Practical Astrostatistics

Statistics is crucial to astronomy. Indeed, statistics is crucial to almost all of science, and we could generalize even further by saying that a decent understanding of probability and statistics is somewhere between useful and essential for *anyone* who wants to participate knowledgeably in society. And yet you wouldn't have to look through many astronomy papers to find one that misused statistics. Some of those misuses might be minor and benign, but others can really confuse a field.

My intent in this course is to get you to ask yourself two questions before you perform any statistical analysis. First:
How *would* I perform this analysis if I had unlimited time and resources?
and then second:
How *can* I perform my analysis, with the least loss of accuracy and precision, given my finite time and resources?

In particular, I would like you to make any approximations or compromises *consciously*, and to understand their consequences. Most people don't do that. This is largely understandable; because of our finite time we often offload parts of a project onto a collaborator or onto some pre-written analysis package. For example, I'm a theorist. When I collaborate with observers I don't double-check everything they do (including learning how to do the relevant analyses from scratch); I try to ask as many questions as possible, but I can't do everything myself. But with statistics, many people go way too far in that direction. They don't ask themselves simple questions about the data, and this can result in their analyses being wrong. I will consider this course to be a success if you come out of it with the practice of asking yourselves those simple, intelligent questions when you embark on an analysis.

One other unusual aspect of this subject is that in astronomy there are many strong and differing opinions about the right way to go about statistical analyses. If I teach a class on radiative processes, then I wouldn't expect an insistent debate on the nature of Compton scattering; it is what it is, and although I might not present the material optimally I would expect that the class would generally go along with the topics.

In contrast, because astronomers analyze data all the time, it is natural that they come to prefer certain methods over others. This can lead to a certain digging in when it comes to statistical methods. I have, for example, had what I'll call "intense" discussions with more than one Ph.D. astronomer who tells me with great heat that binning of data is necessary for statistics to work, or to get the most information out of data. I've seen many such disagreements on this or other statistical topics among colleagues.

I want to make clear that I am not setting myself up as the guru of the One True Path to Statistical Nirvana; there is an aspect of art to statistics, in the sense that the decision

about what practical compromises you might need to make in a particular circumstance will often not be unique. But for this course to work we'll need to pull in the same direction, which will be mine because I'm teaching it :). With that in mind, I want to make clear my point of view, which is that if we *did* have unlimited time and resources, the uniquely correct and rigorous analysis to perform in almost any situation would be based on Bayesian principles. As a result, classes 5−9 in this course will be an all-too-brief overview of those principles. But most of this course will focus on how to decide what to do if you can't do the full analysis; what is feasible and retains the most information?

Before we head into some common mistakes people make during their astronomical analyses, there are two points I'd like to make that might seem pedantic, but that I'll insist on because they eliminate an enormous amount of confusion and can reduce errors:

Error versus uncertainty.—Many times these terms are used semi-interchangeably. But error and uncertainty relate to each other like accuracy and precision. Suppose that you throw darts at a dartboard. Your accuracy would be the distance between the bullseye and the 2-D average location of your throws. Your precision would be how tightly grouped your throws are. With this definition I hope it's clear that you could be accurate but not precise (your throws are all over the place but their centroid is close to the bullseye) or precise but not accurate (your throws are beautifully close to each other, but are sadly two meters from the bullseye). To emphasize this, we will talk about "systematic error" and "statistical uncertainty". Why does it matter? If you think your measurements only have statistical uncertainty then more measurements will give you a better answer. But if you have systematic errors then it is not at all guaranteed that more measurements will help you out; in the example above, if you're always throwing your darts (very precisely!) two meters away from the bullseye then the average of more throws is *still* far away from the bullseye.

Model comparison versus parameter estimation.—Here is one case in which my Bayesian inclinations emerge. If we want to be rigorous, we have to understand that there are two basic statistical tasks. In one, we compare two or more models to determine which one represents the data better. In the other, we *assume* that a model is correct, and try to figure out the best values and uncertainties on the parameters in that model. Many times, people confuse the two. It is easy, for example, to be lured into parameter estimation without realizing that in doing that you are assuming that the model is correct; if it isn't, your work could be meaningless!

With that preamble, let's talk about some common mistakes astronomers (and many others) make in their statistical analyses:

Some Statistical Sins

Ignoring systematics.—There's a saying that in astronomy 3σ happens half the time.

That's a little tongue-in-cheek, but the reason this is said (when really 3σ should happen 0.3% of the time) is that it is very rare indeed that we understand our instruments and contaminating effects perfectly. Maybe that bump in the light curve of your source was a flare, but maybe it was just a cosmic ray that hit your detector. Maybe the detector had a nonlinear response to some photons, or perhaps its calibration isn't perfectly understood. There are also cases in which contaminating sources can intervene. For example, in 1989 a remarkable discovery was announced: an active galaxy had a clear periodic signal in its X-ray emission, with a period of ∼12,100 seconds. Revolutionary! But it turned out to be a cataclysmic variable along the line of sight. Not so revolutionary. Think twice before you rewrite physics...

Not estimating "trials" correctly.—In an otherwise featureless spectrum you see an intriguing bump, which you excitedly calculate to have a statistical probability of 10^{-3}. Wow! Have you just discovered unobtanium? Maybe, but did you take into account that you have 1000 spectral bins and thus that there were 1000 chances to have a bump that is improbable at the 10^{-3} level? Many times people will not account correctly for the number of "trials" they perform, and thus they overestimate the significance of the effect. This can be insidious, in the sense that it may not be obvious how many trials are being performed. For example, in the last several years it has become popular to see planar structures in the distribution of satellite galaxies. One well-publicized result notes that 15 out of 29 satellites of Andromeda are in a plane of thickness around 10 kpc... and 13 of the 15 are orbiting in the same direction. Amazing! But you'd be equally amazed if the structure made an "S" shape or something like that. Here the possible flaw is that the sequence is (1) see something that looks interesting, then (2) calculate the probability that exactly that thing should happen. This is Feynman's "license-plate fallacy": isn't it remarkable that yesterday the car parked next to mine had a license plate that read HSX 495? That exact license plate, out of all possibilities!

