# Assignment

## Data Selection:

You may take any dataset of your choice. It should be a univariate dataset. Also, provide the URL of your dataset. The minimum length of the dataset should be 20.

## Analysis to be done:

### Q.1 Descriptive Statistics

**1.** Type of data, description of the data.
**2.** For the data, compute the following:
   a) Mean, Median, Mode and Variance
   b) Measure of Skewness and Kurtosis
**3.** Create the following plots:
   a) Histogram
   b) One of the following: Line chart/bar chart/pie chart/dot plot/ Stem and leaf plot/scatter plot (whichever is applicable) to visualize data.

Comment on your observations.

### Q.2 Central Limit Theorem (CLT)

Generate a random sample from a given distribution for different sample sizes.

a) Plot its PMF/PDF and CDF.
b) Calculate its first four sample moments and theoretical moments and compare the results.
c) Show that the chosen distribution tends to the normal distribution/another distribution under certain conditions.

---

***Method of Data Selection:***

Make use of your name or surname's initials to choose the data.

**Example 1: Vasudha Upadhyay**

Data may be from the country Vietnam or Uruguay, or any other country starting with U or V.

**Example 2: Komal Yadav**

Data may be related to Kidney Catheter replacements or the Survival time of Kidney patients.

**Example 3: Isha**

Data may be related to insurance, such as insurance claims or infant mortality.

***Assignment Allotment Method***

**(a) Method of selection of Programming language**

- **R**: If (Last 3 digits of Enrolment Number **mod** 3) = 0
- **Python**: If (Last 3 digits of Enrolment Number **mod** 3) = 1
- **MATLAB**: If (Last 3 digits of Enrolment Number **mod** 3) = 2

**(b) Method of Selection of Distribution for Q. 2, and for 2 c): Using the last digit of your Enrolment Number**

- **0:** Binomial Distribution, 2c) Normal approximation to Binomial.
- **1:** Binomial Distribution, 2c) Poisson approximation to Binomial.
- **2:** Poisson Distribution, 2c) Normal approximation to Poisson.
- **3:** Negative Binomial Distribution, 2c) Normal approximation to Negative Binomial.
- **4:** Negative Binomial Distribution, 2c) Poisson approximation to Negative Binomial.
- **5:** Geometric Distribution, 2c) Normal approximation to Geometric.
- **6:** Gamma Distribution, 2c) Normal approximation to Gamma.
- **7:** Hypergeometric Distribution, 2c) Binomial approximation to Hypergeometric.
- **8:** Hypergeometric Distribution, 2c) Normal approximation to Hypergeometric.
- **9:** Hypergeometric Distribution 2c) Poisson approximation to Hypergeometric.

## ***Illustration:***

**Name:** Vasudha Upadhyay
**Enrollment No.:** RSS2023002
**Coding platform:** Remainder of (002/3)=2 ⇒ MATLAB
**Dataset selected:** Distribution of Ethnic groups among women in Vietnam in 2005.
**URL:** https://dhsprogram.com/pubs/pdf/AIS3/AIS3.pdf
**Basis of data selection:** Initial of my name.

