DAVID ROSENCRANS

ATOC5860 - Homework #1 - due January 27, 2022

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)**

Here we use the Bayes theorem instead of the frequentist approach. A frequentist might think that the odds of our friend being negative is 80%, since COVID tests are correct 80% of the time. However, Bayes theorem accounts for a priori knowledge, such as the background rates of COVID in addition to the accuracy of the test.

1) Using Barnes Eq.s 23-26
   Pr(N) is the probability of being negative for COVID.
   Pr(P) is the probability of being positive for COVID.
   Pr(T) is the probability of testing negative.
2) Using Barnes Eq. 30:
   we want to know Pr(N|T).
3) P(N) = 0.9
   P(P) = 0.1
   P(T|N) = 0.8
   P(T|P) = 0.2
4) Using Barnes Eq. 34-35:
   $E_1 = N$
   $E_2 = P$

5) Using Barnes Equation 23:

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

$$= \frac{(0.8)(0.9)}{(0.8)(0.9) + (0.2)(0.1)}$$
$$= 0.972 = 97\%$$

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

First, we explain the steps if a sample contains 1000 independent observations. If our sample size is greater than 30 independent observations and is normally distributed, we use the z-statistic. If using a 95% confidence level, $\alpha = 0.05$. Thus, $\frac{\alpha}{2} = 0.025$ and $1 - \frac{\alpha}{2} = 0.975$. Using the z-statistic table, our critical z is 1.96. Conversely, if using a 99% confidence level, $\alpha - 0.01$. Thus, $\frac{\alpha}{2} = 0.005$ and $1 - \frac{\alpha}{2} = 0.995$. Using the z-statistic table, our critical z is between 2.57 or 2.58, and is thus 2.575. The critical z represents the number of standard errors in which the sample mean must differ from the population mean in order to be statistically significant. After calculating the sample mean, population mean, and population standard deviation, we can solve for z. If z is greater than the critical z, then our sample mean is statistically significant.

For N = 1000:
1) 95% confidence so $\alpha = 0.05$
2) H0: the sample mean is not statistically different from the population mean.
   H1: the sample mean is statistically different from the population mean.
3) We use the z-statistic since we have 1000 observations and a gaussian distribution.
4) The critical z-score is $z_c = 1.96$
5) Using Barnes Eq. 83:
   Evaluate the z-score using $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{N}}$ and compare with z_c.

For N = 1000
1) 99% confidence so $\alpha = 0.01$
2) H0: the sample mean is not statistically different from the population mean.
   H1: the sample mean is statistically different from the population mean.
3) We use the z-statistic since we have 1000 observations and a gaussian distribution.
4) The critical z-score is $z_c = 2.575$
5) Using Barnes Eq. 83:
   Evaluate the z-score using $z = \frac{\bar{x} - \mu}{\sigma/\sqrt{N}}$ and compare with z_c.

Next, we explain the steps if a sample contains only 15 independent observations. Since our sample size is less than 30 independent observations, but is normally distributed, we may use the student's t-statistic. As before, we must find the critical t-value using the t-table. With 14 degrees of freedom at a 95% confidence limit, the critical t-value is 2.1448.

Conversely, with 14 degrees of freedom at a 99% confidence limit, the critical t-value is 2.9768. As before, the critical t-value represents the number of standard errors in which the sample mean must differ from the population mean in order to be statistically significant. After calculating the sample standard deviation, sample mean, and population mean, we solve for t. If t is greater than the critical t-value, our sample mean is statistically significant.

For N = 15

1) 95% confidence so $\alpha = 0.05$
2) H0: the sample mean is not statistically different from the population mean.
   H1: the sample mean is statistically different from the population mean.

3) We use the t-statistic since there are only 15 independent observations and a guassian distribution.
4) The critical t-score is $t_c = 2.1448$ using 14 degrees of freedom.
5) Using Barnes Eq. 96:
   Evaluate $t = \frac{\overline{X_1} - \mu}{\frac{s}{\sqrt{N-1}}}$ and compare with t_c.

For N = 15

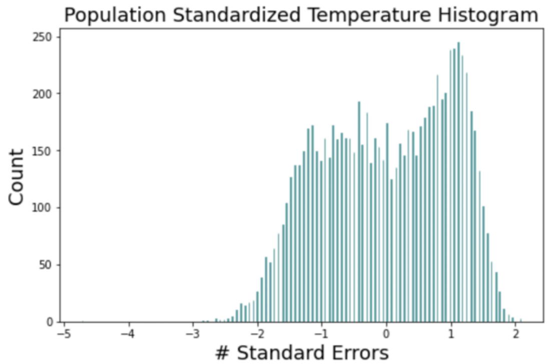
1) 99% confidence so $\alpha = 0.01$
2) H0: the sample mean is not statistically different from the population mean
   H1: the sample mean is statistically different from the population mean
3) We use the t-statistic since there are only 15 independent observations and a guassian distribution.
4) The critical t-score is $t_c = 2.9768$ using 14 degrees of freedom.
5) Using Barnes Eq. 96:
   Evaluate $t = \frac{\overline{X_1} - \mu}{\frac{s}{\sqrt{N-1}}}$ and compare with t_c.

