ASTR 600 Homework #7: Truncated Data (Chapter 10)

1. The provided data set *luminosity.txt.* contains simulated data for a sample of galaxies. The file has three columns: the first is x-ray luminosity, the second is radio luminosity, and the third is distance.

Let us say we detect a galaxy in the x-ray if

$$L_x/d^2 \geq 5,$$

and similarly in the radio if

$$L_r/d^2 \geq 5$$

Otherwise, the value(s) for that galaxy are censored with a limit of $5d^2$.

a. Before doing any calculations, do you expect L_x and L_r to be correlated? That is, if a galaxy has high x-ray luminosity, would you expect it to also have high radio luminosity?

Let's test your expectation.

b. First, throw out any censored data points, and then compute a Kendall's τ correlation coefficient as we learned in chapter 5of the text book.

Does your result confirm your initial expectation?

c. Now, instead of ignoring the censored data, write an R code to compute a modified version of Kendall's τ statistic given by the following (from the Brown *et al*. (1973) reference in the text book):

$$\tau = \sum_i \sum_j a_{ij} b_{ij}$$

where now

$$a_{ij} = \begin{cases} 1 & if\ X_j > X_i \text{ AND } X_j \text{ is not censored} \\ 0 & if\ X_j = X_i\ OR \text{ if it's uncertain (i. e. both censored)} \\ -1 & \text{if } X_j < X_i \text{ AND } X_i \text{ is not censored} \end{cases}$$

And similarly for b_{ij}.

Hint: the statistic is asymptotically normal with variance given by:

$$var = \frac{4}{n(n-1)(n-2)}\left(\sum_i^n\sum_j^n\sum_k^n a_{ij}a_{jk} - \sum_i^n\sum_j^n a_{ij}^2\right)\left(\sum_i^n\sum_j^n\sum_k^n b_{ij}b_{jk} - \sum_i^n\sum_j^n b_{ij}^2\right) + \frac{2}{n(n-1)}\left(\sum_i^n\sum_j^n a_{ij}^2\sum_i^n\sum_j^n b_{ij}^2\right)$$

Does this result agree with your answer from part b? Did including the censored data provide you with any additional information?

2. Read in the data file *abundance.txt.* There should be 68 rows in the file, with three values in each row. The 68 data points correspond to 68 stars, roughly half of which host planetary companions. The first entry in a row indicates whether the star has a planet (1 = yes, 0 = no), the second entry is the star's measured log abundance of beryllium, and the third entry indicates whether the abundance given is a censored value, i.e., an upper limit (0 = censored, 1 = not censored).

It's known that stars with higher metallicity are more likely to be found hosting a planet. Test this fact by comparing the Be distribution of stars with planets to stars without.

Is the Be distribution different between the two groups?

3. Read in the data file from http://data.princeton.edu/wws509/datasets/gehan.dat. The file contains data on length of remission for leukemia patients, and has three columns: the first indicates the treatment (1 is medicine, 2 is placebo), the second indicates weeks in remission, and the third indicates whether a relapse occurred (1 for a relapse, and 0 for censored data).

a. First, compare the length of remission between then groups (placebo and medicated). Is the medication effective? (make sure to define what you mean by effective)

b. Determine the probability that a patient who is receiving the medication will experience a relapse after being in remission for 9 weeks.