---

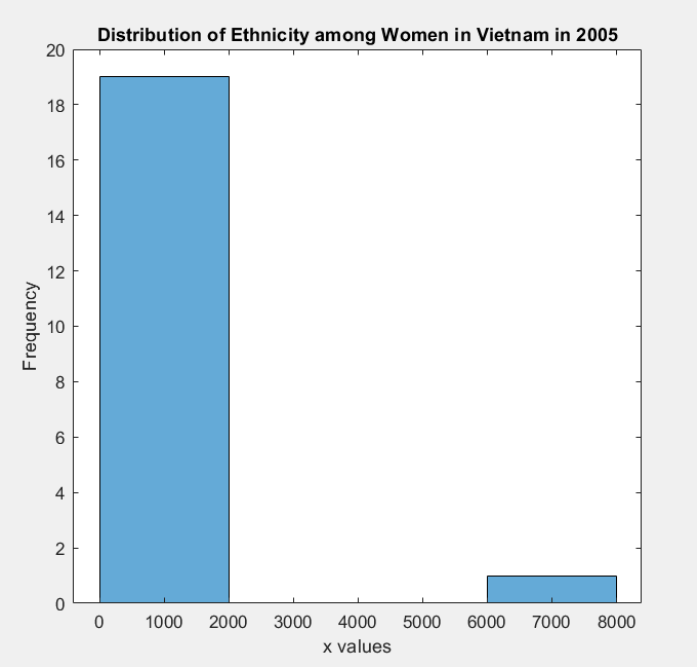
**Solution to Q.1**

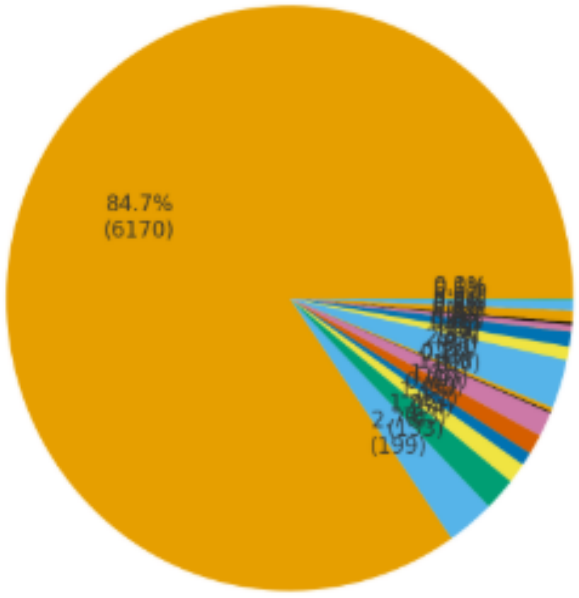
**Data Description:** The dataset contains the distribution of 20 ethnic groups among women in Vietnam as per the Vietnam AIS report, 2005.
**Type of Data:** Discrete
**Size of data:** 20
**Visual representation:** Histogram, Stem-and-Leaf Plot.
**Formulae Used:** <<WRITE THE FORMULAE YOU'VE USED:
{$Mean,\ Median,\ Mode,\ Variance,\ Skewness,\ Kurtosis$ >>

**Code:**<PASTE YOUR CODE HERE.>

**Results:** Mean: 361.95
Median: 45
Mode: 1
Variance: 1872597.8395
Coefficient of Skewness: 16.9412
Coefficient of Kurtosis: 17.9817

