PROJECT – REALSTAT

A Chi-Square Driven Decision Analytics

Table of contents

Data Collection - Dataset 1

In order to build a solid basis for this study, a systematic data collection process was undertaken. This section describes how the data was collected for Data Set 1 (Salaries of employees at a company) and Data Set 2 (Time intervals of people exiting the Chase building at 500 East Border.)

Data collection of Data Set 1 :

In order to compile an accurate representation of employee salaries within the organization, a systematic and comprehensive data collection process was meticulously executed**.** Data Set 1, which pertains to employee salaries in a corporate context, was obtained from an online repository accessible at the following URL: **https://www.kaggle.com/datasets/rkiattisak/salaly-prediction-for-beginer.** This dataset encompassed comprehensive information about the salaries of 375 employees at a company. The dataset was structured with several columns, that included demographic information such as age, education level, years of experience, job title, gender, and salary. Each row within the dataset represents a distinct employee record. However, for the purpose of this descriptive analysis, a judicious selection was made, focusing on solely two pertinent columns: "**Job Title**" and "**Salary**." This subset of data consisting of 140 randomly selected entries was deemed most suited to the objectives of this analysis.

The approach of the Data Set 1 collection:

1. The dataset was obtained from an online repository **https://www.kaggle.com/datasets/rkiattisak/salaly-prediction-for-beginer.** It encompassed comprehensive information about the salaries of 375 employees at a company and other demographic information.

2. The dataset was further cleaned by removing any missing values which happened to be none in this dataset. Selected specific columns: 'Job Title' and 'Salary' and filtered 140 entries from the dataset which were randomly selected and created a subset for further descriptive analysis.

   The main idea for doing a descriptive study on the above dataset is to acquire a comprehensive understanding of the compensation structure within the organisation. This analysis intends to give us insights into salary distribution and factors influencing salary levels such as job titles.

Descriptive Analysis - Dataset 1

Descriptive analysis is a fundamental component of data analysis that aims to summarize and present data in a meaningful and easily comprehensible manner. This analytical approach entails organizing, summarizing, and presenting data using different statistical and graphical tools. Its primary objective is to provide a clear and concise overview of the essential characteristics of a dataset, such as central tendencies (mean, median, mode), variability (range, variance, standard deviation), and the distribution of values (histograms, frequency tables). It allows researchers and decision-makers to acquire useful insights into the underlying patterns and trends in data, allowing for a more in-depth knowledge of the issue under examination.

Data Set 1: Salaries of employees at a company

Measurements of central tendency:

Measurement	Value
Sample Mean	**100392.8571**
Sample Median	**95000**
Sample Mode	**95000**

Table 1.1 – Measurement of central tendency

Sample Mean: The sample mean, is a fundamental measure used to describe the central tendency of a dataset. It is calculated by summing up all the values in a sample and then dividing that sum by the total number of observations in the sample. Mathematically, it can be expressed as: $\bar{x} = (\Sigma x) / n$

Where:
The sample mean is represented by $\bar{x}$.
The total sum of all values in the dataset is denoted by Σx.
n is the total number of observations in the dataset i.e. 140

In Excel, we use the = AVERAGE(B2:B141) formula to calculate the sample mean. Here the sample mean salary is $100,392.85, which suggests that, on average, employees in this organisation earn around $100,392.85.

SAMPLE MEAN =	100392.8571
Formula used =AVERAGE(B2:B141)	

Sample Median: When a dataset is arranged in ascending or descending order, the sample median is a measure that reflects the middle value. It is a measure of central tendency that is less sensitive to extreme outliers compared to the mean.
First, we arrange the datasets in ascending order. This step is crucial to identify the middle value. Then we use the Excel function = MEDIAN(B2:B141) to calculate the sample median. Here the sample median is $95000

SAMPLE MEDIAN =	95000
Formula used =MEDIAN(B2:B141)	

Sample Mode: The sample mode is a measure that represents the value or values that occur with the highest frequency in a dataset. Unlike the mean and median, which focus on central tendencies, the mode identifies the most prevalent or frequently occurring value(s) in the dataset.

