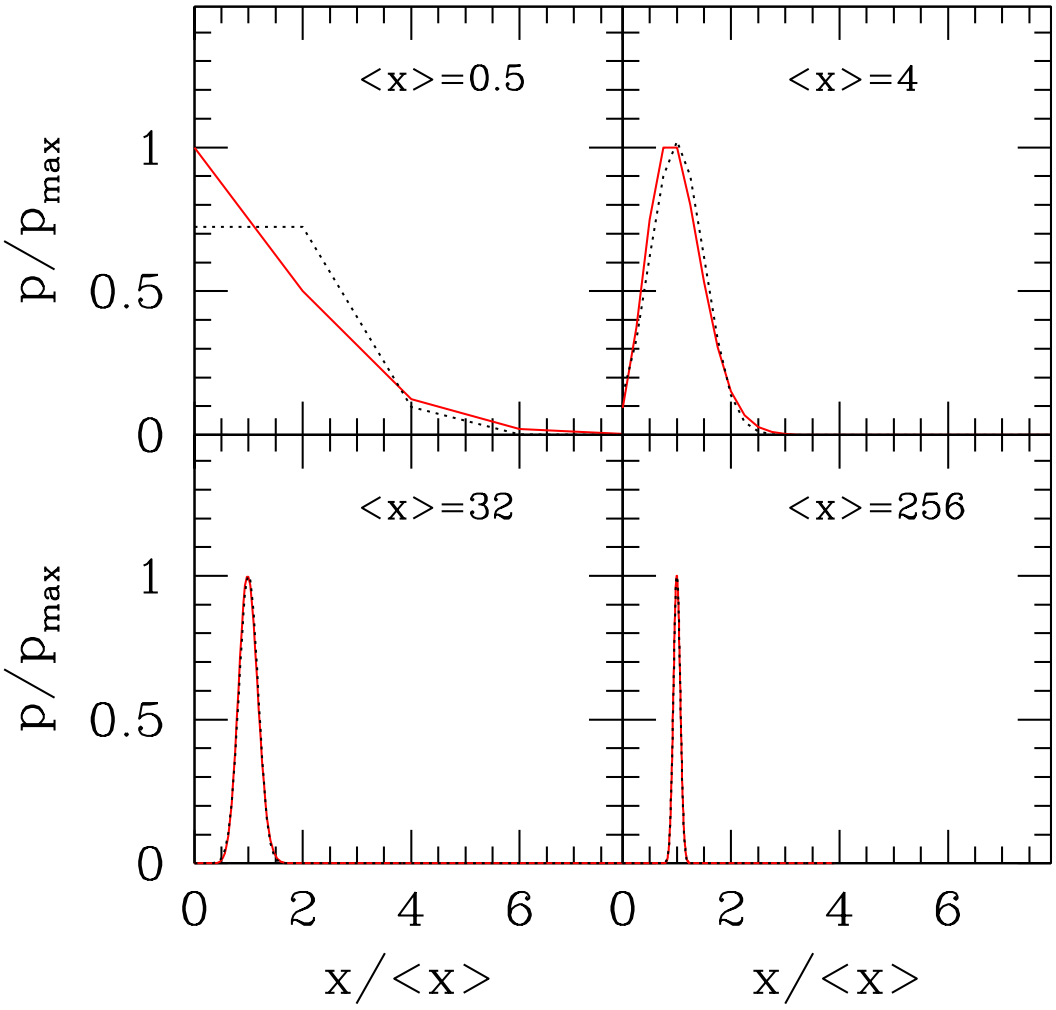

Fig. 1.— Poisson distributions (solid red lines) versus Gaussians with the same mean and variance (dotted black lines). In each case, the vertical axis is normalized to the peak probability in the Poisson distribution, and the horizontal axis is normalized to the expectation value $\langle x \rangle$. Top left: $\langle x \rangle = 0.5$. Top right: $\langle x \rangle = 4$. Bottom left: $\langle x \rangle = 32$. Bottom right: $\langle x \rangle = 256$. We see that for larger $\langle x \rangle$, the Poisson distribution is closer to a Gaussian.

Gaussians of the same expectation value and variance (which is also m for a Poisson distribution). We see that for small $\langle x \rangle = m$, the Poisson distribution differs significantly from a Gaussian, but for larger $\langle x \rangle$, the distributions become progressively more similar to each other near the peak.

