Astro 121, Spring 2014
Week 2 (January 30)

Topic: Uncertainty and statistics
Break: ??? – let me know if you can do it

Reading: All of these books are on reserve in Cornell unless otherwise noted. You don't have to read all of these; read what you need to understand the essential concepts and do the problems. I'd recommend Chromey as a good place to start, plus Taylor Chapter 11 to see some examples of the Poisson distribution in action; definitely read those, and then take a look at some of the other sources as needed. So:

Highly recommended/required:

- *To Measure the Sky* by Chromey, Chapter 2.
- *An Introduction to Error Analysis*, by John R. Taylor. (This is the book with the train wreck on the cover.) Taylor covers much the same ground as the other references above, but more slowly. Definitely read Chapter 11, on the Poisson distribution. Chromey mentions it, but doesn't show any examples of how to use it, so it's useful to see Taylor's examples.

Optional if you want more info:

- Other chapters of Taylor. For anything that isn't clear in Chromey, I'd suggest taking a look at the corresponding Chapter in Taylor.
- *Data Reduction and Error Analysis for the Physical Sciences,* Bevington and Robinson, Chapters 1–4.
- "Practical Statistics for Astronomers – I. Definitions, the Normal Distribution, Detection of Signal", J.V. Wall 1979, *Quarterly Journal of the Royal Astronomical Society*, **20**, 138–152. This contains less theory than some of the other sources, but it is more closely aligned with our goals as astronomers, and it is pretty readable. You probably wouldn't want it to be the only thing you ever read about statistics, but it's a good complement to other sources. You can find it through NASA's Astrophysics Data System, or ADS.

Important terms and concepts:

- Parent distribution vs. sample distribution; mean; standard deviation; standard deviation of the mean; Poisson distribution; Gaussian distribution; propagation of uncertainty

We have a number of questions this week that ask things like "is this unusual?" or "do these agree?". I have asked the questions in those plain-English ways because those are the first-order, natural-language questions we tend to ask ourselves. But as scientists, we need to go a bit farther, and ask (and answer) more carefully posed, quantitative questions.

Now I'll give away the answer to all of these questions: it's "maybe." ;-)

In other words, it is almost never the case in dealing with astronomical questions that you can 100% positive about questions like this. And since that's the case, the only truly useful answer to such questions is a probabilistic one, i.e. a *quantitative* statement of the probability that two numbers are the same, or that a given field would have the population it does simply by chance. For each problem like this, you should try to state a clear, specific hypothesis, and then calculate a quantitative probability that this hypothesis is true. Often this is what's called a *null hypothesis*, meaning that the hypothesis is the simple, uninteresting answer. For example, a null hypothesis might be "The source has a constant flux" or "the stars in these fields are randomly distributed". If you find a small probability that such a hypothesis is true, then you can rule it out with a high degree of confidence.

The other thing to keep in mind as you're working through these problems is that whenever you're counting something (stars, galaxies, photons), you probably want to use the Poisson distribution.

Problems and questions

1. Uncertainty on the mean. What happens to the sample standard deviation *s* (Chromey Eq. 2.5) as the sample size increases? What happens to $\sigma_\mu(n)$, the uncertainty on the mean? Can you give an intuitive explanation for this difference?

2. You take a CCD image of a nearby star. On examining the image, you find that the average background level (due to emission from the sky) is 200 photons per pixel. (This problem assumes a perfect CCD with no intrinsic noise; we'll deal with real CCDs soon.)

    a. If the sky background level is determined from averaging the photon counts in 10^6 pixels, what is the uncertainty on the background level?

    b. The pixel centered on the star contains 1000 photons, which is the sum of the counts from both star and sky. (Assume that all the light from the star falls in this one pixel.) What is the signal-to-noise ratio of your measurement of the flux from the *star*? Note that there is more than one source of noise to consider here.

    c. What would the signal-to-noise ratio be if you observed the same star with *no* background emission? (Since it's the same star and the background is gone, now assume that the pixel with the star has 800 counts.)

3. Rules of thumb, round 1. A useful rule of thumb in astronomy is that an uncertainty of 0.01 magnitudes is roughly equal to a fractional uncertainty of 0.01 (i.e. 1%) in flux. Use the formalism of error propagation discussed in any of the readings to show that this is indeed true. (It turns out that it simply is the result of a numerical coincidence—what is that coincidence?) At what magnitude uncertainty is this approximation no longer correct when rounded to the nearest percent?

