Statistics for Astronomers
Homework #3 (Due before 5:00 PM on Wednesday, 2019.03.20)

Prof. Sundar Srinivasan

March 13, 2019

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) for a control sample (first file) of $N = 51$ subjects with no evidence of heart disease and for a test sample (second file) of $N = 320$ 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)** Compute the empirical distribution function $\hat{F}(x)$ for both the control sample and the test sample. Find the point x_k in each sample such that $\hat{F}(x_k)$ is closest to 0.5. What is the significance of these points?

(b) **(4 points)** Use the bootstrap algorithm to generate distributions for the sample means of the control and test samples. Compute the mean and standard deviation for each distribution.

(c) **(3 points)** Using the results from part (b), generate 95% confidence intervals (CIs) for the population means of the control and test samples. Based on these CIs, are the two samples drawn from distributions with the same population mean?

(d) **(3 points)** If $\hat{\theta} \equiv \hat{\mu_t} - \hat{\mu_c}$, where $\hat{\mu_t}$ and $\hat{\mu_c}$ are estimators of the population means of the test and control samples respectively, what is the expectation value and variance of $\hat{\theta}$? Based on these values, are the two samples drawn from distributions with the same population mean?
Hint: use the results from part (b).

(e) **(5 points)** Winsorize [1] the original control and test samples and use the bootstrap algorithm to generate distributions for the sample **medians (not means!)** of the control and test samples. Compute the mean and standard deviation for each distribution. Construct 95% CIs for the population medians. Based on these CIs, are the two samples drawn from distributions with the same population median?

(f) **(2 points)** Do your conclusions in parts (c) and (e) agree? Why or why not? Based on the results of part (c), can you conclude that there is a connection between cholesterol level and heart disease?

[1] Named after its inventor, this is a censoring technique. Any data outside of the central 95% of the distribution is replaced with the value of the 95$^{\text{th}}$ percentile. For instance, if the value at the 95$^{\text{th}}$ percentile is 5.22, then all numbers greater than this will be replaced with 5.22. In `Python`, an array of data `a` can be Winsorized as follows:
`a_Wins = scipy.stats.mstats.winsorize(a, limits = 0.025)`,
where `limits = 0.025` corresponds to 2.5% on either side of the distribution, or data that is outside the central 95% of the distribution.