Do Diet Drinks Cause Weight Gain?

A project by Shimun Naher

Overview:

Though artificially-sweetened diet drinks (like Diet Pepsi or a cup of coffee with sucralose) contain no calories, it is theorized that drinking sweet diet drinks could increase cravings for other sweet food, or that the artificial sweeteners in diet drinks (like aspartame and sucralose) could directly cause weight gain. This article (http://www.vox.com/2016/11/28/13764656/diet-soda-metabolism-weight-loss-obesity) summarizes some of the recent research activity.

Here I will use hypothesis testing to replicate some of the analysis in this study (http://onlinelibrary.wiley.com/doi/10.1038/oby.2008.284/full). The original dataset is called the San Antonio Heart Study. It tracks 3,371 people living in San Antonio, Texas, over 7-8 years. For each person, it records (among many other things) how many diet drinks they reported drinking in a typical week, and the change in the person's Body Mass Index (BMI, a measure of weight adjusted for height) between the start and the end of the 7-8 year period. A change of 1 in BMI means that the person gained around 4-8 pounds, depending on their height.

I intend to compare the change in weight for those who had no diet drinks to the change in weight for those who did have diet drinks.

Program Code

```
In [2]: # import some helpful packages for data science
        import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        %matplotlib inline
        import helper_functions

        # set some parameters
        plt.rcParams['figure.figsize'] = (9,6)
        pd.options.display.max_rows = 20
```

```
In [3]: diet = pd.read_csv("~/shared/diet.csv")
        diet.head()
```

```
Out[3]:
          ID Typical diet drinks per week BMI change

           0                6             -4.400649

           1                0              0.952995

           2                4              2.710194

           3                6             -0.276764

           4                0              6.120786
```

```
In [22]: def check_nonzero(list_of_numbers):
          list_for_output = []  # This list is used store the results at the end of each loop iteration

         for number in list_of_numbers:
          if number > 0:
          list_for_output.append(True)
          else:
          list_for_output.append(False)

          return list_for_output
```

```
In [5]: new_column_values = check_nonzero(diet["Typical diet drinks per week"])
        diet["Drink"] = new_column_values
```

Null hypothesis: In the population of all people who lived during this 7-8 year period, the average BMI change among diet drink drinkers was the same as or less than average BMI change among nondrinkers.

Alternative hypothesis: The average BMI changes for diet drinkers was not the same as or less than BMI changes among nondrinkers.

```
In [6]: diet_averages = diet[["BMI change"]].groupby(diet["Drink"]).mean()
        diet_averages
```

```
Out[6]:
                BMI change

        Drink

        False   1.019248

         True   1.504488
```

```
In [29]: def my_test_statistic(average_for_drinkers, average_for_abstainers):   #
```

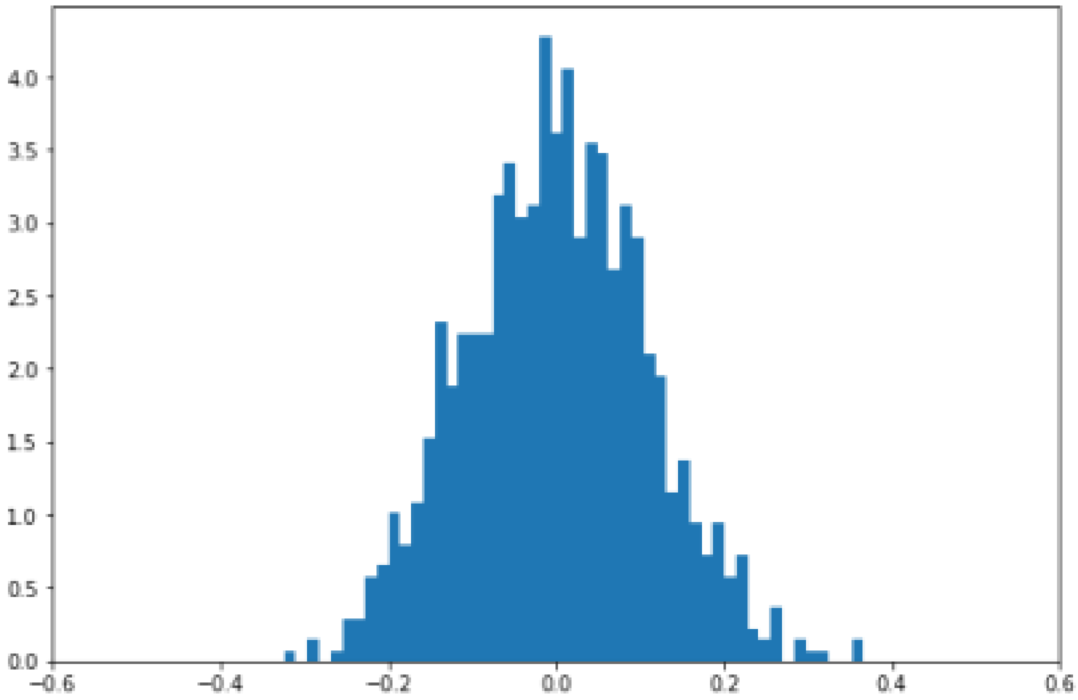
```
    # The test statistic is computed as the difference between the drinker and abstainer averages
      difference = average_for_drinkers - average_for_abstainers
    return difference

observed_statistic = 1.504488 - 1.019248
```

