**PROJECT 1: DESCRIPTIVE STATISTICS**

*Submitted by:* GORREPATI SAI SURYA (UTA ID 1002028664*)*

TABLE OF CONTENTS:

|  |  |  |
| --- | --- | --- |
| Section Number | Section Name | Page Number |
| I | Cover Page | 1 |
| II | Data collection | 2 |
| III | Descriptive analysis | 3-7 |
| IV | Appendices of original data sets | 8 |

**COVER PAGE:**

NAME*:* GORREPATI SAI SURYA

UTA ID: 1002028664

EMAIL: sxg8664@mavs.uta.edu

I Gorrepati Sai Surya  , did not give any or receive any assistance on this project and the report submitted is wholly my own.

I was really fond of iris data because it is a simple dataset and numerous observations can be made. Thus, I have analysed this data for my project. A link is provided for reference dataset 1 (<https://www.kaggle.com/code/jchen2186/machine-learning-with-iris-dataset>)

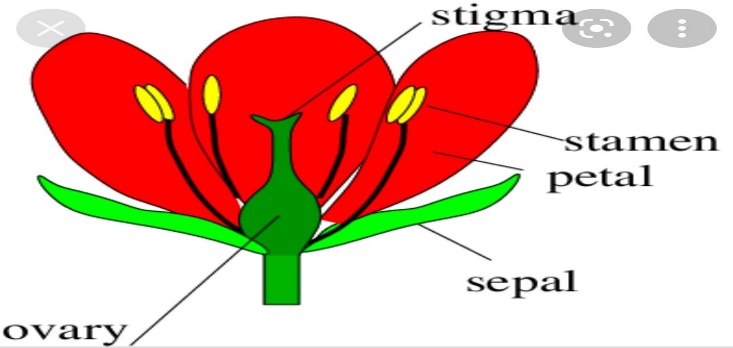
I have used the USA accident dataset as my second data in this project. A link is provided for this below. (https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents)

**DATA COLLECTION:**

DATA SET 1:

I have used iris data as my data set 1 for this project and this data is about iris flowers. There are three disparate kinds of flowers in the data. Iris data has four different attributes: petal width, petal length, sepal width, and sepal length. The targeted population includes all 150 elements from a set of data. A sample of 130 observations is drawn from the population for this project.

Sepal means the outer part of the flower as shown in the below picture, sepal width indicates the width of the outer part of the flower. We can consider the values of sepal width as a continuous random variable because it takes values within the specified range and for each observation of sepal width. Furthermore, when the histogram for sepal width is visualized it is suspected to be normally distributed.



I have used google collab (A online Python compiler) for analyzing the sepal width from the iris data set. When choosing a sample data from the population I have used random\_state= 0 in python because it enables us to maintain the same sample observations instead of having variations in the sample observations.

DATASET2:

I have used accident data set of the USA as my second data set. There are 2.8 million data in this set which covers all the states in USA. The observation was done on start time where accidents take place on a particular day. 102 observations are taken from the data and this data is from 8th February 2016 to 11th February 2016 in Texas.

The 1st observation is start time of first accident that happened on 8th February 2016 and 2nd observation is the start time of a second accident that happened on this day. The time interval was calculated between these times. The events are defined as the time interval of accidents that occurred from 8th February to 11th February(1st accident, 2nd accident, and so on). Descriptive analysis was done in excel for this dataset and Time interval observations are converted into seconds for analysis.

**DESCRIPTIVE ANALYSIS:**

DATA SET1:

Number of observations are 130, n=130

Sample mean **=**

Sample Median

Sample Standard deviation **=**

Sample Range **=- =** 4.4- 2.0 =2.4

Q1 Lower Quartile = 25th percentile =2.80

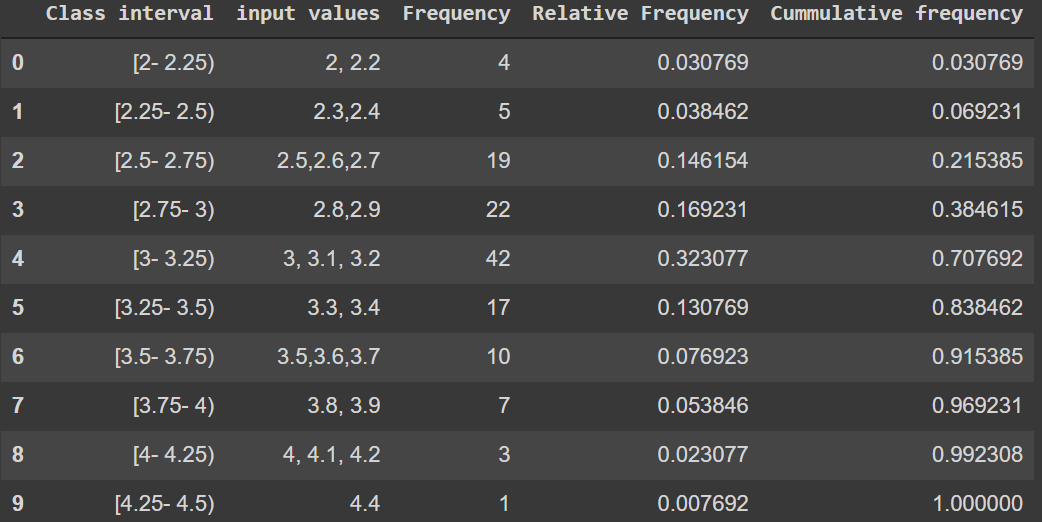
Q2 Middle Quartile = 50th percentile =3

Q3 Upper Quartile = 75th percentile = 3.30

Q2 is same as sample median because Q2 is 50th percentile of observations

|  |  |
| --- | --- |
|  | Values |
| Sample mean | 3.057 |
| Sample median | 3.0 |
| Sample standard deviation | 0.4405 |
| Sample range | 2.40 |
| Quartile Q1 | 2.80 |
| Quartile Q2 | 3.00 |
| Quartile Q3 | 3.30 |