Jen – would you rather explanative paragraphs or the numbered version?

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)**

Your team pulls daily maximum temperature data from an ASOS station in Ajo, Arizona from 19981 to 1984. Using the 1,827 days of good data, you conclude that the average high temperature was 28.2°C with a standard deviation of 8.3°C. Your team then samples data from 2002 to 2006 and finds that the average temperature is now 29.2°C with a standard deviation of 8.6°C. Due to instrument aging, only 1,174 days of good data are collected. Have temperatures warmed significantly over the past 14 years?

The standardized population is Normally distributed with a mean of $4.76\text{x}10^{16} \cong 0$, a standard deviation of 1, and a skew of -0.2.



1) 95% confidence level ($\alpha = 0.05$)
2) H0: $\mu_{1984} = \mu_{2006}$
   H1: $\mu_{1984}! = \mu_{2006}$
3) We will use the z-statistic since the temperature population is Normally distributed with a mean of 0, a standard deviation of 1, a skew of -0.2, and we have greater than 30 data points in each sample. Although we might infer that temperatures could be warmer due to climate change, 4 years of data is not a climatology, so we use a two-sided approach.
4) We will reject the null hypothesis if the z-score is greater than the critical z score: $|z| \geq |z_c| = 1.96$
5) From Barnes Eq. 103:

$$z = \frac{\overline{x_1} - \overline{x_2} - \Delta_{1,2}}{\left(\frac{\sigma_1^2}{N1} + \frac{\sigma_2^2}{N_2}\right)^{\frac{1}{2}}}$$

$$z = \frac{28.2°C - 29.2°C - 0}{\left(\frac{(8.3°C)^2}{1827} + \frac{(8.6°C)^2}{1174}\right)^{\frac{1}{2}}}$$

$$z = -3.03$$

Since $|z|$ is larger than $|z_c|$, we can reject the null hypothesis that the sample means for the two time periods are equal.

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)

Say we have daily temperature data from the period 2001 to 2005. Using the 1,266 data points for this record, the average temperature is 29.5°C with a standard deviation of 8.7°C. What is the 95% confidence interval on the true population mean?
From Barnes Eq. 83:

$$z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{N}}}$$

$$\mu = \overline{x} \pm \frac{z_{crit}\sigma}{\sqrt{N}}$$

$$\mu = 29.5°C \pm \frac{(1.96)(8.7°C)}{\sqrt{1266}}$$

$$29.02°C \leq \mu \leq 29.98°C$$

We know that the true population mean is 29.33°C which is within the confidence limit.

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. (10 points)

Planners want to simulate wind speeds in Boulder to determine if enough power could be produced to install a wind farm. They run a year-long simulation and find that, during their period of interest in July when grid demand is largest, the standard deviation of daily wind speeds is 1.42 m s^{-1}. During the year-long simulation, the standard deviation of wind speeds is 1.5 m s^{-1}.

To quantify uncertainty, the planners run a separate simulation using a different planetary boundary layer scheme. During July, daily wind speeds now have a standard deviation of 1.3 m s^{-1} and the year-long simulation a standard deviation of 1.35 m s^{-1}. Are the standard deviations between runs statistically different?

1) We use the 95% confidence interval so $\alpha = 0.05$.
2) H0: The sample standard deviations are not different.
   H1: The sample standard deviations are different
3) We use the f-statistic since we are comparing the standard deviations of two samples with sizes of 31 days.
4) We will reject the null hypothesis if the f-score is greater than the critical f-score of is 1.82.
5) Evaluate using Barnes Eq. 121:

$$f = \frac{s_1^2/\sigma_2^2}{s_2^2/\sigma_2^2}$$

$$= 0.966$$

Since $f = 0.97 < f_{crit} = 1.82$, we cannot reject the null hypothesis that the sample standard deviations are different. Thus. since the winds can be predicted somewhat, the planners install the wind farm ☺

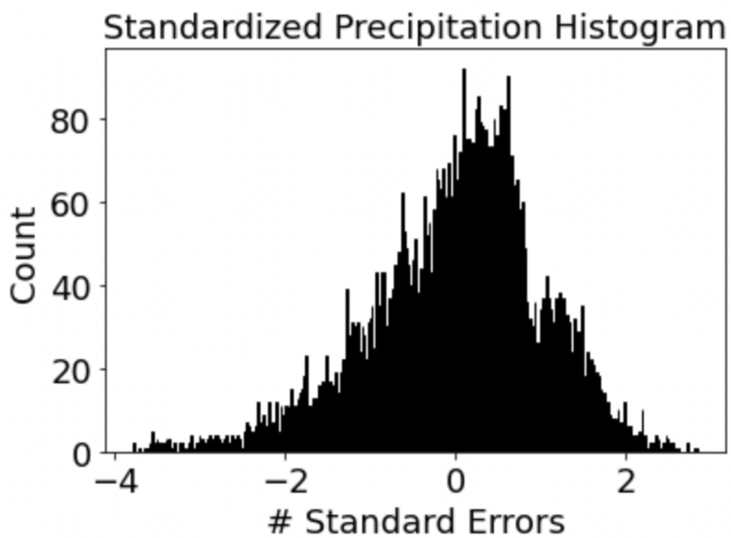
2) Compare composite-averages using t/z tests and bootstrapping. Note: coding is required for this problem. Please use python Jupyter notebooks. It will be helpful to follow the ipython notebook examples introduced in Application Lab #1 and in lectures. (40 points)

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.

