# Statistics

## Question 1

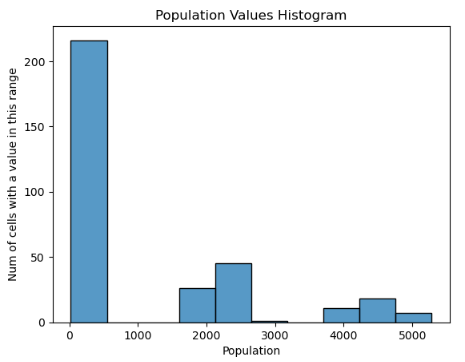
The dataset can be broken into two categories of data: The first category is categorical data which represent descriptive text used to give meaning to what is being viewed within the data set. The second category of data is numerical data. The numerical data itself can be of two types. The first type of numerical data is discrete data. Here the data can usually be counted in a finite matter while the second type of numerical data is continuous data. Here the data is infinite and impossible to count.

Using descriptive statistics, a data analyst has many techniques at the disposal which they may use to describe the data either numerically or graphically. Descriptive statistics usually involves using samples of dataset so the data itself can be transformed into information by organising the data. Having the data organised allows the data analysts to pull variables from the data such as averages that they may be used in functions provided by the various libraries available in python. Having the data organised also allows for the data to be used as an input for inferential statistics.

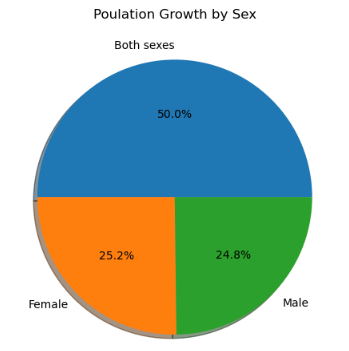
The data analyst needs to choose the right sample of data to obtain accurate information from the data set. Types of samples that may be used include a random sample, stratified random sample and cluster sample. An example of a random sample would be a weekly lottery draw. The result for that particular week would be random where as all the balls drawn out over the year would be a more complete sample. The stratified random sample is similar to the previous example except the data has been broken into groups before the random sample is taken. The cluster sample also breaks the data into groups but the difference here from the stratified random sample is there is individual samples are not taken instead groups of data are selected so all characteristic attributes of that data can be examined within the cluster.

Using these techniques, it was possible to create graphs and charts to give information about the data at a glance. First a histogram was created which shows the frequency of numerical data using rectangles. There were 324 rows of data within the spreadsheet with one column of numerical values. The histogram’s horizontal axis groups the numeric values column while its vertical axis is used to display the frequency of those values within the dataset. The histogram displayed below shows of the 324 rows of data over two hundred of those rows had a value less than one thousand (one million).

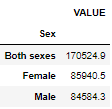
Note: For the dataset to be more readable, the dataset from the CSO displays its numeric values in units of a thousand. The values displayed need to be multiplied by one thousand to give their actual value. For example, 1000 population \* 1000 per unit = 1,000,000



The pie chart shows the totals for all the data within the dataset grouped by category. As the data for both sexes are combination of the male and female values it should equate to fifty percent while the values for the total amount of female and male data may differ form one another. Here we can see that there are more females in Ireland then there are males. From perform manual calculation this is the expected result. Also, from using Python to group the numeric values for each category of gender we are expecting to see the same result displayed within the pie chart.

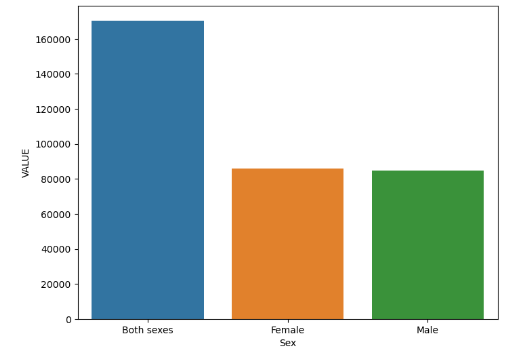


The bar chart’s vertical axis shows the total population values over eighteen years. The data within the CSO spreadsheet gives a total for each year and it also give the total numbers for each region. |With this in mind the figures could be halved as we collect the same data twice. Once as a total for all countries and a second time by summing the data from each region as a total. The horizontal axis shows the categories of data displayed within the chart. As the numbers for total females and males slightly differ it may be hard to see within the chart itself but there is a difference which can be seen in the below table followed by the chart itself.