# Histogram Plot and Pie Chart of the data:



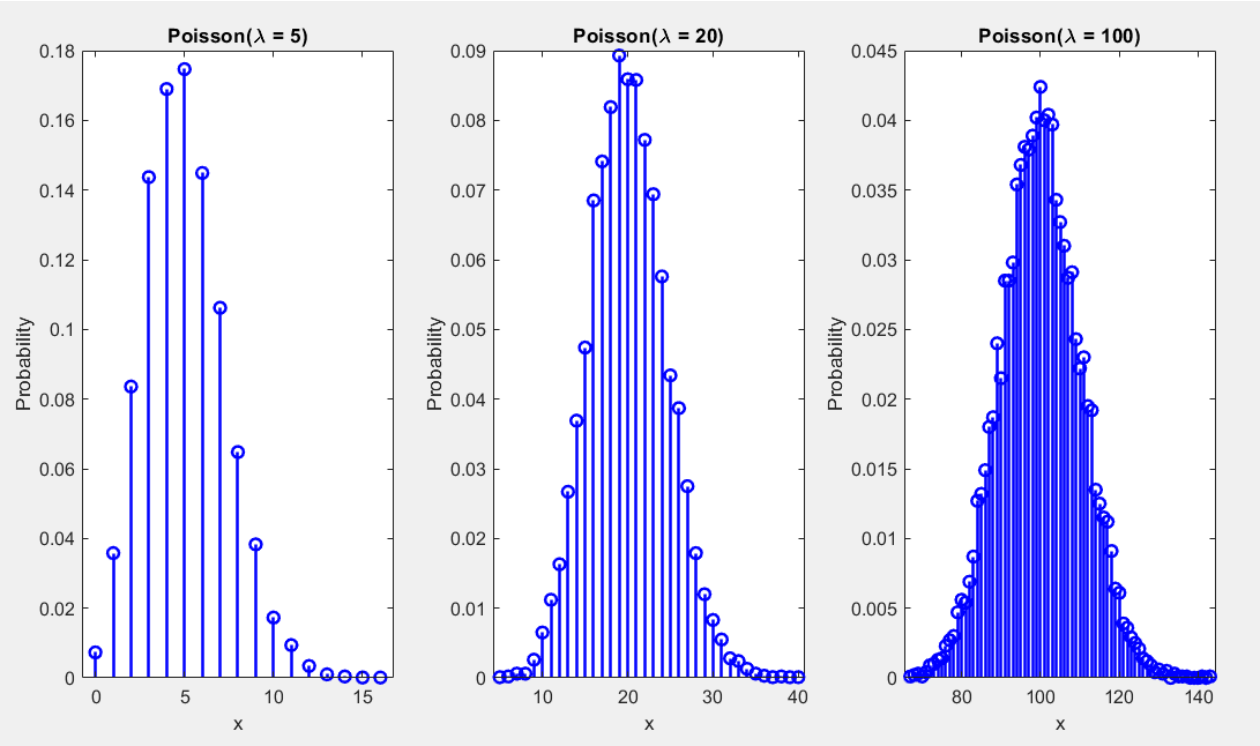
**Observations:** From the statistics calculated, we see that Mean>Median>Mode. This implies that the data is heavily positively skewed. The high skewness in the data can also be observed by the high absolute value of the coefficient of skewness. The variance has an abnormally high value, thereby indicating that the data is heavily scattered, i.e. there is a vast difference in the counts of the ethnic groups among Vietnamese women in 2005. This scatteredness in the data is also visible from the histogram plot. We see that 19 values lie in the interval 0-2000, while only one value lies in 6000-8000.
The high kurtosis value indicates a leptokurtic distribution, with values highly centered around the mean of the data.

