2019

Homework: Probability and Statistics - Week 10

Evan Agovino
CUNY City College

NYC Tech-in-Residence Corps

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

In [2]:
```
education = pd.read_csv('Downloads/Education.csv', encoding='ISO-8859-1', head
er=0)
education = education.iloc[:, :-1]
education = education.dropna().reset_index(drop=True)
```

1) Plot a histogram of the percentage of adults with a bachelor's degree or higher in 2000.

2) Plot a boxplot of the percentage of adults with a bachelor's degree or higher in 2000. Are there any outliers? If so, how many? What is the cutoff for an outlier on either side?

3) Which state has the highest average percentage of adults with a bachelor's degree? Which state has the lowest? (We want to take the average of the percentage of bachelor's degrees in each county per state) / (Save this to a new variable, as we'll be using it a lot)

4) Plot a histogram for the percentage of adults with a bachelor degree by state.

5) Plot a boxplot for the percentage of adults with a bachelor degree by state. Are there any outliers? If so, how many? What is the cutoff for an outlier on either side?

6) Now, let's read in a dataset that tells us whether a particular state voted for Al Gore or George Bush in 2000.

In [ ]:
```
state_results = pd.read_html('https://transition.fec.gov/pubrec/2000presgeresu
lts.htm')[2]
blue_states = state_results[state_results['ELECTORAL VOTE  BUSH'].isnull()]['S
TATE']
red_states = state_results[state_results['ELECTORAL VOTE  GORE'].isnull()]['ST
ATE']
```

7) Create two arrays of the percentage of adults with bachelor degrees for red states and blue states - HINT: you can use the function index.isin(blue_states) and index.isin(red_states) to specifically query blue and red states.

Plot a boxplot showing blue states and red states (hint: you can plot two separate items in a boxplot by plotting an array, i.e. plt.boxplot([a, b])

Do blue states or red states have a higher mean % of bachelor degrees? What are the means of each? What is the mean difference between the two?

Do either groups of states have outliers? Which states are outliers?

8) Now run a boostrapping example using 10,000 simulations. Use np.random.seed(42) to ensure consistency if you run again. Concatenate the blue states and red states, shuffle them, and then break out new blue states and red states, similar to what we did last week in class. Record the difference between the blue states and red states.

Plot a histogram of the 10,000 differences recorded. What is the average difference?