ATOC 5860 - Homework 1

Joshua Lave

1)

a) The frequentist attitude toward our friend's negative test might be that there is an 80% chance that our friend is actually COVID-negative based purely on test accuracy. However, we can apply Bayes' Theorem to account for a priori information and thereby improve our estimation; in this case, the a priori information is the background rate of COVID. Our application of Bayes's Theorem is as follows:

$$Pr(N|TN) = \frac{Pr(TN|N)Pr(N)}{Pr(TN|N)Pr(N) + Pr(TN|P)\mathrm{Pr}(P)}$$

where

Pr(N): Background COVID negative rate = 0.9
Pr(P): Background COVID positive rate = 0.1
Pr(TN|N): COVID test accuracy rate = 0.8
Pr(TN|P): COVID test inaccuracy rate = 0.2

Therefore, the probability that my friend is actually negative is about 0.97 or 97%. The present application of Bayes' theorem uses conditional probability to account for the background rate of COVID among the population to inform the probability that he is actually negative. This approach to statistics is often referred to as the Bayesian approach.

b) Hypothesis testing can be used to test statistical significance at any confidence level by using one or more test statistics. Testing at a that a sample mean is different than zero at a 95% and 99% confidence level involves a five step process as outlined and followed below.

i) Step one: State the significance level (α). To test at 95% confidence, set α = 0.05; to test at a 99% confidence, set α = 0.01.
ii) Step two: State the null hypothesis (H_0) and alternative hypothesis (H_1). In our present case, H_0: the sample mean is equal zero (μ=0), and H_1: the sample mean is not equal to zero (μ≠0).
iii) Step three: State the statistic to be used, and the assumptions required to use it. For sample sizes less than 30 we can use the t-statistic, while for larger samples we can use the z-statistic. Each of these statistics represent the number of standard errors a sample mean lies from a population mean and are representative of standard normal distributions (where the mean = 0 and the standard deviation = 1). To use these statistics we must assume a normal sample distribution.
iv) Step four: State the critical region. This involves using statistics tables. The degrees of freedom (df) is N-1. In the case where N=15 we use the t-statistic, df=14, and so the critical region for α=0.05 is $t > t_{0.025}$=2.145 and for α=0.01 is $t > t_{0.005}$=2.977. In the case where

N=1000 we use the z-statistic, df=999, and so the critical region for α=0.05 is $z > z_{0.025}$=1.6 and for α=0.01 is $z > z_{0.005}$=2.3.

v) Step five: Evaluate the statistic and state the conclusion. We evaluate the data to calculate the test statistic. If the test statistic falls in the critical region then we can reject the null hypothesis. If it does not, then we cannot reject the null hypothesis.

c) Derived brightness temperatures from GOES16 ABI band 10, centered at 7.3 microns, are often interpreted in terms of low-level tropospheric water vapor density. Using the GOES16 ABI dataset available on AWS, determine if morning, low-level water vapor was significantly different across the CONUS domain in September 2021 as opposed to September 2020.

The GOES16 dataset can be retrieved from the s3://noaa-goes16 bucket via the AWS command line interface (CLI). Download the first scan starting after 12UTC on each day in September in 2020 and 2021. These samples represent the population of scans from all times during September 2020 and 2021. The python code to convert the radiance grids into a list of daily (12Z), domain-average brightness temperature is given in the notebook. Now on to the five steps of hypothesis testing:

i) Significance level: α = 0.05
ii) H_0: the 2020 sample mean is equal to the 2021 sample mean
iii) We will use the z-statistic since each of our sample include 30 observations. We will assume that the population distribution of daily, domain-average brightness temperatures across all times follows a normal distribution. We will use a two-sided test.
iv) We will reject the null hypothesis if $z > z_c = z_{0.025} = 1.6$.
v) We use the following equation for the z-score:

$$z = \frac{(\overline{x_1} + \overline{x_2}) - \Delta_{1,2}}{\sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}}$$

We evaluate the z-score at z = 0.6597, so we are unable to reject the null hypothesis that the sample means of the two years are equal.

d) Using the dataset from part c), calculate the 95% confidence interval on the mean of daily, domain-average brightness temperature from 2021.

The number of samples, mean, and standard deviation for our September 2021 brightness temperature data are 30, -19.561°C and 1.019°C, respectively. Assuming a normal distribution about the mean, we can calculate the 95% confidence interval as

$$\mu = \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{N}}$$

For $\alpha = 0.05$, $z_{\alpha/2} = \pm1.96$, so our 95% confidence interval ranges from -20.101°C to -19.369°C.

e) Again, using the dataset from part c), compare the standard deviations from the 2020 and 2021 brightness temperature distributions and assess if they are significant at the 95% confidence level.

