**ATOC7500 - Homework #1 - due Friday September 11, 2020**

**Please e-mail your homework to Jen/Prof. Kay at** [**Jennifer.E.Kay@colorado.edu**](mailto:Jennifer.E.Kay@colorado.edu)**.**

**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 equations used.**

**Write in complete, clear, and concise sentences.**

**Eliminate spelling/grammar mistakes.**

**Label all graph axes. Include units.**

**1) Basic statistics (60 points).**

1. **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)**

Using equation 23 from Barnes, the probability of testing negative and actually being negative is 97%, given both the test accuracy and background COVID rates. A Bayesian approach to this problem is important because if we ignored the background COVID rates, we would just conclude that the probability of testing negative and being negative is 80%. But this assumes that COVID has infected 50% of the population. Given that most of the population does not have COVID, this significantly reduces the likelihood of receiving a false negative.

1. **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. Contrast your approach for a sample with 15 independent observations (N=15) and a sample 1000 independent observations (N=1000). (15 points)**

The 5 steps for hypothesis testing are:

1. State the significance level
2. State the null hypothesis
3. State the statistic to be used (t or z) and necessary assumptions
4. State the critical region
5. Evaluate the statistic and state the conclusion

These 5 steps would remain the same regardless of the sample size. The main difference however is that with a sample size of 1000 independent data points, I would use the z-test, whereas for the smaller sample size I would use the t-test. With a sample size of 15 independent data points, it is important to use the t-test because the statistic takes into account that smaller samples are not perfectly normal and thus is a more conservative statistic. The differences in the significance levels will not influence the hypothesis procedure either but instead will just alter the critical regions in the t and z statistics.

1. **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)**

Question: Is there a significant difference between atmospheric pressures when precipitation is between 0.01 and 0.1 inches per hour and when precipitation is more than 0.1 inches per hour?

The 5 steps for hypothesis testing are:

1. State the significance level: **95%**
2. State the null hypothesis: **There is no significant difference between air pressures when precipitation is between 0.01 and 0.1 inches per hour and when precipitation is more than 0.1 inches per hour**
3. State the statistic to be used (t or z): **Because both samples have more than 30 data points, we can use the z-test. We are assuming they are all independent data points that are normally distributed**
4. State the critical region: **We will reject the null hypothesis if z >= 1.64.**
5. Evaluate the statistic and state the conclusion: **Using Barnes’s equation 108, z = 0.68. Because z < 1.64, we have failed to reject the null hypothesis that there is no significant difference between air pressures when precipitation is between 0.01 and 0.1 inches per hour and when precipitation is more than 0.1 inches per hour.**
6. **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)**

Question: What are the 95% confidence intervals on the true mean of air pressure when precipitation is greater than 0.1 inches per hour?

Because the size our sample is greater than 30, we can calculate confidence intervals using the z-statistic (Barnes equation 90). The confidence interval of the true mean air pressure when precipitation is greater than 0.1 inches per hour is:

Lower limit: 845.02 hPa

Upper limit: 847.92 hPa

1. **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.* (10 points)**

Question: Is there a statistically significant (95%) difference between the standard deviation of air pressure when precipitation is less than 0.1 inches per hour versus more than 0.1 inches per hour?

Both samples come from normal population and have approximately the same variance, therefor we can calculate the F-statistic using equation 122 from the Barnes notes. The critical F-statistic at the 95% confidence interval is 1.51 and the calculated F-statistic is 0.83. F < Fcrit. Thus, we cannot reject the null hypothesis that there is a significant difference between the standard deviations of the two populations.

**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. The data file is called homework1\_data.csv.**

1. **What was the average pressure in 2014 ()? What was the average pressure when it rained ()? (10 points)**

The 2014 average atmospheric pressure in Fort Collins was 846.33 hPa.

The 2014 average atmospheric pressure during rainfall in Fort Collins was 847.03 hPa.

1. **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. (15 points)**

The confidence intervals using the t-statistic (Barnes equation 100) are:

Low: 846.58

High: 847.48

The confidence intervals using the z-statistic (Barnes equation 90) are:

Low: 846.58

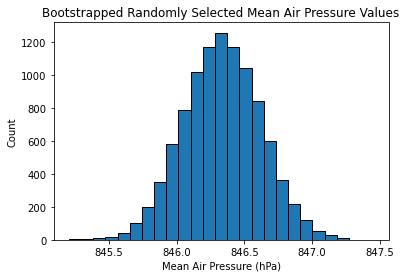
High: 847.48

The mean pressure of the entire time series is not contained in the 95% confidence intervals of mean air pressure when it is raining. Therefor we can conclude that the mean air pressure when it is raining is statistically different than when it isn’t raining.

In this scenario, the z-statistic is more appropriate because we have a very large sample size (N >> 30). Here, it does not matter because for very large sample sizes, the t and z distributions converge, but technically, for very large N, the z-statistic is more appropriate.

1. **Instead of the t/z-test – use bootstrap sampling to determine whether the local pressure is anomalously high during times when it is raining. 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 (PN) 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  by chance alone using percentiles of the boot-strap generated distribution of PN.***

****

Some basic statistics of the bootstrapped distribution are:

Mean: 846.33 hPa

Standard deviation: 0.28 hPa

Minimum: 845.19 hPa

Maximum: 847.28 hPa

Using equation 87 from the Barnes notes, we find that the probability that an average atmospheric pressure of 847.03 hPa would only occur in a randomly sampled pressure time series 0.72% of the time. This is in keeping with our previous findings (using confidence intervals) that there is a statistically significant difference between Fort Collins atmospheric pressures when it is and is not raining.