Intro to Data Science - Lab 5

DATA 1501 — Dr. Mihail
Department of Computer Science
Valdosta State University

September 22, 2021

Introduction

In this lab, you will download a dataset of 1000 height measurements. We will consider this to be our "population". The population's true mean is 72 and its true standard deviation is 4. You will then create samples with an increasing number of observations, and compute the mean and standard deviation. You will record them and plot them.

Part 1 (10 points)

In this part, you will download the dataset of heights. Create a new Colab notebook and run the following code in a code cell:

```
!wget https://cs.valdosta.edu/~rpmihail/DATA1500/lab5/heights.csv
```

Confirm that the file was downloaded and no errors reported from mispellings. Next, create a code cell and run the following code:

```
import pandas as pd
import numpy as np
df = pd.read_csv('heights.csv')
df
```

Confirm you can see a few records in the dataset.

Part 2 (60 points)

In this part, you will create samples of varying size by randomly picking observations from the population. The code below creates a sample of size 10, and prints out the sample's mean and standard deviation:

```
# code to create a sample and compute its mean and standard deviation
observations = 10
indices = np.random.randint(0, 1000, observations)
sample = df.iloc[indices]
```

```
mean = sample.mean()[0]
std =  sample.std()[0]
print("Mean of sample sized %i observations is %.4f" % (observations, mean))
print("St dev of sample sized %i observations is %.4f" % (observations, std))
```

A successful run of the code can be seen below:

```
Mean of sample sized 10 observations is 73.7370
Standard deviation of sample sized 10 observations is 2.7315
```

Your task is to compute and record the mean and standard deviation for different number of observations. To do that, you will first create two empty lists where the observations will be stored. Create a new code cell and insert the following code:

```
means = []
stds = []
```

Next, the code listing below will compute and record means standard deviations of samples with the following sizes: N=2, 3, 5, 10, 15, 20, 25, 50, 100, 200, 500, 1000.

Create a new code cell and copy/paste the following code:

```
observations_list = [2, 3, 5, 10, 15, 20, 25, 50, 100, 200, 500, 1000]
for observations in observations_list:
    indices = np.random.randint(0, 1000, observations)
    sample = df.iloc[indices]
    mean = sample.mean()[0]
    std =  sample.std()[0]
    means.append(mean)
    stds.append(std)

means = pd.DataFrame({"Nr_Obs":observations_list, "Means":means})
stds = pd.DataFrame({"Nr_Obs":observations_list, "Stds":stds})

print(means)
print(stds)
```

Confirm there are no errors, and the code prints out 12 computed means and standard deviations. Now create a new code cell and run the following code to plot the means and standard devations:

```
means.plot("Nr_Obs", "Means")
stds.plot("Nr_Obs", "Stds")
```

Confirm you see the two plots.

Part 3 (30 points)

Create a Microsoft Word document and paste a screen shot of the graphs you just created. Also answer the following question:

- Explain what you observed about the estimated means and standard deviations compared to their true values as the number of observations increases.
- What have you learned in this lab?

Due Date: Before Midnight on Sunday, September 26th.