576 Homework #2

You can work through this problem set however you like, Jupyter notebooks, matlab, root, etc., but please document your thinking clearly in words and plots. The built in statistics packages will be your friend, particularly cdf() and inverse cdf functions (for python use the spicy.stats module). Problems marked (L) are for the light homework option, while those who have opted for the heavy homework option should also do problems marked (H).

Problem 1(L)

Really just getting used to your stats package, we will do a few simple calculations using Gaussian distribution.

a) Create 1e6 random throws from a standard normal distribution (zero mean, variance 1 Gaussian). Generate semilogy histogram.
b) Calculate the fraction of the pdf in the upper tail above integer sigma values 1 through 5. (This should not be the number of events thrown in part a, but the actual numerical integral. Don't actually do the integrals, use your library. There are multiple ways of doing this, think about what is most accurate. In particular subtracting two large numbers to get a small number can lead to numerical error.)
c) Compare the fraction of events from your random sample to what you expected from part b)
d) Reverse the problem. Choose a small probability (something less than 1e-6) and calculate the associated 'sigma' for that probability.

Problem 2 (L)

Choose a non-Gaussian distribution of your choosing, and simulate a set of background events drawn from that distribution (~10^6 is a good number of events). Then choose a candidate event you wish to determine the significance of (null hypothesis).

a) Plot a histogram of the background events and the overlaid analytic distribution on semilogy axes.
b) Clearly describe what 'more signal like' means for your case (probably involves creating a little story about what the measurement is). Convert your story into an integral.
c) Determine the probability of the background producing an event as signal-like or more than your candidate.
d) Determine the 'sigma' of your candidate, and discuss whether this was significant detection or not. (This may involve community standards for your field of study.)

Problem 3 (H)

Reverse Problem 2. Choose a desired 'sigma' and determine how strong the candidate must be to be a significant detection. In effect you are predicting the sensitivity of your measurement.

Problem 4 (L)

Assume you are able to average multiple observations from Problem 2 together (the signal is of constant strength over multiple identical observing runs).

a) Calculate the pdf of the average of N observations (pick N; makes sure distribution is getting narrower).
b) Explore how the significance of your candidate changes as you average more observations together (increase N).
c) How many observations must you average together before you can use Gaussian statistics? Discuss some of the subtleties and approximations needed to answer this question.

Problem 5 (H)

This problem is inspired by folks in class doing exoplanet spectra. We are going to imagine we have spectrally white (constant power with frequency) light that has passed through a thin gas containing some spectral lines. Some of the lines will be in emission (add power) and others in absorption (remove power), and our goal is determine if a certain line exists. The spectrum of the combined light is dispersed with a grating and detected as a measure of power vs. frequency (or energy; this is just a standard spectrum).

a) Assume that our power detector after the dispersion grating works in the photon counting regime, and the background distribution is dominated by the counting statistics of the white backlight. (Note this is not currently true of most visible and IR spectrometers which have dark current and other contributions to the observed background. However this is true of many x-ray and gamma ray spectrometers and near future visible and IR spectrometers built around superconducting transition edge sensors or stepper CCDs may become photon counting limited.) What is the expected statistical distribution of the measured power when there is no spectral line?
b) Build a specific model for your background, including an expected photon rate when there is no emission or absorption.
c) You have a spectral location where you expect to see a line in emission. Come up with a candidate signal that adds X photons to the background; describe the integral you need to do; and calculate the 'sigma' of your candidate. (It is okay to fiddle with your candidate if you happened to pick a wildly weak or strong signal the first time, we're just playing with statistics here.)
d) Using the same number X as you did in part c; repeat the problem for a line seen in absorption.
e) Unless you chose a hugely strong signal or background, you will have noticed that the sigma of the emission and absorption signals in c & d did not agree, despite the strength (±X) of the signal being the same. Discuss why this is.
f) Now imagine you have an emission line that may be in either emission or absorption. What is the integral you need to perform when calculating sigma? Explain your thought process carefully.

Optional

If you want a jump on next week's homework, start looking at your data and determining how to obtain a signal-free background distribution and how to model it.