IT166 Lab 6

In this lab, we will use some basic Python programming skills as specifued below to calculate some statistic parameters (e.g., mean, standard deviation) of a given population.

5:44

Objectives

- Be able to write Python programs to find minimal, maximal, mean, median, and standard deviation values of a collection of numerical data.
- Be able to write Python programs that conduct sampling on a given porpulation.
- Be able to write Python programs that verify Central Limit Theorem.

Preparation

- Launch the Jupyter notebook.
- Rename the notebook page as “Lab6”.

Please provide solutions to the problems below:

Background:

Statistic properties (e.g., mean, standard deviation) of a population is important for data analysis. In reality, we usually cannot collect the complete dataset of a whole population. Thus, we often take random samples of the population and estimate these statistic parameters based on the samples.

Given a sample of n observations from a population, we can calculate estimates of the population mean, standard deviation, and various other population characteristics (parameters). Prior to obtaining the complete dataset, there is uncertainty as to which of all possible samples will occur. Because of this, estimates such as $\bar{x}$, and s will vary from one sample to another. The behavior of such estimates in repeated sampling is described by what are called **sampling distributions**. Any particular sampling distribution will give an indication of how close the estimate is likely to be to the value of the parameter being estimated.

The **Central Limit Theorem** states that if you have a population with mean μ and standard deviation σ and take sufficiently large $n \geq 30$ random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed.

Sampling can be done with or without replacement. If with replacement, the same data point could appear more than once in a sample.

Problem 1

In this lab, we assume the following 100 weights to be a complete population, and thus can accurately calculate some important parameters of this dataset:

[112.99,136.49,153.03,142.34,144.3,123.3,141.49,136.46,112.37,120.67,127.45,114.14,125.61,122.46,116.09,140,129.5,142.97,137.9,124.04,141.28,143.54,97.9,129.5,141.85,129.72,142.42,131.55,108.33,113.89,103.3,120.75,125.79,136.22,140.1,128.75,141.8,121.23,131.35,106.71,124.36,124.86,139.67,137.37,106.45,128.76,145.68,116.82,143.62,134.93,147.02,126.33,125.48,115.71,123.49,147.89,155.9,128.07,119.37,133.81,128.73,137.55,129.76,128.82,135.32,109.6[illegible]7,132.75,103.53,124.73,129.31,134.02,140.4,102.84,128.52,120.3,138.6,132.96,115.62,12[illegible]4.63,121.9,155.38,128.94,129.1,139.47,140.89,131.59,121.12,131.51,136.55,141.49,140.[illegible].14,133.46,131.8,120.03,123.1,128.14,115.48]

5:44

Data source:
http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Dinov_020108_HeightsWeights

Create a Python List to store the population; and then calculate the maxum, minumum, mean, standard deviation)

Mean: $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$

Standard deviation: $\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$

(Here, x_i is a data point of the population)

Please note: in this lab, you should use ***for loop*** to interate all data and perform required calculations. You are **not** allowed to directly use built-in functions max(); min(); sum(); and mean() from statistics module. You can use len() if needed.

The example of the expected output is shown below:

The mean of this population is 129.21

The min and max of this population is 97.90 and 155.90

The standard deviation of this population is 12.06

Problem 2

Please do sampling with and without replacement from the population.

- Sampling with replacement (the same data point could appear more than once in a sample)
- Sampling without replacement (the same data point could appear only once in a sample)

For each sampling method, please use the following sampling sizes 5, 10, and 30, respectively. Please print each sample.

One example of the expected output is shown below:

The sample with size 5 is: [140.1, 123.49, 128.14, 125.79, 142.47]

Problem 3

Let's observe if the Central Limit Theorom works in this population.

Please do sampling with replacement 100 times with the same sampling size 30 from the population. For each sample, calculate its mean, and then convert the mean result to int and save to a list. Please print this list.

One example of the expected output is shown below:

The mean values of 100 samplings: [129, 127, 129, 132, 129, 131, 130, 130, 129, 125, 128, 128, 131, 129, 131, 129, 130, 127, 131, 127, 128, 134, 129, 128, 125, 133, 127, 123, 127, 126, 125, 128, 129, 132, 131, 127, 134, 130, 128, 129, 129, 131, 129, 131, 124, 125, 129, 125, 127, 127, 127, 126, 132, 128, 128, 131, 129, 131, 131, 125, 127, 126, 133, 127, 126, 127, 126, 125, 128, 129, 127, 128, 126, 126, 128, 129, 133, 130, 128, 130, 127, 129, 126, 131, 128, 130, 130, 130, 129, 133, 130, 127, 129, 128, 131, 126, 130, 130, 129, 128]

Assuming the mean list above called sampleMeanIntList, use the following code to observe if it roughly follows a normal distribution:

```
from collections import Counter

Counter(sampleMeanIntList)
```

One example of the expected output is shown below:

```
Counter({123: 1,
         124: 1,
         125: 7,
         126: 9,
         127: 15,
         128: 15,
         129: 19,
         130: 12,
         131: 12,
         132: 3,
         133: 4,
         134: 2})
```