

# Simulating from the Null

December 5, 2017

## 0.0.1 Simulating From the Null Hypothesis

Load in the data below, and follow the questions to assist with answering the quiz questions below.

```
In [1]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

%matplotlib inline
np.random.seed(42)

full_data = pd.read_csv('coffee_dataset.csv')
sample_data = full_data.sample(200)
```

```
In [37]: sample_data.head()
```

```
Out[37]:
```

	user_id	age	drinks_coffee	height
2402	2874	<21	True	64.357154
2864	3670	>=21	True	66.859636
2167	7441	<21	False	66.659561
507	2781	>=21	True	70.166241
1817	2875	>=21	True	71.369120

1. If you were interested in if the average height for coffee drinkers is the same as for non-coffee drinkers, what would the null and alternative be? Place them in the cell below, and use your answer to answer the first quiz question below.

$$H_0 : \mu_c = \mu_{nc} \quad H_1 : \mu_c \neq \mu_{nc}$$

2. If you were interested in if the average height for coffee drinkers is less than non-coffee drinkers, what would the null and alternative be? Place them in the cell below, and use your answer to answer the second quiz question below.

$$H_0 : \mu_c \geq \mu_{nc} \quad H_1 : \mu_c < \mu_{nc}$$

3. For 10,000 iterations: bootstrap the sample data, calculate the mean height for coffee drinkers and non-coffee drinkers, and calculate the difference in means for each sample. You will want to have three arrays at the end of the iterations - one for each mean and one for the difference in means. Use the results of your sampling distribution, to answer the third quiz question below.

```

In [29]: cym = []
        cnm = []
        diffs = []
        for _ in range(10000):
            s = sample_data.sample(200, replace=True)
            cy = s[s["drinks_coffee"] == True]["height"].mean()
            cn = s[s["drinks_coffee"] == False]["height"].mean()
            cym.append(cy)
            cnm.append(cn)
            diffs.append(cy - cn)

        print("done!")

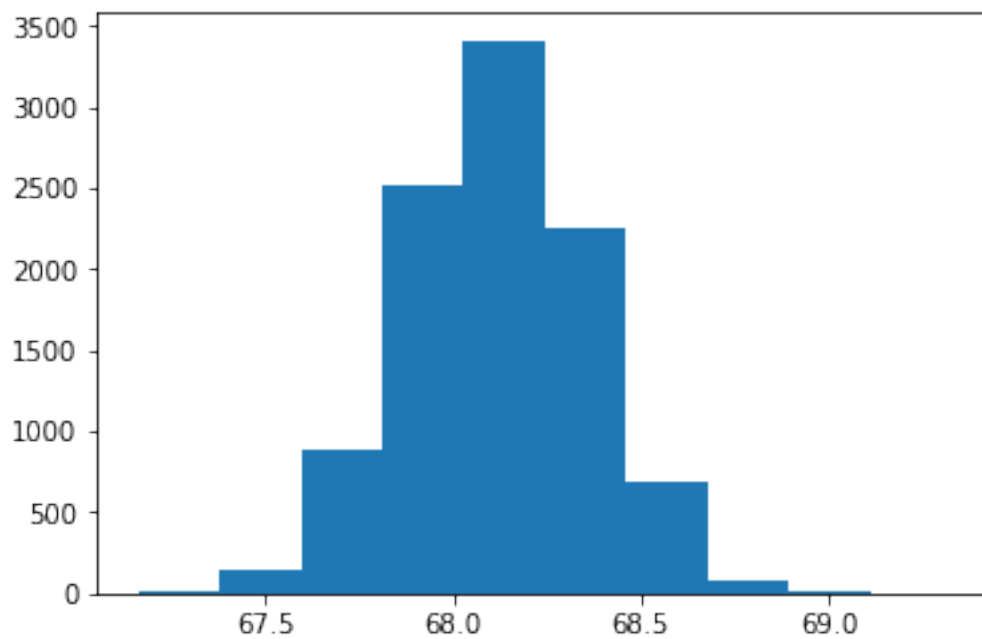
```