To solve the problem we follow the five steps of hypothesis testing:

i) Significance level: $\alpha = 0.05$
ii) H_0: the 2020 sample standard deviation is equal to the 2021 sample standard deviation
iii) We will use the f-statistic to compare sample standard deviations. We assume that the populations have the same variances and are both normally distributed.
iv) Using an f-statistic table where df = 29 for both samples, our critical region is $f > f_c = 1.86$.
v) Given the assumptions in part iii), we can use the simplified version of the f-score:

$$f = \frac{s_1^2}{s_2^2}$$

Our two standard deviations were calculated in part c) as 1.024 and 1.019, so f = 1.01. We are not able to reject the null hypothesis that the two standard deviations are equal.

2)

a) We can load our data into a pandas dataframe and use integrated functions to retrieve the desired means. The average observed surface pressure during all observations was 846.332 hPa, while the average observed surface pressuring during rain observations was 847.031 hPa.

b) First, we can calculate sets of descriptive statistics for all and rain only observations:

For all observations -
Mean pressure: 846.332
Pressure standard deviation: 5.617
Number of observations: 8760

For rain observations -
Mean pressure: 847.031
Pressure standard deviation: 5.386
Number of observations: 384

Next, we can set up our five step hypothesis test:
i) Significance level: $\alpha = 0.05$
ii) H_0: average pressure during rain observations in Fort Collins is equal to average pressure during all observations
iii) We will use the z-statistic. The problem asks us to conduct this test with both the z- and t-statistic but, with N >> 30, the t-statistic has converged to the z-statistic. We assume a

normal distribution of the population dataset, even though the distribution looks to be somewhat skew-negative. This skew is consistent with atmospheric dynamics, which predicts that low pressure systems are capable of digging much deeper than high pressure systems are capable of building.

iv) The critical z-score at 95% confidence for a two-tailed test is $z > z_{0.025} = 1.96$.

v) We will use the following formula to calculate the z-score

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{N}}$$

Using the descriptive statistics calculated above we obtain z = 2.439, meaning that we are able to reject the null hypothesis at the 95% confidence level.

Finally, we can set up confidence intervals using both a z-statistic and t-statistic. At a 95% confidence level, the equations for the statistics are as follows:

$$\mu_z = \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{N}}$$
$$\mu_t = \bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{N-1}}$$

For these equations:

$\bar{x}$ = 846.332 hPa
N = 384
σ = s = 5.617
$z_{0.025}$ = 1.96
$t_{0.025}$ = 1.96

Given the large number of observations, our confidence intervals for both the z- and t-statistic will be nearly identical. In any case:

z-statistic 95% confidence interval: 846.469 to 847.593
t-statistic 95% confidence interval: 845.769 to 846.895

Therefore the average pressure during rain, 847.031 hPa, is significantly different at the 95% confidence level. We know that the t-statistic converges to the z-statistic at values over 30, so while either statistic can be used on this large sample, the z-statistic is more appropriate.

c) First we will generate a set of 1000 mean pressure values from random samples of 384 observations, corresponding to the number of observations available during rain. Once we have this distribution of means, we can calculate a set of descriptive statistics for the bootstrapped distribution:

Bootstrap mean: 846.335
Bootstrap standard deviation: 0.288

With the mean and standard deviation available, we can now follow a similar five step hypothesis test as in part c):

i) Significance level: $\alpha = 0.05$
ii) H_0: average pressure during rain observations in Fort Collins is equal to average pressure during the bootstrapped samples
iii) We will once again use the z-statistic, but this time our assumption that the background population distribution is normal is accurate. Normalizing a skewed dataset is a common use of bootstrapping.
iv) The critical z-score at 95% confidence for a two-tailed test is $z > z_{0.025} = 1.96$.
v) We will (again) use the following formula to calculate the z-score

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{N}}$$

Using the descriptive statistics calculated above we obtain z = 2.417, meaning that we are able to reject the null hypothesis at the 95% confidence level.

The z-scores using both the z-test and bootstrapping are similar. However, it is noteworthy that the exact value of the z-score will change based on the exact bootstrapped distribution. Of course, with a high enough sample value we expect to find a steady state normal distribution but that level of samples is too great to quickly run on my computer.