Note: the values in the above table are double what the actual figure would be over the eighteen years. This is due to the data being collected twice. Once as a total for all counties and secondly as total from the following regions:

* EU14 excluding Ireland (countries in the EU pre 2004 excluding UK & Ireland)
* EU15 to EU27 (accession countries joined post 2004)
* United Kingdom
* Ireland
* All countries excluding Ireland, United Kingdom and EU272020



## Question 2

The Binomial Distribution is a discrete probability distribution. Distributions like Binomial tend to involve statistical analysis of counts to determine the probability of an event occurring. For this assignment Binomial distribution was used to find the probability of a given number of people being a certain gender within ten people. To get valid results for these probabilities three functions were used. These functions were made available after the binom library had been imported from scipy.stats.

The probability value used by the binomial functions was calculated by counting the total population and then dividing that result by the count for the total female population giving a number which represents the probability of the next person you meet in Ireland being a woman.

First the probability mass function was used. To use this function properly three parameters, have to be passed. The first parameter ‘k’ represents the amount of an occurrence we are looking for. The second parameter ‘n’ represents the total number of occurrences while the third parameter represents the probability of an event happening per iteration. A simple calculation to answer what the probability of three out of ten random people being female was.

The next function used was the cumulative distribution function. Using three parameters the probability of observing less than or equal to x can be obtained, ‘n’ represents the successes in trials, with the probability of success on a single trial being represented as ‘p’. To calculate the probability of having more than ‘k’ out of ‘n’ iterations the survival function was used which can be viewed as 1 – cdf.

The Poisson distribution also uses the same functions except here there are only two parameters passed. The first parameter ‘k’ represents the value we are looking to find the probability of by using the second parameter ‘mu’ which an average value pre calculated. To use Poisson first the average growth of the total population over eighteen years was found. This average was set as ‘mu’ and then ‘k’ was given a value of which we want to find the probability of. For example, ‘k’ was set to 100 which represents 100K population ‘mu’ the average growth value was ran against ‘k’ to give the probability to the total population within Ireland growing over one hundred thousand in a year.

## Question 3

Normal distribution also know as Gaussian distribution is a continuous distribution. Graphs created by using normal distribution tend to look like a bell and are commonly called “the Gauss Bell Curve”. The distribution takes into consideration an infinite number of data points where the expected value and variance are finite numbers. The distribution is symmetric so the curve displayed will never be to one side instead the expected average will be in the middle giving us the bell curve previously discussed. By importing the norm library from scipy.stats we gain access to the probability mass function, the cumulative distribution function and the survival function which were previously used for the Binomial and Poisson distributions.

## Question 4

Explain the importance of the distributions used in point 3 and 4 in your analysis. Justify the choice of the variables and explain if the variables used for the discrete distributions could be used as normal distribution in this case.

# Data Prep

## Question 1

The boxplot was created by importing the Seaborn library. The boxplot itself is very basic and is used to visualize distributions. When comparing data between different groups the boxplot is quite useful as it does not overload the viewer and gives accurate information about the different groups within the dataset. The lines extending from the rectangles are called whiskers so, sometimes a boxplot may be called a box-and-whisker plot. The box/rectangle in the graph shows the quartiles of the dataset while the whiskers show the rest of the distribution. In the example below the data was split into each three groups with two of the groups representing genders and the third group represents both genders.

The vertical axis displays the population numbers and the horizontal axis displays the name for each group. The lower whisker for each box shows the minimum value while the upper whisker shows the maximum value. Then the box itself is split between the first quartile and the third quartile with the median being displayed as a line separating the two quartiles.