done!

```

In [30]: plt.hist(cym);

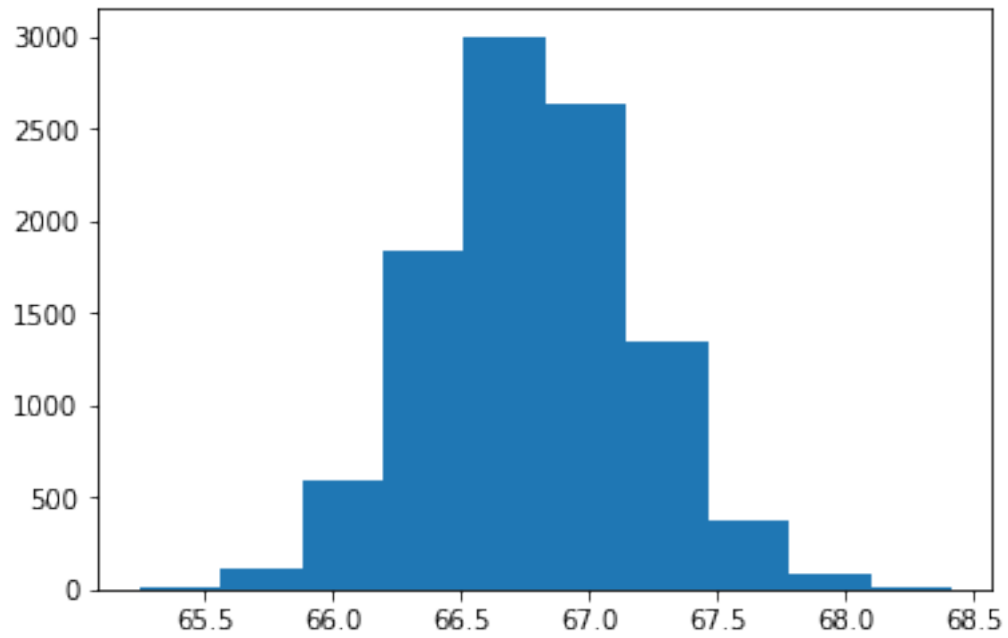
```



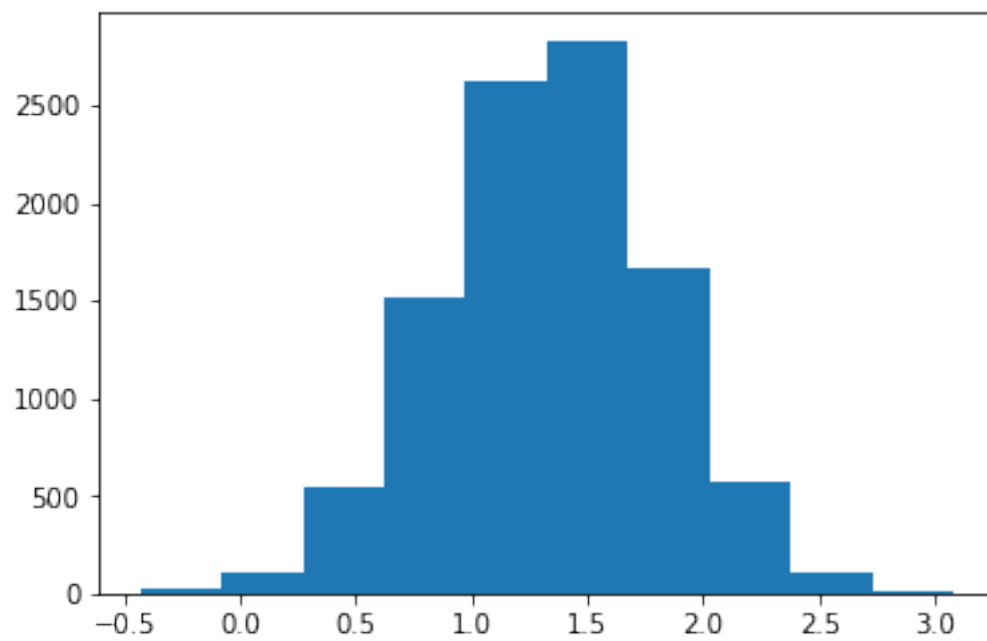
```

In [31]: plt.hist(cnm);

```



```
In [32]: plt.hist(diffs);
```



```
In [34]: np.std(cnm)
```

```
Out[34]: 0.40473398138941474
```

```
In [35]: np.std(cym)
```

```
Out[35]: 0.24091955569986864
```

```
In [38]: sd = np.std(diffs)
         sd
```

```
Out[38]: 0.46953537512085691
```

4. Now, use your observed sampling distribution for the difference in means and [the docs](#) to simulate what you would expect the sampling distribution to be if the null hypothesis is true. You can do this by recentering your distribution at zero. Also, calculate the observed sample mean difference in `sample_data`. Use your solutions to answer the last questions in the quiz below.

```
In [39]: null_vals = np.random.normal(0, sd, 10000)
         print("done")
```

done

```
In [43]: plt.hist(null_vals);
         plt.axvline(x=0, color="red");
```

