Sampling Distributions - Difference in Means

April 23, 2019

0.0.1 Confidence Interval - Difference In Means

Here you will look through the example from the last video, but you will also go a couple of steps further into what might actually be going on with this data.

```
In [10]: 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)
         sample_data.head()
```

Out[10]:

	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. For 10,000 iterations, bootstrap sample your sample data, compute the difference in the average heights for coffee and non-coffee drinkers. Build a 99% confidence interval using your sampling distribution. Use your interval to start answering the first quiz question below.

```
In [13]: diff = []
         for x in range (10000):
             bootstrap = sample_data.sample(2000, replace = True)
             drink_coffee = bootstrap[bootstrap['drinks_coffee'] == True]['height'].mean()
             drink_not_coffee = bootstrap[bootstrap['drinks_coffee'] == False]['height'].mean()
             diff.append(drink_coffee - drink_not_coffee)

         np.percentile(diff, 0.5), np.percentile(diff, 99.5)
```

```
Out[13]: (0.9495709687100535, 1.7139738289648443)
```

2. For 10,000 iterations, bootstrap sample your sample data, compute the difference in the average heights for those older than 21 and those younger than 21. Build a 99% confidence interval using your sampling distribution. Use your interval to finish answering the first quiz question below.

```
In [17]: diff = []
         for x in range (10000):
             bootstrap = sample_data.sample(2000, replace = True)
             under21 = bootstrap[bootstrap['age'] == '<21']['height'].mean()
             over21 = bootstrap[bootstrap['age'] != '<21']['height'].mean()
             diff.append(over21 - under21)

         np.percentile(diff, 0.5), np.percentile(diff, 99.5)

Out[17]: (3.9761777574331427, 4.5213706863427126)
```

3. For 10,000 iterations bootstrap your sample data, compute the **difference** in the average height for coffee drinkers and the average height for non-coffee drinkers for individuals **under** 21 years old. Using your sampling distribution, build a 95% confidence interval. Use your interval to start answering question 2 below.

```
In [27]: diff = []
         for x in range (10000):
             bootstrap = sample_data.sample(2000, replace = True)
             under21_drink_coffee = bootstrap.query("age == '<21' and drinks_coffee == True")['h
             under21_not_drink_coffee = bootstrap.query("age == '<21' and drinks_coffee == False
             diff.append(under21_not_drink_coffee - under21_drink_coffee)

         np.percentile(diff, 2.5), np.percentile(diff, 97.5)

Out[27]: (1.605003831398858, 2.0840539253537296)
```

4. For 10,000 iterations bootstrap your sample data, compute the **difference** in the average height for coffee drinkers and the average height for non-coffee drinkers for individuals **over** 21 years old. Using your sampling distribution, build a 95% confidence interval. Use your interval to finish answering the second quiz question below. As well as the following questions.

```
In [32]: diff = []
         for x in range (10000):
             bootstrap = sample_data.sample(2000, replace = True)
             over21_drink_coffee = bootstrap.query("age != '<21' and drinks_coffee == True")['he
             over21_not_drink_coffee = bootstrap.query("age != '<21' and drinks_coffee == False"
             diff.append(over21_not_drink_coffee - over21_drink_coffee)

         np.percentile(diff, 2.5), np.percentile(diff, 97.5)

Out[32]: (2.713948534534922, 3.5108244904130945)

In [ ]:
```