MAT 128 Lab 11: Central Limit Theorem

Central Limit Theorem

As the sample size n increases without limit (n $\rightarrow \infty$), the shape of the distribution of the sample means taken from a population with mean μ and standard deviation σ will approach a normal distribution. This distribution will have mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$

Motivating Question: Can we see the Central Limit Theorem in action with real data?

Data: We will use the 311 Service Request data available from OpenDate NYC:

https://nycopendata.socrata.com/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9

Use the filters to select and download (as a CSV file) 1 year's worth of requests with a Descriptor value of Pothole.

Exploring the Data:

1) As we have done many times, read the data from your CSV file into your Python program.

   You might get a message in the shell saying:

```
DtypeWarning: Columns (17,18) have mixed types.  Specify dtype
option on import or set low_memory=False.
```

This message means that some of the columns (in this case, columns 17 and 18 where the numbering starts at 0) contain data of two different types (ex. integer and string). Pandas starts reading in the file assuming the column data is the type first encountered, but when it encounters the other type, it has to update its assumption and re-read the entire CSV file. This takes a lot of time.

Instead, we can tell pandas the the types of those two columns (which are `‘Landmark’` and `‘Facility Type’`) will be strings (`‘str’`):

```
dtypes = {‘Landmark’:’str’,’Facility Type’:’str’}

requests = pd.read_csv(“311_Service_Requests.csv”,dtype = dtypes,
parse_dates=[‘Created Date’])
```

We have also added the option `parse_dates =[‘Created Date’]` which tells the `read_csv` command that the column `Created Date` should be interpreted as a date and time.

Otherwise, to interpret the column Created Date as a date and time, we would have added the extra line:

```
requests[‘Created Date’] = pd.to_datetime(requests[‘Created
Date’])
```

2) For each date, count the number of requests there are:

```
pothole_counts = requests[‘Created Date’].dt.date.value_counts()
```

3) Plot a histogram of how many requests were made each day:

```
pothole_counts.plot.hist(bins = 20,title=”Number of pothole requests per day”)
```

What do you notice about this plot?

4) Plot a line graph of the number of pothole requests made each day during the year. Do you see any patterns?

Testing the Central Limit Theorem

In this section, the population will be all pothole service requests made during the year.

1) As in Lab 5, print out the population mean.

2) Print out the population standard deviation (as with the variance, we use the option `ddof=0` since this is the population and we do not need to account for the sample bias):

```
sigma = pothole_counts.std(ddof=0)
print(“Population standard deviation (sigma):,”, sigma)
```

3) As in Lab 5, create a loop to repeatedly sample from the population, calculate the sample mean, and store it in a list.

4) Plot a histogram of the sample means. What shape does this histogram have?

5) Compute the mean of the sample means. Does this match the Central Limit Theorem?

6) Compute the standard deviation (using the option `ddof=0`) of the sample means. Does this match the Central Limit Theorem? (Hint: you may have to do a further computation)

7) What happens to the mean and standard deviation of the sample means when you increase the sample size? Does this match the Central Limit Theorem?