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)**
The average pressure in 2014 was 846.33 hPa. The average pressure when it rained was 847.03 hPa.

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. (15 points)**
The underlying population is Normally distributed with a standardized mean of $6.57\text{x}10^{-15} \cong 0$ and a standardized standard deviation of 1.



z-score

1) We will use a 95% confidence interval so $\alpha = 0.05$
2) H0: The average pressure during precipitation events is the same as the overall average pressure
   H1: The average pressure during precipitation events is not the same the overall average pressure
3) We will use the z-score since there are 384 data points in the sample and it is Normally distributed with a skew of -0.5.
4) We will reject the null hypothesis if the z-score is greater than the critical z score of 1.96.
5) Using Barnes Eq. 83:

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

$$z = \frac{(846.82\ \text{hPa}) - (846.33\ \text{hPa})}{\frac{5.37\ \text{hPa}}{\sqrt{254}}}$$

$z = 2.54$

Using the z-score, we can reject the null hypothesis that the mean pressures are different.

t-score

1) We will use a 95% confidence interval so $\alpha = 0.05$
2) H0: The average pressure during precipitation events is not different from the overall average pressure.

   H1: The average pressure during precipitation events is different from the overall average pressure.
3) We will use the t-score since the data is Normally distributed with a skew of -0.5 and 254 data points. Although the z-score is sufficient with 254 data points, the t-score converges to the z-score.
4) We will reject the null hypothesis if the t-score is greater than the critical t-score of 1.96 with 253 degrees of freedom.
5) Using Barnes Eq. 96:

   $$t = \frac{\overline{x_1} - \mu}{\frac{s}{\sqrt{N-1}}}$$

   $$t = \frac{(846.82\ \text{hPa}) - (846.33\ \text{hPa})}{\frac{5.33\ \text{hPa}}{\sqrt{253}}}$$

   $t = 2.54$

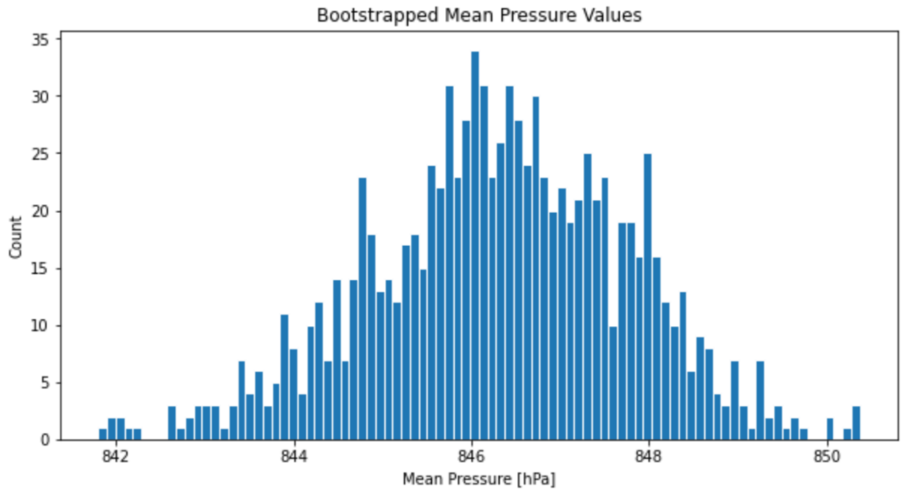
   Using the t-score, we can reject the null hypothesis that the mean pressures are the same.

Either test is appropriate in this case since the t-statistic converges to the z-statistic at large data lengths. This is why we see that the z- and t-scores, and their critical values, are equal in this case (rounding to two decimals).

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. 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.



We incorporate bootstrap sampling with 1000 sets of 384 individual data points to create a distribution of means. We choose to average 384 data points per sample since there are 384 days that contain precipitation greater than 0.01 inches. Using the bootstrapped average data, the mean is 847.03 hPa, the standard deviation is 0.28 hPa, the minimum is 845.13 hPa, and the maximum is 847.11 hPa.

From this methodology, using a percentile of 2.5% and 97.5% (to contain 95% of the distribution), the interval is from 845.7 hPa to 849846.8.0 hPa. Since the average pressure of 847.03 hPa is not within this range, we can reject the null hypothesis once again.