**Solution to Q.2**

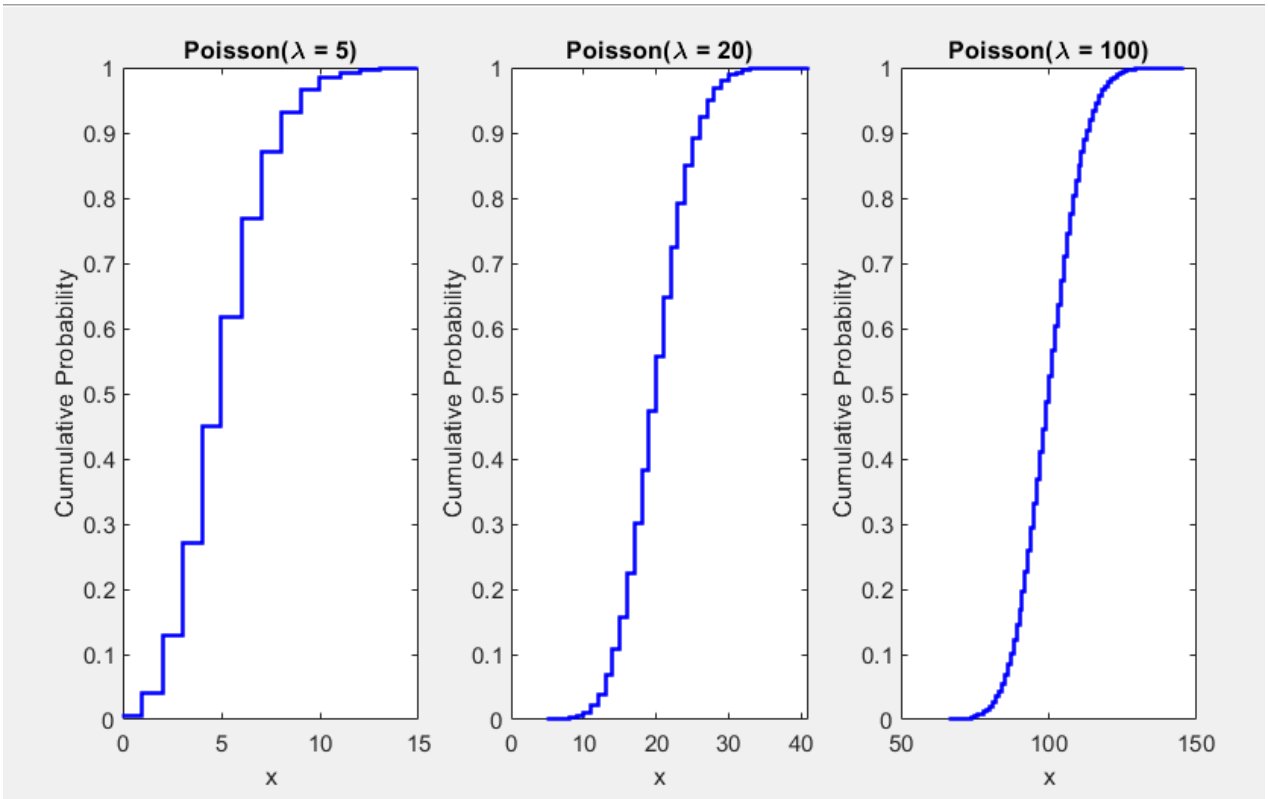
Last digit of Enrolment Number: 2

Distribution: Poisson

a) PMF and CDF plots of the Poisson Parameters: (5, 20, 100) for sample size 10000.
Code:<PASTE YOUR CODE HERE.>




**PMF Plots for different values of λ.**




**CDF Plots for different values of λ.**

**b) Code:**<PASTE YOUR CODE HERE.>

--- Poisson(lambda = 5) ---
Sample size = 10000
Raw 1st moment: sample = 4.9759, theoretical = 5.0000
Raw 2nd moment: sample = 29.8577, theoretical = 30.0000
Raw 3rd moment: sample = 205.1053, theoretical = 205.0000
Raw 4th moment: sample = 1572.0125, theoretical = 1555.0000

--- Poisson(lambda = 20) ---
Sample size = 10000
Raw 1st moment: sample = 19.9455, theoretical = 20.0000
Raw 2nd moment: sample = 417.4449, theoretical = 420.0000
Raw 3rd moment: sample = 9124.1463, theoretical = 9220.0000
Raw 4th moment: sample = 207461.8521, theoretical = 210820.0000

--- Poisson(lambda = 100) ---
Sample size = 10000
Raw 1st moment: sample = 100.0698, theoretical = 100.0000
Raw 2nd moment: sample = 10113.5132, theoretical = 10100.0000
Raw 3rd moment: sample = 1032101.0784, theoretical = 1030100.0000
Raw 4th moment: sample = 106338227.8400, theoretical = 106070100.0000

c) **Normal approximation to Poisson:** If the Poisson parameter $\lambda \geq 10$, then the normal approximation to Poisson can be used.