Null hypothesis testing.—This is a little tricky, and it has some relation to the issue of trials. In an introductory statistics class you are often told that this is *the* way to test a hypothesis. That is, you have a model, and you determine how likely it is that you would see some data if your model is correct. For example, your model might be that a signal is constant, and you use some statistical approach to determine whether the data are consistent with your model. If you judge the model to be inconsistent with the data at some significance level, then you reject the model at that significance level. This may sound reasonable, but the reason it is tricky is that this approach compares, in a nebulous way, a specific model with *all other models combined*. It could easily be that no other specific model does better than your model, which would mean that you were incorrect to reject your model. That's why in Bayesian statistics there is an insistence on doing model comparison between *precisely specified* models. That being said, here is a respect in which I differ somewhat from Bayesian

orthodoxy; I think it's not a bad idea to have *some* way to determine whether the model you're considering is an adequate fit to the data, in an absolute sense. More about this in lecture 9.

Thinking that you need to bin.—As we will discuss in class 3, it is very common in statistical analyses to assume a Gaussian distribution for some quantity. Many tools require that assumption (e.g., this underlies the calculation of χ^2). But people usually understand that when one has a small number of points, the distribution will typically *not* be Gaussian. So they take their data and group it so that there are larger number of data points per group, and thus so that the statistics are closer to Gaussian. I have, incredibly, had Ph.D. scientists tell me that this *improves* the precision of the resulting statistical inference. No, no, no! By grouping data you lose track of where in the group the data originated, so you are guaranteed to lose information. Now, it could be that the information you lose is of negligible importance, or that it is computationally infeasible to use all the data in their original form, but if you are somehow forced to bin you should do so with eyes open.

Confirmation bias and the elimination of "outliers".—It's easy to want certain results from an analysis. But because we do know that glitches occur, sometimes an observation or a point in that observation might not really be representative of the source. As a result, we can be tempted to try to identify those "outliers" and eliminate them, to get "clean" data. But beware! This leads to a statistics version of confirmation bias, by which we reinforce our prejudices when we see something we like, and dismiss evidence that contradicts our prior conclusions.

Subtracting a background rather than modeling it.—Suppose that you're looking for an excess above a background; maybe there is some overall sky glow, and you're looking for evidence of a dim high-redshift galaxy. Or, maybe a source has some constant level of emission and you're interested in whether it has flared in a particular time. A common and incorrect procedure is to *subtract* the constant level from the emission when either assessing the case for the existence of the source or flare, or determining the parameters and their uncertainties for the phenomenon. Indeed, at least until recently this was automatic in the analysis package XSPEC, which is standard in the X-ray community. Why is this wrong? Because fluctuations in the data depend on the *total* intensity or number of counts. Suppose, for example, that we're in the Gaussian regime, where if the average number of counts in some interval is N, we'd expect $N \pm \sqrt{N}$ in a particular observation. Then if (for instance) we use χ^2 statistics, $\sqrt{N}$ is what we use for the standard deviation of the data. If the background has 99% of the counts, then if we subtract the background then we erroneously conclude that the fluctuation level (and the standard deviation we use in our χ^2 analysis) is $\sqrt{0.01N}$, or only 1/10 of the correct value. The right procedure is to include a model of the background as part of the overall modeling of your data.

Using a black box code.—As discussed above, we have finite time and thus we naturally focus our personal resources on a limited set of things. But when we do statistical analyses, this can come back to bite us. Someone points us to a particular statistical package, which is used for our type of analysis. Yay! These can save us a lot of effort; for example, who wants to spend a huge time writing their own code from scratch to interpret data from a particular instrument? But the analysis performed by the package will usually make certain assumptions, and those won't always be valid. The XSPEC example above is a case in point: if you just stuff your data into the code and ask for an answer, it does things (like background subtraction) that are actually wrong, and you'd never know. It is your responsibility to determine the assumptions used in any package you employ, and to understand the consequences of those assumptions.

Not thinking about whether your answers make sense.—Try actually *looking* at your data! Do the conclusions you drew from your analysis pass the gut check test? If not, think again. For example, it can easily be that you do an analysis, estimate parameters, and end up with some clear conclusions, but actually your model doesn't fit the data. Or, you can do something in a formally right way that leads to an answer that is actually absurd. As an example, many years ago I saw a paper in which the authors computed a correlation coefficient between two quantities, call them A and B. They concluded that the two are stunningly well-correlated; the coefficient was 0.9997! But they had a graph of the quantities, and it was pretty much a scatter plot. What's going on??? It turns out that they had a log-log plot, and because there was one very high point and one very low point, a straight line fit beautifully (basically, it's a line between two points). But they didn't comment on it. Remember, you are the master of the statistics; statistics shouldn't boss you around!

A word about the coding exercises

There is nothing like hands-on exercise to help in any topic! For this class, this will involve coding, which I'll try to make reasonable. The idea is that, say, you would perform the "Coding exercise for class 2" on the website by the beginning of class 2. Because the early exercises involve the basic concepts I'll use synthetic data sets to make the relevant points, but I hope to use real astronomical data sets for most of the examples.