In Excel, we use the = MODE(B2:B141) formula to calculate the sample mode. The sample mode is $95000 which tells us that this is the most common salary of employees in the company.

SAMPLE MODE =	95000
Formula used =MODE(B2:B141)	

Measurements of Variability :

Measurement	Value
Sample Range	**220000**
Sample Variance	**1946067575**
Sample Standard Variation	**44114.25591**

Table 1.2 – Measurements of Variability

Sample Range: The sample range is a basic measure that quantifies the spread or variability of a dataset. It is calculated as the difference between the maximum and minimum values in the sample.

In Excel, we use the = MAX(B2:B141)-MIN(B2:B141) formula to calculate the sample range. The sample range is $220000, Where the maximum value is $250,000 and the minimum value is $30,000.

SAMPLE RANGE =	220000
Formula used =MAX(B2:B141)-MIN(B2:B141)	

Sample Variance: Sample variance is a measure that quantifies the spread or dispersion of data points in a sample. It provides insight into how individual data points deviate from the sample mean.

In Excel, we use the = VAR.S(B2:B141) formula to calculate the sample variance. The sample variance is 1946067575, which tells us that there is greater variability in salaries among employees.

SAMPLE VARIANCE =	1946067575
Formula used =VAR.S(B2:B141)	

Sample Standard Deviation: The sample standard deviation is a statistical measure that quantifies the amount of variation or dispersion in a dataset. It is particularly useful for understanding how individual data points deviate from the sample mean. The sample standard deviation is calculated as the square root of the sample variance.

In Excel, we use the = STDDEV.S(B2:B141) formula to calculate the sample Standard Deviation The sample Standard Deviation is 44114.25, which tells us that salaries are more spread out from the mean.

SAMPLE STANDARD DEVIATION =	44114.25591
Formula used =STDDEV.S(B2:B141)	

Percentiles:

Measurement	Value
QUARTILE 1 (25th Percentile)	**70000**
QUARTILE 2 (50th Percentile)	**95000**
QUARTILE 3 (75th Percentile)	**120000**

Table 1.3 – Measurement of percentiles

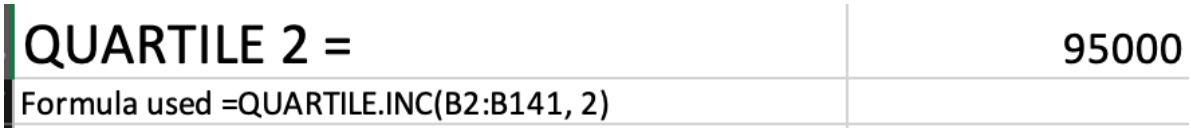
Quartile 1: The first quartile, often denoted as Q1 is a measure that represents the 25th percentile of a dataset when it is ordered in ascending or descending order. In other words, it divides the dataset into 4 equal parts, with Q1 marking the boundary between the lowest 25% of the data.

In Excel, we use = QUARTILE.INC(B2:B141, 1) formula to calculate the Q1, where 1 in the formula represents the 25th percentile. The Q1 value is 70000, which indicates the salary level below which a quarter of the employees earn their wages.

QUARTILE 1 =	70000
Formula used =QUARTILE.INC(B2:B141, 1)	

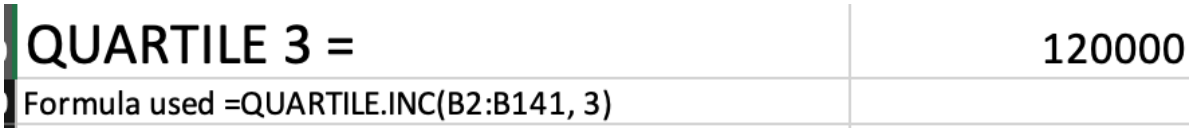
Quartile 2: The second quartile, often denoted as Q2 and commonly referred to as the median is a measure that represents the 50th percentile of a dataset when it is ordered in ascending or descending order. It divides the dataset into two equal halves, with Q2 representing the midpoint or middle value.