Given the assumptions of discreteness and independence, the Poisson distribution can be derived as a limiting case of a binomial distribution. We do this in the appendix at the end of the lecture, for the mathematically curious.

Chi squared

We now turn to the χ^2 statistic, which is ubiquitous in astronomy. It has almost magical properties: when it is applicable, use of χ^2 is advertised as being able to tell you (1) whether a model is a good fit to data in an absolute sense, (2) whether one model is better than another model, at least when the models are nested (i.e., one model is a special case of the other), and (3) the best values and uncertainties in each of the parameters of the model. With such power and the ease of use that we will describe, no wonder it is such a favorite! But as always, we need to think carefully about whether χ^2 *is* applicable in any given circumstance.

First, we'll describe it. There are actually some variants in the way to calculate χ^2, but we'll use the most common one (Pearson's χ^2). Suppose that we have i bins of data. These might be bins in wavelength, or time, or something different. Suppose that we observe d_i counts in bin i, and our model predicts that there will be m_i counts in that same bin. Summing over all bins, then

$$\chi^2 = \sum_i \frac{(d_i - m_i)^2}{\sigma_i^2} \ , \tag{2}$$

where σ_i^2 is a variance. In Pearson's formulation, $\sigma_i^2 = m_i$, because an important implicit assumption in the whole χ^2 formalism is that d_i and m_i are both large enough that a Gaussian distribution is appropriate.

Suppose we consider a single bin, so that $\chi^2 = (d-m)^2/m$. If the actual probability distribution of d given m is a Poisson distribution, so that

$$\mathrm{Prob}(d|m) = \frac{m^d}{d!} e^{-m} \ , \tag{3}$$

then because $\sum_{d=0}^{\infty} \mathrm{Prob}(d|m) = 1$ (i.e., the Poisson distribution is normalized properly), then the expectation value of any function $f(d)$ is given by

$$\langle f \rangle = \sum_{d=0}^{\infty} f(d) \mathrm{Prob}(d|m) \ . \tag{4}$$

For example, the expectation value of χ^2 is

$$\langle \chi^2 \rangle = \sum_{d=0}^{\infty} \frac{(d-m)^2}{m} \frac{m^d}{d!} e^{-m} = 1 \ , \tag{5}$$

where the sum can be evaluated using, e.g., Wolfram Alpha. Thus the arithmetic mean of χ^2 for a single bin is *always* 1, *regardless* of the value of m! That's a surprise.

Similarly, we find that

$$\langle (\chi^2)^2 \rangle = \sum_{d=0}^{\infty} \left[\frac{(d-m)^2}{m} \right]^2 \frac{m^d}{d!} e^{-m} = 3 + 1/m \ , \tag{6}$$

where again the final evaluation can be obtained from a symbolic manipulation program. This means that the variance in χ^2 is $\langle (\chi^2)^2 \rangle - \langle \chi^2 \rangle^2 = 2 + 1/m$. The chi squared tables you see are all constructed in the limit $m \to \infty$; in such tables, you look up the number of bins minus the number of parameters (this difference is called the *number of degrees of freedom*), and then from the tables you can judge how probable it would be, *if the model is correct*, that you would by chance have a value of χ^2 as large or larger than what you found for that number of degrees of freedom. Therefore, if you blindly compute χ^2 and look up the probability in the table, but m isn't big enough, you'll mislead yourself.

This, as I say, is how χ^2 *should* be used: by using the variance of the model in the denominator. But in most cases I've seen in astronomy, it is the variance of the *data* that is put in the denominator! For example, in the manual for XSPEC (the default fitting program in astronomy for X-ray data) we read: "...in general, we do not know the true variance and have to estimate it... The default option (weight standard) is to use the observed number of counts as an estimator for the underlying variance (equals the underlying mean)." This is incorrect! The authors of XSPEC realized this, so they followed with "It is important to realize that this introduces a bias. Downward fluctuations will be weighted more heavily than upward fluctuations because, while the numerator of chi-squared for the bin will be the same, the denominator will be smaller for the downward fluctuation. An obvious alternative to try is to use the predicted counts from the model as an estimator for the Poisson variance (weight model). This does not have the bias problem of the standard method however in practice it turns out to be unstable and can drive the fit away from the best parameters."