4. Statistics from real CCD data. Since we'll be using the telescope and CCD throughout the semester, let's take a first look at the statistical behavior of one aspect of our CCD's data. Take at least 4 bias frames with the CCD, and transfer them to the Linux computers in the astro lab. Be sure to allow the CCD temperature to stabilize before taking your biases, since there is a slight temperature dependence to the bias level. Take 2x2 binned images. Use IRAF to look at the statistical behavior of the pixel values. You can do this with the *imstat* task, which will calculate mean, median, standard deviation, etc. of the pixel values in a given image. You may also find it useful to use *display* followed by *imexam* to look at the images.

    a. What does a single bias frame look like? Do you see any patterns or features of note? Print out one image to turn in.

    b. What are the mean and standard deviation of the pixel values in a single bias frame?

c. Before doing any further analysis, predict what you expect the mean and standard deviation to be when you add together two bias images. What about when you subtract two images? Justify your predictions in writing. (Really write them down! If you don't, and a prediction doesn't come true, it's easy to persuade yourself after the fact that you "knew it all along".) You can think of each pixel as a sample from a distribution; each is a measurement of the bias level in the CCD.

d. Now add and subtract the bias frames, using *imarith*. Look at the combined frames (one with the sum and one with the difference) using *imstat*, and note the mean and standard deviation.

e. Were your predictions correct? If not, can you come up with an explanation? And can you think of a way to test your explanation?

f. Finally, form two difference images (e.g. bias1 – bias2, and bias3 – bias4), and then take the difference of *those* resulting difference images. Now what do you find for the standard deviation of the pixel values in the resultant image, compared to the input images? You may be surprised at the result – can you figure out a way to explain it?

5. In a 10-pc-radius volume around the star TW Hya, there are three stars that lie well above the main sequence and have space velocities of less than 10 km/s relative to the Local Standard of Rest. In 28 other fields selected in the same way, there are 22 fields with no stars that meet these criteria, 5 fields with one such star, and 1 field with two. Is the TW Hya field unusual? (This is a problem I really had to solve. If you want, see Jensen, Cohen, & Neuhäuser 1998, *Astronomical Journal* [you can get the exact reference from ADS] to check your answer against mine *after* you've worked the problem out for yourself.) When you work the problem, think carefully about whether or not to include the TW Hya field in your calculation of the mean number of stars per field. Should it be included, or omitted?

6. In Preibisch et al. (2001; *Astronomical Journal* 121:1040), the authors primarily examine the low-mass population of the Upper Scorpius OB association, but in Section 5.2.2, they briefly discuss the high-mass population. Based on previous studies, they say that the expected number of high-mass stars in a field that is 2° in diameter is 1.3 ± 0.1 stars. In the 2°-field that they study in this work, they find two massive stars, which they say is "consistent with the expected number within the uncertainties." Given that one could *never* find a field with 1.2–1.4 stars (the range that naively appears to be allowed by the uncertainties), what does it mean to be consistent with the expected number? You might want to find the probability that the authors would find what they did in a random field, as well as finding the most likely number of stars in a field.

7. Look at the paper "Infrared radiation from an extrasolar planet", by Deming et al., *Nature* 434:740 (2005), which presents the first detection of photons from an extrasolar planet. On the top of page 742, the authors say "Because about half of the 1,696 points are out of eclipse, and half are in eclipse, and the SNR≈111 per point, the error on the eclipse depth should be $0.009 * 2^{0.5}/848^{0.5}$ = 0.044% of the stellar continuum." Derive that expression for the uncertainty on the eclipse depth. Be sure to look at the eclipse data–note that you could never see the eclipse with just a few data points; it's the power of averaging many noisy measurements that makes it possible.

8. Survival analysis. Do Chromey, problem 1 from Chapter 2. The type of statistical analysis necessary to deal with questions like this formally is called "survival analysis". But astronomy doesn't generally deal with survival times—so why should we care about questions like this? (As I'm sure you have guessed, we *should* care.)

9. You are reading the *Astrophysical Journal*, and you find one paper that gives a value of 32 ± 2 km/s for the radial velocity of a star, and another paper that gives a value of 28 ± 2 km/s for the radial velocity of the same star. Do these numbers agree with each other? Is the radial velocity of the star variable? Explain your answer carefully, defining exactly what you mean by "agree."