In Excel, we use = QUARTILE.INC(B2:B141, 2) formula to calculate the Q2, where 2 in the formula represents the 50^{th} percentile. The Q2 value is 95000, which indicates the salary level at which exactly half of the employees earn less and half earn more.



Quartile 3: The third quartile, often denoted as Q3 is a measure that represents the 75th percentile of a dataset when it is ordered in ascending or descending order. In other words, it divides the dataset into four equal parts, with Q3 marking the boundary between the upper 25% of the data and the lower 75%.

In Excel, we use = QUARTILE.INC(B2:B141, 3) formula to calculate the Q3, where 3 in the formula represents the 75^{th} percentile. The Q3 value is 120000, This implies that it denotes the point at which 75% of employees earn less and 25% earn more.



Box-and-Whisker Plot:

A box and whisker plot, commonly referred to as a box plot, is a graphical depiction of the distribution and important summary statistics of a dataset. It shows the data's central tendency, spread, and any probable outliers. Box and whisker charts are effective for displaying data shape and variability.

<u>To calculate boxplot</u> :

Q1= 70000

Q2= 95000

Q3= 120000

IQR = Q3 – Q1 = 50000

<u>Calculating Lower Whisker</u> :

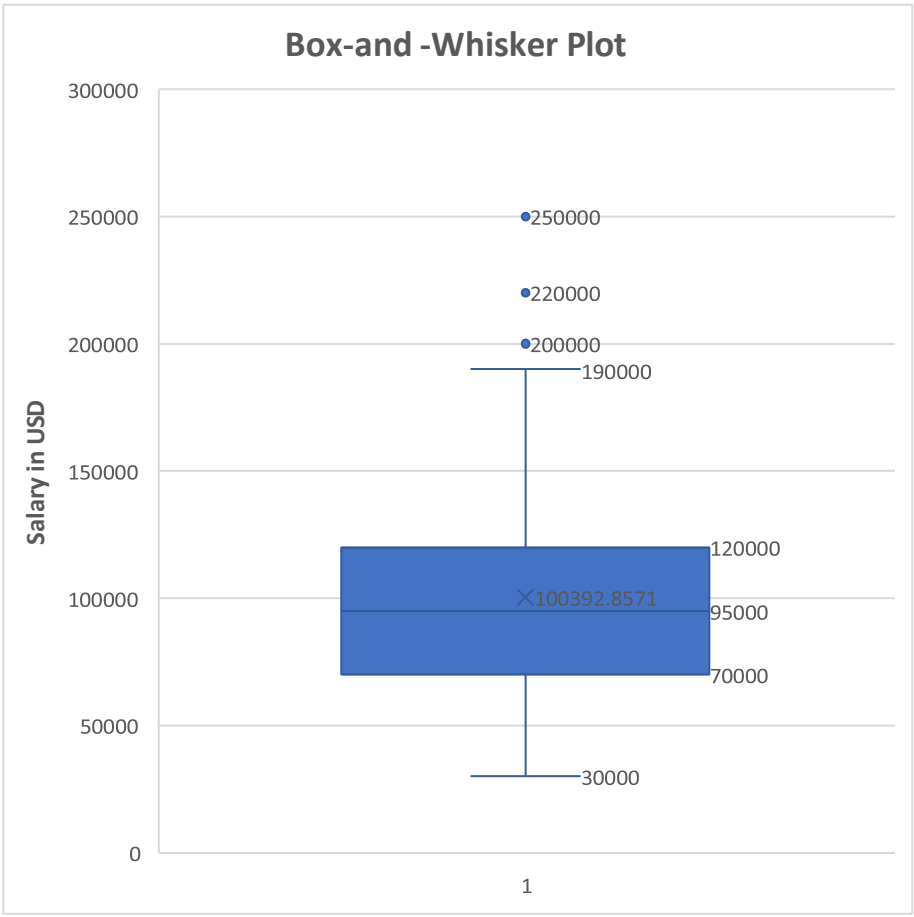
Draw the whisker down to the smallest value > Q1 – 1.5*IQR

The lower whisker for our dataset is 30000.