Unpacking this, it means that unless you specify otherwise, in XSPEC (and, as it turns out, in a lot of other astronomy data analysis packages) you implicitly assume that the *data* have variance; this is not the way the test was designed, and it produces bias. Even if you are alert enough to correct this and use the model variance, you have problems. Note, by the way, that for data that do not come in numbers of counts (e.g., galaxy observations are more likely to be in magnitudes), astronomers will often make a separate estimate of the uncertainty, which they usually associate with the data rather than the model.

If you swallow hard and ignore these potential problems, how should you use χ^2?

Let's suppose that your model (which produces the expected values m_i for each bin) has some number of parameters; call that number k. Suppose also that there are n total bins. Then as we say above the number of *degrees of freedom*, or dof, of the data is $n - k$. Adjusting the parameter values in your model to *minimize* χ^2 gives you the best fit. If the fit is good, then you expect the minimum chi squared to be approximately equal to the number of degrees of freedom.

What if your minimum χ^2, $\chi^2_{\rm min}$, is *not* close to $n - k$? Then:

1. If $\chi^2_{\rm min} \gg n - k$, then you have a bad fit. Something is missing from your model.
2. If $\chi^2_{\rm min} \ll n - k$, it does *not* mean that you have a fantastic model. Instead, it means that you have overestimated the uncertainties (i.e., your σ_i^2 are too large).

Many astronomers do not appear to understand the second point: I've had discussions with researchers who think it's fine to have a tiny χ^2. No, it's an indication of another problem.

It is, of course, possible to be more quantitative than simply comparing χ^2 with $n - k$. There are tables you can check, and websites that compute the probability that if the model is right you will have a χ^2 as large or larger than what you saw (and even relatively simple formulae). A decent rule of thumb, if both χ^2 and $n - k$ are large enough, is that $\sqrt{2\chi^2}$ is distributed roughly as a normal distribution with a mean of $\sqrt{2(n-k)-1}$ and unit variance (result proven by R. A. Fisher in 1922).

But you can do even more with χ^2. Suppose you have two models that are *nested*, in the sense that one contains the other. For example, maybe you are fitting a spectrum. One model of the spectrum is that it is just flat, but in the other model the spectrum is flat except for a small portion that has a Gaussian emission line. The line model contains the flat model (because the amplitude of the Gaussian could be zero), but it has three extra parameters (the amplitude, centroid, and variance of the Gaussian). Clearly the minimum χ^2 of the line model can't be any larger than the minimum χ^2 of the flat model, but how much smaller does it be so that you will accept the more complicated model?

Swallowing caveats for the moment (and there are plenty!), the answer is that you take the *difference* $\Delta\chi^2$ between the minimum chi squared of the flat model, and of the line model, and then compare that difference with a number of degrees of freedom that in this case is the number of extra parameters. In our case, we could look up that table to find that for three degrees of freedom, an improvement of 6.251 in χ^2 would cause us to favor the more complicated model at the 90% confidence level, and an improvement of 11.34 in χ^2 would cause us to favor the more complicated model at the 99% confidence level.

This all sounds very straightforward and easy. Indeed, this is a reason why χ^2 is a workhorse of astronomical data analysis. As long as you take care not to have too few counts in your bins, you're set!

Or are you? Remember, our goal in this course is to *think* about our results and methods rather than just applying recipes blindly. If you're very alert, you might have noticed that when we considered the problem of determining whether there is good evidence for a Gaussian line, we said nothing about the range of parameter values we plan to consider. But, intuitively, that *should* matter. If we would be happy for our line to be anywhere in the vast range of wavelengths that we have in our spectrum, it should be easier for some random bump in the spectrum to be interpreted as a line than if we instead focused our attention only on a very narrow portion of the spectrum. But the $\Delta\chi^2$ approach we described doesn't take that into account at all. That's a pretty significant flaw in the method!

Another flaw is more insidious. Many methods such as χ^2 implicitly assume that you have so many counts in bins (or whatever you're measuring) that you can treat the statistics as locally Gaussian (in our case above, that means that we can ignore $1/m$ compared with 2 in the variance). Lots of astronomical data analysis packages are arranged using those assumptions. But sometimes you have enough resolution in your data that many bins have only a few, or even no, counts. Then, you can't use methods that assume Gaussians.

