MAT 128 Lab 18 Continued

We have already estimated the probability that an accident causes injury or death.

The steps that remain are:
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

4) count the number of accidents that occurred at night and had an injury or fatality in our dataset
5) 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?

Here are some more details for each step:

2) Make a smaller dataset that only contains accident that happen at night, and find its length:

a) To compare the times more easily, we make the TIME column be a Pandas datetime column:

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

b) Decide what hours are considered night. Here we use 6pm - 6am. The datatime type uses a 24 clock, so this becomes 18:00 - 6:00.

Set up a filter to find all accidents where the hour is 18 or more, or 5 or less. (Since we only check the hour, this will catch both 18:01 (6:01pm) and 5:59am, but not 6:01am.)

```
night_filter = (accidents['TIME'].dt.hour <= 5) | (accidents['TIME'].dt.hour >= 18)
```

c) Use this filter to create a new dataset of just the night accidents and find its length. This number is called N below.

---

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

This is similar to what we did in Lab 17. Suppose that in step 2, we calculated that there are 4000 accidents at night. Suppose in step 1 you estimated the probability of a dangerous accident (one causing injury or death) as 0.2. (You should use the numbers you calculated in steps 1 and 2.)

We are going to simulate 4000 accidents that happen at night, assuming that each one has a 0.2 probability of causing injury or death.

a) As in Lab 17, first set up the simulation. To make it easier to count how many dangerous accidents there are in each simulation, we will use 1 for dangerous and 0 for if there were no injuries or deaths.

```
population = [1,0]
weight = [0.2,0.8]
```

One sample of 4000 accidents using this distribution:

```
sample = np.random.choice(population, p=weight,size = 4000)
```

b) We can count the number of accidents causing injury or death by summing up the array (since each such accident contributes 1 to the sum):

```
dangerous_count = sum(sample)
```

c) Add in a loop to repeat this sampling procedure many times. Store the counts in a list.

d) Plot a histogram of the simulated counts.

---

4) Count the number of accidents that occurred at night and had an injury or fatality in our dataset.

We can do this by applying both our night filter and our dangerous (injury or death) filter to the original data set. This gives us a dataset of just the dangerous night accidents, and we can compute its length.

---

5) 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?

Answer this question on the classwork sheet.