ASTR/PHYS 356 - Problem Set 2

Posted: February 9, 2018 Due: February 21, 2018

Problem 1 – Testing the Central Limit Theorem (10 pts)
Test the central limit theorem by creating plots similar to the figures shown in the Wikipedia page for "central limit theorem". That is: **(a)** show the distribution of mean values of N random variables obtained from a uniform distribution for increasing numbers of N (i.e., N = 2, 5, 20, 200) and compare each distribution with a Gaussian with the same mean and variance as the obtained distribution. **(b)** Do the same as (a) but using the distribution in problem 4 of problem set 1, with $a = 2$.

Problem 2 – Distribution of Exoplanets (15 pts)
Go to exoplanets.org and obtain the distribution of stellar [Fe/H]. For this, go to the "Plots" tab and click on the "Histogram Plot" and "Advance" tabs (on the right). Using the "Data" pull-down menu (which is below the "Simple" and "Advance" tabs) choose FE (almost at the end of the menu, under "Stellar Properties"). Choose the min and max of the [Fe/H] values to be -0.5 and 0.5, respectively, and a binwidth of 0.04 (this should give you 25 bins), using all stars with exoplanets and measured values of [Fe/H].

Use this distribution as the parent pdf. Write a code to obtain 10^5 randomly distributed points from this parent distribution. Plot a histogram of the value of the points obtained from your code and compare it with a plot of the parent distribution. Are the two distributions similar? How would you test if these two distributions come from the same parent distribution?

FYI: [Fe/H] is the ratio of the Fe abundance to the H abundance, relative to that of the Sun in a logarithmic scale. Hence, a value of [Fe/H] = 0 indicates that the [Fe/H] value of the star is the same as that of the Sun, while a value of [Fe/H] = 0.1 indicates the star has a value of [Fe/H] that is one-tenth that of the Sun. In many instances this quantity is used to estimate the "metallicity" of a star. In astronomy metals refer to any element heavier than Helium.

Problem 3 - Fitting a line to data using the MLE (10 pts)
Suppose you have bivariate data $(x_1,y_1),\ldots,(x_n,y_n)$. A common model is that there is a linear relationship between x and y, so in principle the data should lie exactly along a line. However since data have random noise and our model is probably not exact this will not be the case. What we can do is look for the line that best fits the data. To do this we will use a simple linear regression model.
For bivariate data the simple linear regression model assumes that the x_i are not random but that for some values of the parameters a and b the value y_i is drawn from the random variable
$Y_i \sim ax_i + b + \varepsilon_i$
where ε_i is a normal random variable with mean 0 and variance σ^2. We assume all of the random variables ε_i are independent.

Notes:

- The model assumes that σ is a known constant, the same for each ε_i.
- We think of ε_i as the measurement error, so the model says that:

  $y_i = ax_i + b + \textit{random measurement error}$
- Remember that (x_i, y_i) are not variables. They are values of data.

(a) The distribution of Y_i depends on a, b, σ and x_i. Of these only a and b are not known. Give the formula for the likelihood function $f(y_i \mid a, b, x_i, \sigma)$ corresponding to one random value y_i.

(b)

i- Suppose we have data (1, 8), (3, 2), (5, 1). Based on our model write down the likelihood and log likelihood as functions of a, b, and σ.

ii - For general data (x_1,y_1), ..., (x_n,y_n) give the likelihood and log likelihood functions (again as functions of a, b, and σ).

(c) Assume σ is a constant, known value. For the data in part b(i) find the maximum likelihood estimates for a and b. Give confidence intervals for your estimates of a and b.

(d) Use python to plot the data and the regression line you found in 1c.

Problem 4 - Estimating parameters of a uniform distribution (5 pts)

(a) Suppose we have data 1.2, 2.1, 1.3, 10.5, 5 which we know is drawn independently from a uniform(a, b) distribution. Give the maximum likelihood estimate for the parameters a and b. Recall that the pdf for uniform(a,b) is $f(x \mid a, b) = 1/(b-a)$ for $a \le x \le b$, and 0 otherwise. [Hint: in this case you should not try to find the MLE by differentiating the likelihood function.]

(b) Suppose we have data $x_1, x_2, \ldots, x_n$ which we know is drawn independently from a uniform(a, b) distribution. Give the maximum likelihood estimate for the parameters a and b in mathematical form.