2019

Test: Probability and Statistics - Practice Final

Evan Agovino
CUNY City College

NYC Tech-in-Residence Corps

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats
%matplotlib inline
```

In [2]:
```
df = pd.read_csv('all_things.csv')
```

The above is a CSV file that tracks the per capita GDP for each country in the world (where available), along with the percentage of people in each country who pray on a daily basis, based on survey data. There are four columns:

Country/Territory: The name of the country
GDP: The GDP per capita of that country
Pray_Daily: The percentage of respondents from that country who pray daily
Continent: The continent that country is located in

Instructions:

Answer all questions. For any graphs, labelling the axes, titling the graph and changing the size of the graph are **NOT** required unless specifically noted.

In [3]:
```
df.head()
```

Out[3]:

	Country/Territory	GDP	Pray_Daily	Continent
0	Ireland	78785.0	19.0	Europe
1	Norway	74356.0	18.0	Europe
2	Switzerland	64649.0	8.0	Europe
3	United States	62606.0	55.0	North America
4	Netherlands	56383.0	20.0	Europe

1) Plot a histogram of the percentage of adults who pray daily by country in the world. What type of distribution is the result closest to?

2) What is the mean and standard deviation of the percentage of people who pray daily per country?

3) If you were to build a **normal distribution** with the above mean and standard distribution, what would the 95th percentile value be of that distribution?

4) Which five countries have the highest percentage of respondents who pray daily? Which five countries have the lowest?

5) Find the average time that responders from each continent pray. Which continent prays the most on average? And which prays the least?

6) Plot a boxplot showing how much different respondents in Europe pray. Are there any outliers? If so, which countries are outliers and how much do respondents from those countries pray?

7) Now let's do a hypothesis test between North American and Asian countries to see if there's a statistically significant difference between the two. First, what is the mean difference between the percentage of respondents who pray daily between the average North American country and the average Asian country?

8) Given that North American countries, on average, a higher percentage of respondents who pray daily than Asian countries, we should do a one-sided bootstrap test with a significance level of 0.05. What is the null hypothesis? And what is the alternate hypothesis?

9) Now let's actually run the null hypothesis. Set a random seed of 42 and run 10,000 simulations. At the stated signifiance level, what is the cutoff value for the rejection region? Given the observed value, can we reject the null hypothesis? At what percentile of the distribution is our observed value? What is the minimum significance level (integer value) with which we'd be able to reject the null hypothesis?

10) Now, graph a scatterplot between the GDP for a country (on the X-axis) and the percentage of respondents who pray daily from those countries (on the Y-axis). What does the relationship looks like?

11) What is the correlation between these two variables? What does that correlation imply about the relationship between the two variables? Is the correlation statistically significant?

12) What is the R-squared value between these two variables?

13) If a country has a per-capita GDP of $50,000, what is the predicted percentage of its respondents who will pray daily according to the model?

14) Which country has the closest per-capita GDP to $50,000 and what is the actual amount of time its respondents pray daily?

15) Describe the residuals plot for this relationship. Is it heteroskedastic or homoskedastic?

16) Look at the histograms for each of the variables. Which of the two would you suggest transforming to get a more linear relationship? How would you suggest transforming it?

17) Take the log value of the appropriate variable and find the new linear relationship. What is the R-squared value now?

18) Describe the residuals plot for this relationship. Is it heteroskedastic or homoskedastic?

19) Which of the two plots (the untransformed variable and the response variable, or the transformed variable and the response variable) is more appropriate for linear prediction? Why might this be counterintuitive?

BONUS: What is a structural error with this dataset? (Hint: are all of the values unique?)

In [ ]: