Spring 2019

Homework: Probability and Statistics - Week 11

Evan Agovino
CUNY City College

NYC Tech-in-Residence Corps

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

In [2]:
```
df = pd.read_csv('https://github.com/fivethirtyeight/data/raw/master/candy-power-ranking/candy-data.csv')
```

1) How many candies are in this dataset?

2) How many candies have chocolate in them? How many don't?

3) Let's call the last column - the 'winpercent' equivalent to an approval rating for a given candy.

What is the mean 'approval rating' for candies with chocolate? What is the mean 'approval rating' for candies without chocolate? What is the difference in these mean approval ratings?

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

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

5) Say you are running a two-tailed hypothesis test, where the null hypothesis is that there is no difference in the approval rating for chocolate candies and non-chocolate candies, and the alternate hypothesis is that there is a difference in the approval rating.

If the test is at the 0.05 significance level, what are the rejection regions?

Where do the results you found earlier fit into the distribution? What is the percentile value? What is the p-value (remember, two-tailed)?

Can we reject the null hypothesis?

6) Now let's use what we've learned the past two weeks. Plot a scatter plot between the percentage of sugar for *all* candies and the approval rating for *all* candies.

7) Do the histograms for either of the variables look skewed? If so, what transformation should we apply?

8) What is the covariance between these two variables? What is the correlation? What does the correlation indicate about the strength of the relationship? Is the correlation statistically significant (Note you do NOT have to calculate this, it is in the correlation output via pearsonr).

9) Using the 'evaluate linear relationship' package we found in class this week, find the slope, intercept, prediction and residuals for this relationship. Re-plot the scatterplot with the predictive slope on top of it like we did in clas.

In [3]:
```
def evaluate_linear_relationship(a, b):
    slope = np.cov(a, b, bias=True)[0][1] / np.var(a)
    intercept = np.mean(b) - (slope * np.mean(a))
    predictions = (slope * a) + intercept
    residuals = b - predictions
    return slope, intercept, predictions, residuals
```

9) What is the R-squared value of the relationship between these two variables? What does this say when considered with the correlation of the two variables?

BONUS

10) Plot the confidence and prediction intervals of the linear relationship between these two variables.

In [4]:
```
def get_intervals(a, b, residuals):
    t = stats.t(df=len(a) - 2).ppf(0.975)
    s_err = np.sum(np.power(residuals, 2))
    confidence_interval = t * np.sqrt((s_err/(len(a) - 2))*(1/len(a) + (np.pow
er((a - a.mean()), 2)/((np.sum(np.power(a,2))) - len(df)*(np.power(a.mean(),2
))))))
    prediction_interval = t * (np.std(residuals))
    return abs(confidence_interval), prediction_interval
```

In [ ]:
```

```