MAT 128 Lab 18: Hypothesis Testing

Steps in Hypothesis Testing:
1) Write down the *null hypothesis* and the *alternative hypothesis*. The null hypothesis says that the data was generated randomly from some simple model. You should be able to simulate data by assuming your null hypothesis is true. The alternative hypothesis says that something other than chance caused the data.

2) Find a *test statistic* that we can use to decide between the null and alternative hypothesis. The test statistic should be something we can measure or compute that would tend to have a different values if the alternative hypothesis is true than if the null hypothesis is true.

3) Simulate the distribution of the test statistics under the null hypothesis. Use a histogram to visualize it.

4) Decide whether the data is likely when the null hypothesis is true. If it is not, we can reject the null hypothesis.

Data

All motor vehicle accidents in NYC:
https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95

On the website, filter the data by data to download one month's worth of accidents.

Motivating question:
Are accidents at night more dangerous?

Hypotheses:
Null hypothesis: Accidents at night are just as dangerous as accidents during the day.
Alternative hypothesis: Accidents at night are more dangerous than accidents during the day.

We will use data from February 2018, and thus will define night as 6pm - 6am, based on sunrise and sunset times during that month.

Test Statistic:
Number of accidents that have an injury or a fatality.

Therefore, we can make our null hypothesis more precise: accidents at night have the same proportion of injuries/fatalities as accidents during the day.

Using what we have already learned in the course, we can:
1) estimate the probability (p) that an accident has an injury or fatality (ignoring time of day)
2) count the number (N) of accidents that occurred at night in our dataset (call it N)
3) use simulation to find the distribution of accidents with an injury or fatality at night under the null hypothesis:

- Repeatedly simulate N accidents where each accident has p probability of having an injury or fatality
- Plot a histogram of the number of accidents in each simulation that had an injury or fatality

3) count the number of accidents that occurred at night and had an injury or fatality in our dataset
4) compare the number of night accidents with an injury or fatality in our dataset with the histogram. Does this data fit the model we used to generate the histogram?

Hints:

1)
To make a filter/condition variable with multiple conditions, put brackets around each one:

```
injury_fatality_filter = (accidents['NUMBER OF PERSONS INJURED'] > 0) |
(accidents['NUMBER OF PERSONS KILLED'] > 0)
```

Note: The above code is all on one line.

2) We can convert the TIME column into a datatime one:

```
accidents['TIME'] = pd.to_datetime(accidents['TIME'])
```

This will add a date, but we can ignore it.

To get the hour:

```
accidents['TIME'].dt.hour
```

3) We can simplify the steps for sampling from the night accidents as follows:

```
population = [0,1]       # 1 will represent injuries/fatalities
weight = [prob_no_injuries_fatalities, prob_injuries_fatalities]

sample_array = np.random.choice(population,p=weight,size=num_night_accidents)
count = sum(sample_array)   # counts the number of 1's in the array, which is the number
                            # of accidents with injuries/fatalities in this sample
```

Challenges:
Make and test another hypothesis. For example, is one borough more dangerous for driving than another?