CS 696 Intro to Big Data
Fall Semester, 2019
Assignment 2

San Diego State University -- This page last updated 1/29/19

Due Feb 24 23:59

Issue 1 - Auto Fuel Economy

You can access the U.S. official fuel economy for vehicles sold in U.S. at:

https://fueleconomy.gov/feg/download.shtml

You are to investigate the the fuel economy for vehicles using **gasoline** from the years 2000 to 2019. We are interested in the vehicles sold be the companies General Motors, Ford, Chrysler, Honda and Toyota. Note that some companies like GM and Chrysler have sub-brands. For example Cadillac and Chevrolet are GM products. The U.S. government reports MPG (miles per gallon) in three way: city, highway and combined. We are only interested in **combined MPG**.

1. For each company collect the MPG sold by each company in the years 2000-2019. Produce the box plots per company for the MPG over those years. How do the companies compare?

2. Plot the yearly mean in the years 2000- 2019 with confidence interval of the mpg for each company. That is for each company compute the mean mpg over all vehicles sold by that company per year. What changes have there been in those years? How do the companies compare?

3. Plot the mpg for each company per year of their most fuel efficient vehicle each year. What changes have there been in those years? How do the companies compare?

Issue 2 Diet and Death

The data files for this part are available from the assignment web page. The date files we will be using are causes_of_death.csv, framingham.csv.

Causes of death. Plot the death rate for each disease over time from the data set causes_of_death.csv.

Diabetes and the population. The data set in framingham.csv contains information from the Framingham Heart Study of 5,209 adults.

First to check if the sample of people in the study is representative of the general population. We will use diabetes to test this. The CDC indicates that prevalence (percent) of diabetes was 0.93% at the time of the study. Our hypothesis:

Null Hypothesis: The probability that a participant within the Framingham Study has diabetes is equivalent to the prevalence of diagnosed diabetes within the population. (i.e., any difference is due to chance).

Alternative Hypothesis: The probability that a participant within the Framingham Study has diabetes is different than the prevalence of diagnosed diabetes within the population.

In the framingham.csv file the column DIABETES contains 1 for people with diabetes and 0 for those without.

4. What is the percentage of people in the study that have diabetes?

Now we need to compare this to the general population. Either a person is diagnosed as having diabetes or not. We can use the multinomial distribution to generate a sample of two values. Say we have an event that has .75 probability of occurring. Then the following will count the number of times the event does not occur and not occur in a sample of 1000.

```
two_value_probabilities = [0.25, 0.75]
sample_size = 1000
np.random.multinomial(sample_size, two_value_probabilities)
```

Using this we can compute the number of people we would expect to have diabetes in a sample of 5,000, which we need to convert to a percentage. Now do this 200 times.

5. Produce the histogram of the percent of people in your 200 samples with diabetes.

6. Compute the 95% confidence interval of the 200 values in #5

7. Is the study representative of the general population? Why or why not?

In the file framingham.csv the column TOTCHOL gives the total cholesterol of each person in the study. The column ANYCHD indicates if the person has any heard disease.

8. Plot the cholesterol values for the people with heart disease, for the people with out heart disease.

9. Compute the 95% confidence interval of the cholesterol values for the people with heart disease, for the people with out heart disease.

10. What can we deduce about cholesterol values and heart disease?

Instructions

You assignment 2 jupyter notebook should be self contained. All calculations and answers to the questions are to be in one notebook. This assignment requires you to use files, some of which are provided. You notebook needs to read the unmodified files, including names. Any needed modification to the files needs to be done in the notebook.

At the beginning of your notebook you should create variable that hold the path (plus name) of any input files that you use. It is likely that for grading purposes those paths will need to

change. I should be able to run your notebook using my input files by just changing the path to files at the top of your notebook.

Grading

10 points per problem.

What to turn in

You are to turn in a Jupyter Python notebook containing the code and answers to the questions above. Since Jupyter notebooks can contain text and code, before each problem indicate which problem it is in text, not in code comment.

Late Penalty

An assignment turned in 1-7 days late, will lose 5% of the total value of the assignment per day late. The eight day late the penalty will be 40% of the assignment, the ninth day late the penalty will be 60%, after the ninth day late the penalty will be 90%. Once a solution to an assignment has been posted or discussed in class, the assignment will no longer be accepted. Late penalties are always rounded up to the next integer value.