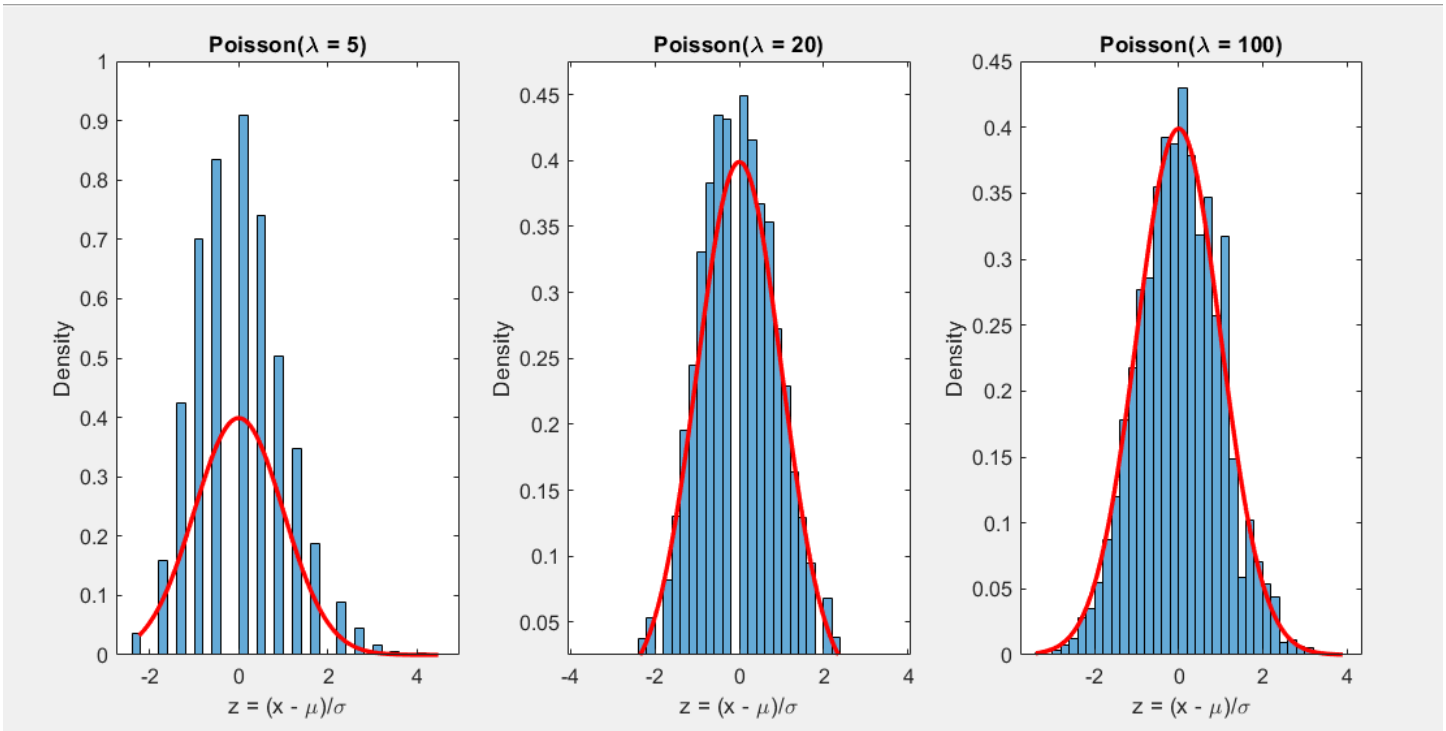
**Code:**<PASTE YOUR CODE HERE.>

**Results:**
Poisson Parameters: 5, 20, 100.
Sample size: 10000.

--- Poisson(lambda=5) ---
Sample size: 10000
Mean(Z): 0.0099  SD(Z): 1.0002

--- Poisson(lambda=20) ---
Sample size: 10000
Mean(Z): 0.0011  SD(Z): 1.0060

--- Poisson(lambda=100) ---
Sample size: 10000
Mean(Z): -0.0139  SD(Z): 0.9872




From the graph shown above, we see that the Normal approximation to the Poisson distribution becomes better with an increasing value of λ.