Homework 5: Sampling, Hypothesis Testing, and Confidence Intervals

Due: 3.05.2021 (NOTE: You have *two* weeks to complete this homework. However, it is recommended to begin this homework before the exam to practice concepts.)

This homework will give you practice in hypothesis testing and confidence intervals with Python. It is the first homework that has a significant writeup component.

Goals

In this homework you will:

1. Formulate hypotheses and carry out appropriate statistical tests
2. Compute confidence intervals based on appropriate assumptions
3. Work with a real dataset (student behavioral data)

Background

Before attempting the homework, please review the notes on sampling and hypothesis testing on the course website. Feel free to copy and modify any of the code we have provided for you here.

Some motivation and helpful information on hypothesis tests with an example:

A hypothesis test (particularly a mean hypothesis test in this example) takes the sample you collected and compares how its mean (i.e., the sample mean `xbar`) looks with respect to the sampling distribution (distribution of the sample means `p(xbar)`). So with your data you collect as your sample, you calculate a single 'mean' value (`xbar`) to test against the 'population of means' of hypothetically possible samples having the same number of datapoints. This is often done because if one has enough data sampled, the mean is distributed as a Gaussian distribution (per the Central Limit Theorem), or even if not much data could be collected, you have the option to use Student's t test. The beauty of this is that virtually no matter what data you are testing (rainfall, test scores, salaries, errors, etc.), the methods you're being taught are applicable, regardless of the type of data you collect.

The 'assumed population mean' (null hypothesis) is the value you test against. For example, your friend Jake says "I reckon there's 3.43957792347 cm of rainfall here daily in spring." If you want to prove him wrong (by running a hypothesis test), you would:

1. Collect samples of rainfall each day in spring at that location.

2. Compute their sample mean.
3. Temporarily assume Jake is correct. (Null Hypothesis H0: mu = 3.43957792347 cm, Alternative Hypothesis H1: mu != 3.43957792347 cm (!= means we only care whether Jake is wrong, if we wanted to prove there was more rainfall, H1 would instead be mu >= 3.43957792347 cm)
4. Hold Jake to a standard (alpha = 0.1, 0.05, 0.01, whatever seems appropriate to prove "your sample was rare beyond a doubt.")
5. Calculate the test statistic (the z-score or t-score, based on the number of datapoints you have collected or information on the true standard deviation value) from the equation in the lecture slides. Rule of thumb: z if greater than 30 datapoints or knowing the true population standard deviation, t for less than 30 datapoints.
6. Determine the p-value from your test statistic. Note: For each p-value, there corresponds a 'critical test statistic' value (or pair of values if a 2 tail test, but you really just care about the side your test point is on for which of the two you'd look at in that case.) so you could just find that value and see if your value passes it in the 'more rare' direction.
7. Compare the p-value to the alpha value you set. If the p-value is smaller than alpha, you have proved Jake wrong (You reject the null hypothesis). If the p-value is larger than alpha, you have failed to prove Jake wrong, but he is not necessarily right because you have only "failed to reject" the null hypothesis.

Confidence Intervals

Confidence intervals are somewhat a flipped perspective compared to hypothesis testing. Hypothesis testing yields claims like "I have rejected that the average wind speed is 20km/hr under significance value alpha=0.05." Whereas a confidence interval for the same claim might say something like "I am 95% confident that the true average wind speed falls in the range (15, 19) km/hr." Here, we notice that the value 20 did not fall in the confidence interval, and so if hypothesis tested at alpha = 1 - 0.95 = 0.05 like we had, this confidence interval already shows the result of the test would be to reject H0: mu=20km/hr.

Why do we have both hypothesis testing and confidence intervals then? There's a number of reasons, but arguably one of the main distinguishing factors is that hypothesis testing is used for making or disproving claims, and confidence intervals don't have to involve proving or disproving a claim but can instead provide a valuable way of bounding an unknown value to some level of certainty.

Python Functions

Reading .txt files

There are several different file formats for data, including .csv and .json which we will cover later. One of the simplest is storing in text (.txt) files, which is how the data is provided to you in this homework. To get each line of a text file `sample.txt` stored as a separate element in a list `data`, you can write:

```
myFile = open('sample.txt')
data = myFile.readlines()
myFile.close()
```

Each element of `data` will be a string. To convert them to floats, we can use a list comprehension:

```
data = [float(x) for x in data]
```

Mean and Standard Deviation

While they are relatively easy to write manually, the `numpy` library in Python has built-in functions for finding the mean and standard deviation of a list. To import it, we can write:

```
import numpy as np
```

The mean of `data` is found as

```
avg = np.mean(data)
```

To find the standard deviation, type

```
sd = np.std(data, ddof=x)
```

where `x` is the differential from the number of samples `N` to determine the degrees of freedom. Typically, we want `ddof=1` (which divides by `N-1` instead of `N`) unless we know the population mean (in which case `ddof=0`).

Standard Normal and Student's t Distributions

The two distributions you will rely on heavily in this homework are the `standard normal (z)` and the `student's t` distributions.

To import the standard normal distribution, type

```
from scipy.stats import norm
```

Then, to find the probability that a value lies below a particular point `z_c`, type

```
p = norm.cdf(z_c)
```

Inversely, to find the point `z_c` below which the probability is `p` (i.e., the inverse cdf), type

```
z_c = norm.ppf(p)
```

To import the Student's t distribution, type

```
from scipy.stats import t
```

Then, to find the probability that a value lies below a particular point `t_c`, type

```
p = t.cdf(t_c, df)
```

where `df` is the degrees of freedom for the t distribution.

