Statistics for Astronomers
Homework #6 (Due before 10:00 PM on Monday, 2020.12.14)

Prof. Sundar Srinivasan

December 7, 2020

Notes: (1) You are welcome to use `Python` functions to evaluate probabilities for various distributions, and `Mathematica/Wolfram Alpha` to compute integrals if necessary. **Just mention your source in each case!** (2) Email me your `Python` scripts and any/all resulting output plots/images.

For each question, assume independent datasets, and that all hypothesis tests require 5% significance.

1. The questions below require the files available here. These tables contain data from Scott et al. (1978) on the levels (in mg dL^{-1}) of plasma cholesterol (first column) and triglycerides (second column, which you won't use) for a control sample (first file) of $N_1 = 51$ subjects with no evidence of heart disease and for a test sample (second file) of $N_2 = 320$ subjects who had narrowing of the arteries (arteriosclerosis/atherosclerosis).
   You will use these data to determine whether there is a correlation between plasma cholesterol level and heart disease.

   (a) (**3 points**) Compare the distributions of cholesterol levels among the two samples using a box-and-whisker plot that shows the median and mean of each sample. Based on the box plots, can you comment on whether or not each of the distributions is symmetric?

   (b) (**4 points**) Perform a 1-sample Kolmogorov-Smirnov test on each sample to determine whether they are Gaussians. Can you use the 2-sample F- or t-tests on these samples?

   (c) (**3 points**) Can you reject the null hypothesis "there is no difference in cholesterol levels in the two samples (and, therefore, no connection between cholesterol level and heart disease)" using the Mann-Whitney U test? In order for there to be a connection between cholesterol level and heart disease, we must demonstrate that the cholesterol level is **higher** in patients in the second sample (that is, perform a one-sided test). Use this as your alternate hypothesis.

2. **Comparing globular cluster luminosity functions.**
   We want to know if the properties of globular clusters are the same between the Milky Way and the Andromeda Galaxy. Download K-band data for globular clusters in the Milky Way and M31. The Milky Way data is corrected for distance (*i.e.*, they are absolute K magnitudes), but the M31 data aren't. Correct the latter assuming a distance modulus of 24.9 mag. Use these two datasets to answer the following questions.

(a) (**5 points**) Perform the 2-sample Kolmogorov-Smirnov and Anderson-Darling tests to determine whether the two samples are drawn from the same distribution. Compare the empirical distributions of the two samples to explain why the two tests disagree.

(b) (**1 points**) Perform a two-sided Mann-Whitney U test to determine whether the two samples are drawn from the same distribution.

(c) (**5 points**) Use the sample mean and sample standard deviation to studentise each dataset. Perform the 1-sample Anderson-Darling test to see if the studentised data in each case are consistent with a standard normal distribution. Which dataset is less likely to be drawn from a standard normal?

3. (**6 points**) Cite one example of a refereed astronomy paper since 2015 that uses **either** the likelihood-ratio test **or** the F-test in its analysis. Briefly explain **in your own words** for what purpose the test is used. Identify the null and alternate hypotheses, and comment on their conclusion – for example, what is their p-value? Are they able to reject their null hypothesis?