Gravitational-Wave Data Analysis Exercises – Day 4

Peter S. Shawhan
CGWA Summer School, May 31, 2012

There are many books and articles about statistics, of course. A nice concise overview of both frequentist and Bayesian statistics, by Fred James, can be found at tinyurl.com/GWAdata, or in its original place: http://www.ippp.dur.ac.uk/Workshops/02/statistics/proceedings/james1.pdf .
I recommend that you read it now, especially if your background in statistics is not so strong. You may also wish to refer to it when answering the questions below.

1. Bayesian probability example

(Adapted from an example in the article by Fred James)

Suppose you are a doctor, and you have at your disposal a flu screening test that has an 8% false-negative rate and a 2% false-positive rate. In other words, in a patient who has the flu, the test will come back negative 8% of the time, and in a patient who does *not* have the flu, the test will come back positive 2% of the time.

a. In Maryland in the wintertime, about 3% of the population seems to have the flu at any given time. If you test a random person and the test comes back positive, what do you conclude about the probability that that person has the flu? If the test comes back negative instead, what is the probability that that person has the flu?

b. The people who are sick enough to come to your office are more probable to have the flu than the population in general; let's say that 25% of those people have the flu. Now, if you test one of these patients and the test comes back positive, what do you conclude about the probability that that person has the flu? What if the test comes back negative instead?

c. On South Padre Island in the summertime, only 0.1% of the population has the flu at any given time. If you test a random person and the test comes back positive, what do you conclude about the probability that that person has the flu? If the test comes back negative instead, what is the probability that that person has the flu?

2. Frequentist upper and lower limits for Poisson process

Suppose there is a Poisson random process which will, on average, produce μ events in an experiment, where μ is unknown.

a. What is the 90% frequentist upper limit on μ if N=0 events are detected in the experiment? In other words, what value of μ yields a 90% probability of yielding *more than* zero events?

b. What is the 90% upper limit on μ if N=1 events are detected?

c. What is the 90% upper limit on μ if N=2 events are detected?

d. What is the 90% ***lower*** limit on μ if N=1 events are detected?

e. What is the 90% ***lower*** limit on μ if N=2 events are detected?

3. Frequentist estimation of the mean

a. Generate an array of N=100 "measured" values, $x_j = s + n_j$, where s is a constant signal equal to 0.3 and the n_j are random values associated with a white, Gaussian-distributed random noise process with zero mean and unit variance.

b. To do frequentist statistics, you define a procedure for setting a *confidence interval* on some parameter of interest based on the data you have. In this case, you have a set of 100 measured values and want a statistical estimate of the true signal value (0.3, but suppose you don't know that). Assuming you know in advance that the noise has unit variance, an appropriate interval-setting procedure is to take the mean of the 100 values plus-or-minus 1.65 sigmas, where sigma is the standard error on the mean, because that central slice of a Gaussian distribution contains 90% of the area. Calculate that interval for your set of 100 values; does it contain the true signal value? It probably will, but not necessarily, depending on your set of random numbers.

c. The real test of a frequentist statistical method is whether it satisfies the "coverage" condition, meaning that in an ensemble of experiments, the interval contains the true value the advertised fraction of the time. To check this, repeat the above two parts for at least 1000 random realizations of the "data", keeping track of the number of realizations for which the interval contains the true value. Is it around 90% of them?

d. Repeat part c for N=1000. The interval you set gets narrower; does it still contain the true value the desired fraction of the time?

4. Bayesian estimation of the mean

(Adapted from an exercise written by Joe Romano)

a. Let $x_j = s + n_j$ be an array of N=100 random values similar to what you generated for the previous problem.

b. Assuming you know in advance that the noise has unit variance, calculate and plot the likelihood function $\mathcal{L}(s) := p(x_1, x_2, \ldots x_N|s)$ for a particular realization of the data x_j. Hint: for a *single* random variable with mean s and standard deviation σ, the likelihood of getting some particular value x is $\frac{1}{\sqrt{2\pi}\sigma}\exp(-(x-s)^2/2\sigma^2)$.

c. Suppose it would be unphysical for the true value to be negative. So, with a prior of the form

$$p(s) = \begin{cases} 1 & s \geq 0 \\ 0 & s < 0 \end{cases}$$

calculate and plot the posterior distribution $p(s|x_1, x_2, \ldots, x_N)$ for s. How does it differ from the likelihood function calculated in part b?

d. From the posterior distribution, calculate the mode (i.e., the most-likely value) of s and the Bayesian 90% probability interval centered on the mode. This is a bit tricky to implement.

e. Repeat the above four parts for several more realizations of the data, noting how the mode and Bayesian confidence interval change from realization to realization.