Inversely, to find the point `t_c` below which the probability is `p` (i.e., the inverse cdf), type

```
t_c = t.ppf(p)
```

Real-World Example of Hypothesis Testing:

There is a branch of research concerned with analyzing data collected about students to design machine learning-based recommendation systems or personalization engines aimed at improving learning outcomes. Several academic conferences, such as the Educational Data Mining conference, and companies in industry, such as Zoomi Inc., have been created to pioneer this mission statement.

A popular thrust of this research is investigating relationships between how a student interacts with the learning content in a course (their *behavioral* data) and the knowledge that they gain from the course (their *performance* data); intuitively, higher levels of interaction should translate to more knowledge transfer. For online learning courses (on platforms such as Coursera), one recently proposed measure of interaction is *engagement*, which translates factors like time spent, length of content, number of clicks, length of annotations, and other application usage information into a single measure between 0 (no engagement) and 1 (maximum engagement). If you are interested in learning more about educational data mining, start with this journal paper.

Very Important Notes for Your Analysis:

In this homework, you will work with the engagement data of about 3,000 students who took an online course. We divide them into two groups: those who demonstrated sufficient knowledge of the material after the course (about 1,000), and those who did not (about 2,000). This determination is made based on whether they passed their final exam. Viewing "knowledgeable" and "unknowledgeable" students as two different populations, your task in Problem 1 will be to formulate and test different hypotheses about their engagement levels.

In short, you will be working with two sampled groups, those who passed the final exam (stored in engagement_1.txt) and those who did not (stored in engagement_0.txt). What is stored in these files is the engagement values of the students from those groups. You will be doing hypothesis tests to determine such things as 'is the typical passing student's engagement level 0.75?' and 'was there a difference in how engaged passing and nonpassing students were?'

Instructions

0) Set up your repository

Click the link on Piazza to set up your repository for HW 5, then clone it.

The repository should contain two files aside from this readme, both of which you will use in Problem 1:

1. `engagement_0.txt` , a text file containing the engagement scores of students who did not demonstrate knowledge of the course material.
2. `engagement_1.txt` , a text file containing the engagement scores of students who demonstrated knowledge of the course material.

1) Problem 1: Hypothesis Testing

This problem concerns the datasets of student engagement in `engagement_0.txt` and `engagement_1.txt` :

1. Suppose the instructor of the course is convinced that the mean engagement of students who become knowledgeable in the material (i.e., the `engagement_1` population) is 0.75. Formulate null and alternative hypotheses for a statistical test that seeks to challenge this belief. What are the null and alternative hypotheses, and what type of test can be used?

2. Carry out this statistical test using the `engagement_1` sample. Report the sample size, the sample mean, the standard error, the standard score, and the p-value. Are the results significant at a level of 0.1? How about 0.05? How about 0.01? What (if anything) can we conclude?

3. What is the largest standard error for which the test will be significant at a level of 0.05? What is the corresponding minimum sample size? (You may assume that the population variance and approximation does not change.)

4. Suppose the instructor is also convinced that the mean engagement is different between students who become knowledgeable (the `engagement_1` population) and those who do not (the `engagement_0` population). Formulate null and alternative hypotheses that seek to validate this belief. What are the null and alternative hypotheses, and what type of test can be used?

5. Carry out this statistical test using the `engagement_0` and `engagement_1` samples. Report the sample sizes, the sample means, the standard error, the z-score, and the p-value. Are the results significant at levels 0.1, 0.05, or 0.01? What (if anything) can we conclude?

2) Problem 2: Confidence Intervals

In this problem, consider the following dataset of the number of points by which a sports team won in its last 11 games:

```
[3, -3, 3, 15, 15, -16, 14, 21, 30, -24, 32]
```

In other words, a `3` means the team won by `3` points, and a `-3` means the team lost by `3` points.

1. Use the sample to construct a 90% confidence interval for the number of points by which the team wins on average. Report whether you will use a z-test or t-test and report the sample mean, the standard error, the standard statistic (t or z value), and the interval. (Think, which distribution should you use here if very few datapoints are available?)

2. Repeat part 1 for a 95% confidence interval for the number of points by which the team wins on average. Report whether you will use a z-test or t-test and report the sample mean, the standard error, the standard statistic (t or z value), and the interval.Note: Report your results and compare your results to part 1 (i.e., what is different or similar?).

3. Repeat part 2 if you are told that the population standard deviation is `16.836`. (Think, which distribution should you use here now that you have the true population standard deviation?). Report whether you will use a z-test or t-test and the values for the standard error, standard statistic, and confidence interval. Report your results and compare your results to part 1 and part 2 (i.e., what is different or similar?).

4. Assume you no longer know the population standard deviation. With what level of confidence can we say that the team is expected to win on average? (Hint: What level of confidence would you get a confidence interval with the lower endpoint being 0?)

What to Submit

For each problem, you must show your work, including the Python code and written explanations of your answers. You can either submit Python files with a separate writeup as a pdf, or a single jupyter notebook containing both the code and the writeup. Your writeup must clearly indicate which problems and questions you are answering (for example if you are answering question 3 of problem 2, we need to be able to know that). The suggested format is "Prob 1, Qs. 1", "Prob 1, Qs. 2", ..., "Prob 2, Qs 4". If you do a separate writeup, make sure to indicate in each question, what code file you used to answer that question. In short, a grader needs to be able to easily match answers to questions, if they cannot, you will lose points.

Submitting your code

Please commit and push the latest version of your code and writeup as you have done in previous assignments. Please verify your submitted files by looking at GitHub online. The submission deadline is 11:59pm ET on the March 5th, 2021.