Confidence Intervals - Part I

First let's read in the necessary libraries and the dataset. You also have the full and reduced versions of the data available. The reduced version is an example of you would actually get in practice, as it is the sample. While the full data is an example of everyone in your population.

In [2]:
```
import pandas as pd
import numpy as np

np.random.seed(42)

coffee_full = pd.read_csv('coffee_dataset.csv')
coffee_red = coffee_full.sample(200)#this is the only data you might act
ually get in the real world.

coffee_red.head()
```

Out[2]:

	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. What is the proportion of coffee drinkers in the sample? What is the proportion of individuals that don't drink coffee?

In [3]:
```
coffee_red['drinks_coffee'].mean()
```

Out[3]: 0.5949999999999997

2. Of the individuals who drink coffee, what is the average height? Of the individuals who do not drink coffee, what is the average height?

In [4]:
```
coffee_red[coffee_red['drinks_coffee']==True].height.mean()
```

Out[4]: 68.119629908586163

In [5]:
```
coffee_red[coffee_red['drinks_coffee']==False].height.mean()
```

Out[5]: 66.784922799278775

3. Simulate 200 "new" individuals from your original sample of 200. What are the proportion of coffee drinkers in your bootstrap sample? How about individuals that don't drink coffee?

In [6]:
```
bootstrap = coffee_full.sample(200, replace=True)
bootstrap.head()
```

Out[6]:

	user_id	age	drinks_coffee	height
930	5963	<21	False	63.718178
288	5179	<21	False	63.164816
723	7915	<21	False	63.382105
2940	7525	>=21	False	70.266762
2319	7125	<21	False	73.939576

In [7]:
```
bootstrap['drinks_coffee'].mean()
```

Out[7]: 0.56999999999999995

In [8]:
```
1 - bootstrap['drinks_coffee'].mean()
```

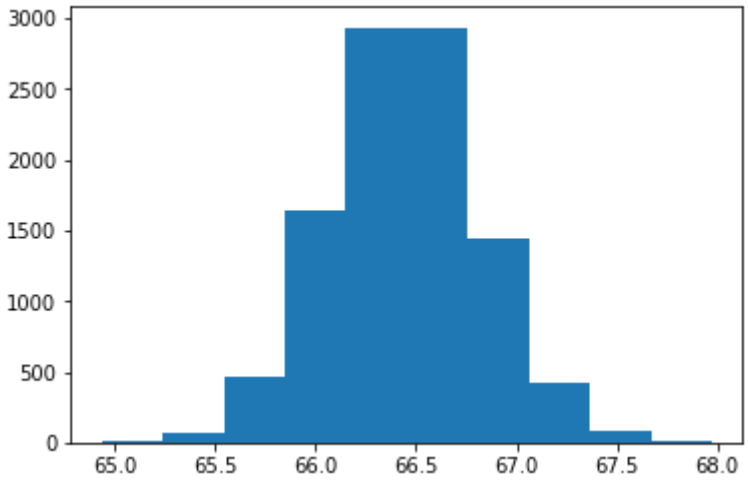
Out[8]: 0.43000000000000005

4. Now simulate your bootstrap sample 10,000 times and take the mean height of the non-coffee drinkers in each sample. Each bootstrap sample should be from the very first sample of 200 data points. Plot the distribution, and pull the values necessary for a 95% confidence interval. What do you notice about the sampling distribution of the mean in this example?

In [15]:
```
import matplotlib.pyplot as plt
%matplotlib inline
bootstrap_samples = []

for _ in range(10000):
    sample = coffee_full.sample(200, replace=True)
    bootstrap_samples.append(sample[sample['drinks_coffee']==False].heig
ht.mean())

plt.hist(bootstrap_samples);
```



In [16]:
```
np.percentile(bootstrap_samples, 2.5), np.percentile(bootstrap_samples,
97.5)
```

Out[16]:
```
(65.704303149654521, 67.188680594170521)
```

In [ ]:
```

```

5. Did your interval capture the actual average height of non-coffee drinkers in the population? Look at the average in the population and the two bounds provided by your 95% confidence interval, and then answer the final quiz question below.

In [17]:
```
coffee_full[coffee_full['drinks_coffee']==False].height.mean()
```

Out[17]:
```
66.443407762147004
```

In [18]:
```
# YES,
# our interval did consist of the mean of the original non-coffee drinke
rs
```

In [ ]:
```

```