I will use a simulation to study the value of the test for different groupings. Every time I run the simulation I will get an estimate of the difference between drinkers and non drinkers. Doing this many times allows me to estimate almost all the possible differences between the drinkers and nondrinkers assuming there is no difference between the two groups. If I find that the difference for the real data is much larger than what I find in the simulation, it is strong evidence that the difference between the two groups is not due to chance.

```
In [11]: simulation_results = helper_functions.simulation(diet, my_test_statistic
          , 1000)
```

```
In [12]: plt.hist(simulation_results, bins = 50, density = True)
          plt.xlim([-0.6,0.6]);
```

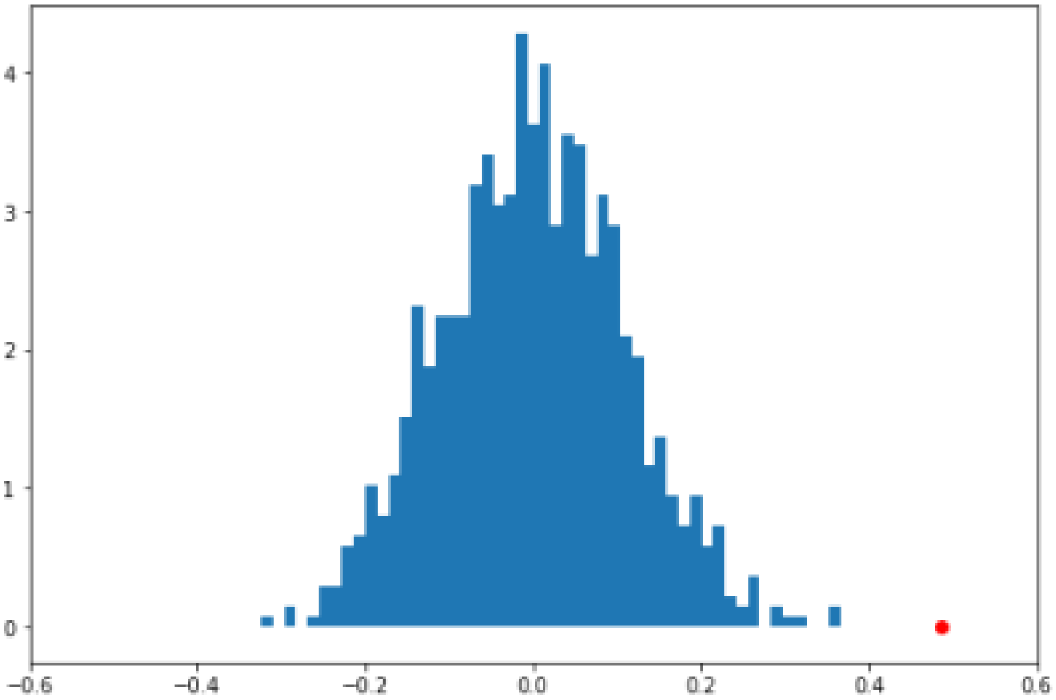


I plotted the observed test statistic as a red dot on the horizontal axis.

```
In [14]:
          plt.hist(simulation_results, bins = 50, density = True)
          plt.xlim([-0.6,0.6])

          plt.scatter([observed_statistic],[0], color = 'red')

Out[14]: <matplotlib.collections.PathCollection at 0x7f493a5e3590>
```



Before I reject or fail to reject the null hypothesis, I computed the p-value.

```
In [27]:  num_greater = 0
         for entry in simulation_results:
          if entry >= observed_statistic:
          num_greater = num_greater + 1
          else:
          num_greater = num_greater + 0
```

```
In [28]: p_value = num_greater / 1000
          p_value

Out[28]: 0.0
```

Based on the p-value, I reject the null hypothesis. Therefore, it can be claimed that diet drinks do cause weight gain.