<u>ATOC5860 - Homework #1 - due January 30, 2024</u>

Please send your homework to Jen/Prof. Kay on Slack as a direct message.

Please Name Your Homework Files: "ATOC5860_HW1_LastName.pdf, .html, .ipynb"

Your submissions should include: 1) A .pdf document with responses to the questions below, 2) Your code in both .ipynb and .html format.

Show all work including the equations used (e.g., by referring to the Barnes Notes).
Write in complete, clear, and concise sentences.
Eliminate spelling/grammar mistakes.
Label all graph axes. Include units.

1) Basic statistics (60 points).

a) **Bayes Theorem. Assume background rates of COVID are 90% negative, 10% positive AND COVID tests are accurate 80% of the time, but fail 20% of the time. Your friend goes and gets a COVID test. Your friend test negative. What is the probability that your friend is actually negative? Explain to your friend how you are using Bayes theorem to inform your thinking. *Hint: Review Lecture #1 and the 1.2.2.2 of the Barnes Notes.* (10 points)**

b) **Explain how to test whether a sample mean is significantly different than zero at the 95% confidence level and the 99% confidence level. State each of the 5 steps in hypothesis testing that you are using. For step 4, calculate the specific critical value assuming a two-tailed test. Contrast your approach for a sample with 15 independent observations (N=15) and a sample 1000 independent observations (N=1000). (15 points)**

c) **Design your own homework problem to compare two sample means using data of your own choice. In other words, test whether two sample means are statistically different. Follow all five steps of hypothesis testing. *Hint: See page 26 of Barnes notes for an example.* (15 points)**

d) **Design your own homework problem to place 95% confidence intervals on the mean value of a data variable of your choice. Use the non-standardized variable. *Hint: See Barnes notes on Confidence Intervals.* (10 points)**

e) **The F-statistic is used to compare two sample standard deviations. Design your own homework problem to compare two sample standard deviations and assess if they are different at the 95% confidence interval. *Hint: See page 38 of the Barnes notes.* Note: When calculating the f-statistic Barnes Chapter 1 Equation (122), the larger variance should always be on top (numerator) and the smaller variance should always be on bottom (denominator). i.e., F = Larger variance / Smaller variance. (10 points)**

2) Compare composite-averages using t/z tests and bootstrapping. Note: coding is required for this problem. Please use python Jupyter notebooks. It will be helpful follow the ipython notebook examples introduced in Application Lab #1 and in lectures. (40 points)

Your friend living in Fort Collins tells you that the air pressure is anomalous when there is measurable precipitation (greater than or equal to 0.01 inches). To test your friends' hypothesis, use hourly observations from Fort Collins in 2014. The data include both the precipitation amount in units of inches and pressure in units of hPa at hourly frequency. The data file is called homework1_data.csv.

a) **What was the average pressure in 2014 ($\overline{P}$)? What was the average pressure when it rained ($\overline{P}_{R>=0.01}$)? (10 points)**

b) **Test your friends' hypothesis by generating confidence intervals using both a t-statistic and a z-statistic. Is the average pressure different when it is raining? What is more appropriate to use as a statistical test – a t- or a z-statistic? Use 95% confidence interval. Write down and address all five steps of hypothesis testing. (15 points)**

c) **Instead of the t/z-test – use bootstrap sampling to determine whether the local pressure is anomalously high during times when it is raining. Write down and address all five steps of hypothesis testing. How does your answer compare with your results using the t/z-test? (15 points)**

Instructions for Bootstrapping: Say there are N hourly periods when R >= 0.01 inches. Instead of averaging the pressure P in those N hours, randomly grab N pressure values and take their average. Then do this again, and again, and again 1000 times. In the end you will end up with a distribution of mean N pressures (P_N) in the case of random sampling, i.e., the distribution you would expect if there was no physical relationship between P and N. Plot a histogram of this distribution and provide basic statistics describing this distribution((mean, standard deviation, minimum, and maximum). Then quantify the likelihood of getting your value of $\overline{P}_{R>0.1}$ by chance alone using percentiles of the boot-strap generated distribution of P_N.

Aside: The name bootstrapping comes from the saying "pulling yourself up by your boot straps", the idea of getting something for nothing. For this method you do not need to know the true distribution underlying your data. You just re-use the data you have to try to calculate the statistics you need.