Data Inferencing for Capstone Project #1

The dataset hardly has any numerical data other than the timestamps for each row. Because of this, it would be difficult, if not impractical, to brainstorm statistical questions for testing hypotheses. Fortunately, we can use the timestamps to group data into separate populations and perform tests based on populations.

The first question I asked myself was whether there is an increase of the mean daily number of robocalls (I used the term loosely to refer to any unsolicited calls). To approach this, I used the earliest and latest full years, 2015 and 2019 respectively, as the two independent populations and then extract the dates for each timestamp.

```
# The column ‘Issue Occurred’ contains datetime values
date_s = phone_df['Issue Occurred'].apply(lambda dt: dt.date())

>> 2014-10-31
   2014-10-31
   ...
```

Then, I extract all that contain the years 2015 and 2019 and count the occurrences for each unique date:

```
mask2015 = (date_s >= date(2015,1,1)) & (date_s <= date(2015,12,31))
mask2019 = (date_s >= date(2019,1,1)) & (date_s <= date(2019,12,31))

phn2015 = date_s.loc[mask2015].value_counts()
phn2019 = date_s.loc[mask2019].value_counts()
```

This test must ensure that `phn2015` and `phn2019` account for every possible date (including those without an occurrence) of the year, so both must have a length of 365. Also, to compute the observed mean, we would compute it as so:

```
mean_rbcall = np.sum(date_s.value_counts()) / <time_delta>
```

The bootstrap method will be used for this test for each population. It will perform 10000 bootstrap repetitions in respect the assumption that the means of 2015 and 2019 are equal (the null hypothesis is true) with $\alpha = 0.05$ **one-tailed**.

The results we get is that the p-value, $p = 0.352$, is greater than α, implying that we cannot reject the null hypothesis. While there may be an increase, it does not provide sufficient evidence that the average daily amount of robocalls is greater in 2019 than in 2015.

The next question was whether there were proportionally more complaints reported by wireless users over time. As with the last question, I used the years 2015 and 2019 as the populations to compare with $\alpha = 0.05$ as the statistical significance. However, since it only focuses on the "Method" column, we only need to extract the methods columns associated with each year.

To do this, we select only the rows with timestamps matching each year, then extract the "Method" column from the selected rows. Since the question focuses on the "Wireless" method, we compute the following as provided below:

$$\widehat{p}_{2015} = \frac{\text{"}Wireless\text{" } in\ 2015}{Entries\ of\ 2015} = \frac{k_{2015}}{n_{2015}}$$

$$\widehat{p}_{2019} = \frac{\text{"}Wireless\text{" } in\ 2019}{Entries\ of\ 2019} = \frac{k_{2019}}{n_{2019}}$$

$$\widehat{p} = \frac{k_{2015}+k_{2019}}{n_{2015}+n_{2019}}$$

For testing the difference of proportions, we would use the z-test and compute as below:

$$z = \frac{\widehat{p}_{2019}-\widehat{p}_{2015}}{\sqrt{\widehat{p}(1-\widehat{p})(\frac{1}{n_{2015}}+\frac{1}{n_{2019}})}}$$

In order to reject the null hypothesis, we would need the $z > z^*$. The critical value z^* in respect to α in this case would be 1.64. However, we got $z = -1.92$. Clearly, $z < z^*$, implying that we cannot reject the null hypothesis. In fact, the observed proportion of complaints from wireless users was smaller in 2019 than in 2015.

The limited numerical data limited our ability to perform any kind of hypothesis testing, even more from testing the correlation of two independent variables. Nonetheless, we can only hope the findings will be vital to battling unsolicited calls.