The *right* thing to do is to then use other methods. But I've had countless conversations with astronomers who try to convince me that the proper thing to do is to group the bins until the bins each have lots of counts. They want to do this because it means that their Gaussian-based methods (such as χ^2) can work in that limit. But as we said before, grouping in this way loses information. Perhaps the information isn't important, but you should *not* make specious arguments to justify incorrect assumptions!

With all of that as background, we will now, in the next several lectures, talk about a more rigorous way to make statistical inferences: Bayesian statistics.

Appendix: Derivation of the Poisson distribution from the binomial distribution

In the first part we follow closely a nice derivation posted by Andrew Chamberlain at https://medium.com/@andrew.chamberlain/deriving-the-poisson-distribution-from-the-binomial-distribution-840cc1668239.

Suppose that we have n very narrow contiguous intervals (of time, or wavelength, or ...) that together make up the bin of interest. The total expected number of events in the n intervals combined is m. Suppose for this derivation we assume a constant rate as well as independence. Then, as a result, the expected number of events in any *one* of the intervals is $p = m/n$. We are interested in the limit $n \to \infty$; that is, we take a bin of observation

(e.g., one second, or 100Å) and divide it into a very large number n of observations. In that limit, $p = m/n \to 0$ because m is fixed while $n \to 0$. Thus we can assume that in a given interval, there can be one event, with probability p, or zero events, with probability $1 - p$.

What is the probability that, in those n intervals, we will see $d \ll n$ events? If we cared about the order (e.g., if event, no event, no event, event was different from no event, event, event, no event and similar combinations), then the probability would be $p^d(1-p)^{n-d}$ (i.e., the probability that those particular d intervals had events while the other particular $n - d$ intervals did not have events). However, we don't care about the order; we only care about the total. There are, in fact, $\binom{n}{d} = \frac{n!}{d!(n-d)!}$ ways to have d events in n intervals. Thus the probability that we would observe d events in the n intervals is given by the binomial distribution:

$$\text{Prob} = \binom{n}{d} p^d (1-p)^{n-d} \ . \tag{7}$$

We now expand this out, starting with $\binom{n}{d}$:

$$\binom{n}{d} = \frac{n!}{(n-d)!d!} = \frac{n \times (n-1) \times \ldots \times (n-d+1)}{d!} \approx \frac{n^d}{d!} \ . \tag{8}$$

In the last step above, we take advantage of our limit $n \to \infty$ to realize that because $n \gg d$, all the factors in the numerator (n, $n-1$, ..., $n-d+1$) are very close to n.

Thus the first two factors in Equation (7), $\binom{n}{d} p^d$, become

$$\binom{n}{d} p^d \approx \frac{n^d}{d!} p^d = \frac{n^d}{d!} (m/n)^d = \frac{m^d}{d!} \ . \tag{9}$$

For the last factor, and using the substitution $x \equiv 1/p$, we get

$$(1-p)^{n-d} \approx (1-p)^n = \left(1 - \frac{1}{x}\right)^n = \left(1 - \frac{1}{x}\right)^{m/p} = \left[\left(1 - \frac{1}{x}\right)^x\right]^m = e^{-m} \ . \tag{10}$$

Here we use the identity that as $x \to \infty$, $(1 - 1/x)^x \to e^{-1}$.

Thus, putting everything together, we find

$$\text{Prob} \to \frac{m^d}{d!} e^{-m} \tag{11}$$

in the limit $n \to \infty$.

This proves that the Poisson distribution is the limit of the binomial distribution, but we need to think about what we have shown. We have shown that if we expect m counts (where, again, m is a positive real number), the probability that we observe d counts (where d is a non-negative integer) is $(m^d/d!)e^{-m}$. This does *not* require that we have infinitely

fine resolution! The supposition of infinitely fine resolution, and the use of the binomial distribution, were only intermediate steps along the way to our derivation.

We have also, as part of our derivation, assumed that the rate of events is constant, but this is not necessary. As is shown formally in, e.g., http://www.randomservices.org/random/poisson/Nonhomogeneous.html, if the rate of events varies with time (or wavelength or whatever your independent variable is), it is straightforward to do a transformation of the independent variable such that the system obeys Poisson statistics, if the discreteness and independence of events still hold. Thus the Poisson formula holds in that circumstance.