**TABULAR SUMMARY:**

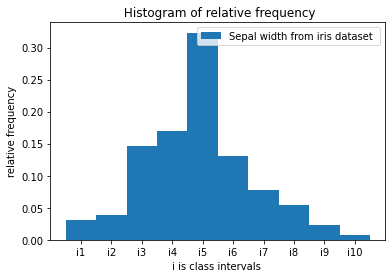


Let I1 be the class interval from 2 to 2.25 where 2.25 is excluded in this interval, I have sepal width of 2 and 2.2 which belongs to this class. Both values are repeated four times i.e it is the frequency of interval.

Similarly, I2 is a class interval from 2.25 to 2.5 where 2.5 is excluded from the class. The sepal width of the iris flower has 2.3 and 2.4 which belong to this class and their frequency is 5.

Let I3, I4, I5, I6, I7, I8, I9, I10 be class intervals of [2.5, 2.75) [2.75,3) [3, 3.25) and so on.

For each class Relative Frequency is computed by using the below formula

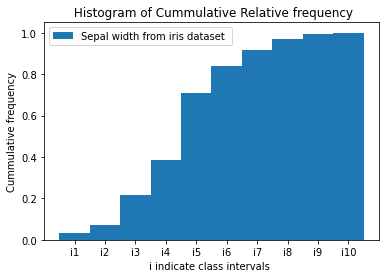


Observations made from the above histogram:

1. The frequency is highest in the class interval where mean lies because in interval i5 [3, 3.25) as the number of observations was more in this range when compared to other intervals.
2. The frequency of sepal width is minimum at the outliers.
3. When we move from class interval i1 to i5 (closer to sample mean) the relative frequency has constantly incremented. In opposition to this for class interval i6 to i10 (after the sample mean) a decline in relative frequency took place.
4. The histogram is a right-skewed graph as most observations from a taken sample are in the intervals i1 to i5. We can prove the graph is a right-skewed as (median < mean.)
5. The value of skewness is 0.352. Therefore, is the taken sample data is not a perfect normal distribution however it is very close to be perfect normal distribution. Skewness value is obtained using stats.skew command in google collab.

For each class, Cumulative relative frequency (CRF) is computed by using the below formula

For instance, CRF of i3 = , Similarly CRF are calculated for all intervals.



Observations made from the above histogram:

1. 38.46 % of the sample data of is till the interval [2.75,3) which implies 38.46 % of data is from intervals [2,2.25) [2.25, 2.5) [2.5, 2.75) [2.75,3). A similar interpretation can be made for the remaining CRF.
2. The value of CRF is highest in interval [4.25,4.5).

DATASET 2:

The number of observations n=102

Sample mean **=**2944.068627

Sample Median 1122.5

Sample Standard deviation **=** 5650.288385

Sample Range **=- =** 30270

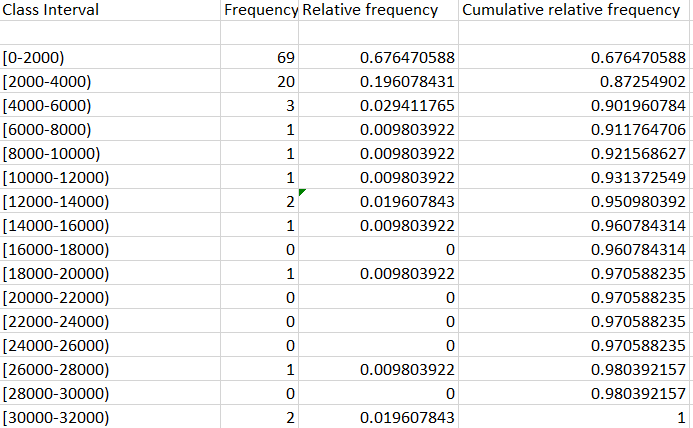
Q1 Lower Quartile = 25th percentile =173

Q2 Middle Quartile = 50th percentile =1122.5

Q3 Upper Quartile = 75th percentile = 2879.5

Q2 is same as sample median because Q2 is 50th percentile of observations

**TABULAR SUMMARY:**



I have taken the range of class interval to be 2000 and formed a table as show above. For instance, consider interval [0-2000). The frequency of this interval is 69 which tells us 69 accidents took place when time interval is between 0 to 2000.

Relative frequency of each interval is

Observations:

1. Most accidents occur when time interval is in the range of [0, 2000) and zero accidents have took place when time interval is from range of [16000, 180000), [20000-22000), [22000-24000), [24000-26000).
2. When time interval value is more we can understand the time gap between consecutive accidents is more. Similarly, when time interval value is less we can tell that start time between consecutive accidents is less.

Observations:

1. 67.6 % of the sample data of is upto the interval [2.75,3) which implies 67.6 % of data is from intervals [0-2000. A similar interpretation can be made for remaining CRF.
2. The value of CRF is highest in interval [30000-32000).

**APPENDICES**:

SAMPLE DATSET1 (SEPAL WIDTH)



SAMPLE DATSET2 ( TIME INTERVALS IN SECONDS)

