Spring 2019

Homework: Probability and Statistics - Week 8

Evan Agovino
CUNY City College

NYC Tech-in-Residence Corps

The income per capita per US State in 2018 is approximately a normal distribution with a mean of $48,500 and a standard deviation of $8,500.

1) Create a PDF graph of this distribution with the proper labels on the axes.

2) Say we wanted to create a rejection region for any states with below average income. If we set the rejection region at 40,000, what would the percentile value of that be given the distribution?

3) Say that I'm a Senator from New York, which has an income per capita of $64,500.

I believe that my state has an income that's conceivably different from the rest of the country - so different that I could convince someone that we're not even part of the United States. I want to use hypothesis testing to prove it.

Note that I am not surveying people - this is a single data point. In this case:

```
   a) What is the null hypothesis?

   b) What is the alternative hypothesis?

   c) Say I want to do a two-tailed test at a significance level of 0.05 to see if
   my hypothesis is true. What is my rejection region? Can I reject the null hypoth
esis here?

   d) Seeing that the average income per capita for New York is obviously higher th
an the
   income per capita for the United States, I want to change my test to a one-taile
d test.

   At the 0.05 significance level, what is my rejection region? Can I reject the nu
ll hypothesis here?

   e) Say I want to keep the test one-tailed, but change my significance level to
 0.01. What is my rejection region now? Can I reject the null hypothesis here?
```

4) Say now that I don't actually know the distribution of the income per capita per US State - I'm just making it up to sound smart. But I do survey ten state representatives from New York and get the following data points for income per capita of our state: [62837, 64511, 58821, 63971, 62837, 57084, 64579, 62837, 58821, 63971]

```
    a) What is the sample mean for this distribution?
    b) What is the sample variance and sample standard deviation for this distribut
ion?
    c) Say I again want to see if I can reject the null hypothesis that the income
 per capita for New York
    is the same as the income per capita for the United States.

       Note that this is based off of a survey of ten, so we would want to effectiv
ely see if the ten people we surveyed could have effectively been pulled from anywh
ere in the US, not just New York.
    d) Given a T-distribution with a mean of 50,000, our sample standard deviation,
and nine degrees of freedom, what is the one-tailed rejection region at the 0.01 si
gnificance level? Can we reject the null hypothesis that New York has a different i
ncome per capita than the rest of the United States? Note that we are using our sam
ple mean to determine this, not the previously stated value of $64,500.
```

5) This information was actually pulled from the Wikipedia here - https://en.wikipedia.org/wiki/List_of_U.S._states_by_GDP_per_capita (https://en.wikipedia.org/wiki/List_of_U.S._states_by_GDP_per_capita).
Using the pd.read_html function, read in the table of U.S. State by GDP capita and plot a histogram and boxplot of the 2018 data in the first table. Are there any outliers, and if so, which states are outliers?

BONUS 6) Remove the outlier (if there is one), and replot a histogram and boxplot of the data. Are there any