Calculating Upper Whisker :

Draw the whisker up to the largest value < Q3 + 1.5*IQR

The upper whisker for our dataset is 190000.



1. The interquartile range (IQR) is represented by the box in a box and whisker plot, which encompasses the middle 50% of the data.
2. A vertical line or "whisker" at the median (Q2 or the second quartile) divides it into two portions. (Q2 = 95000)
3. The box's bottom border indicates the first quartile 70000 (Q1 or the 25%ile), while its top edge represents the third quartile 120000 (Q3 or the 75%ile).
4. The whiskers extend from the box's borders to display the data range. The values above and below the upper whisker and lower whisker are called outliers and are commonly shown as isolated points or dots.
5. A line or vertical mark within the box represents the median (Q2= 95000).

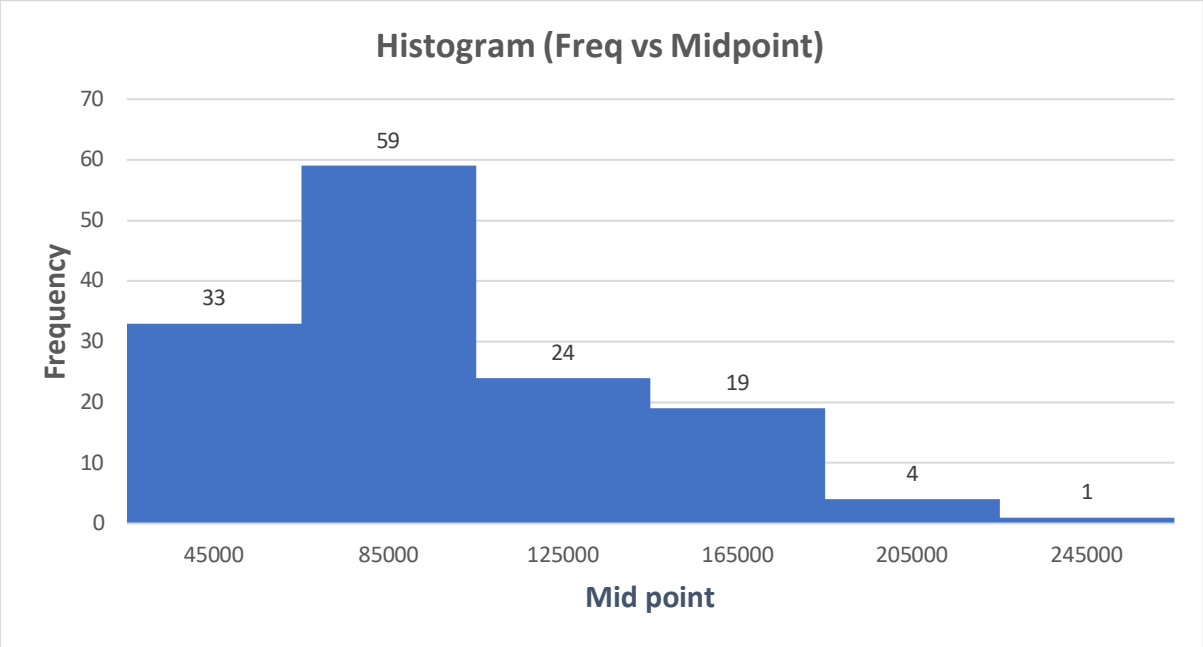
To construct a box plot in Excel, select column B data -> go to Insert -> select the statistical graph and choose Box and Whisker. Your box plot will be constructed. In the above box plot, the outliers are 200000, 220000, and 250000.

Tabular Summary :

Lower bound	Upper bound	Mid-point	Frequency	Relative Frequency	Cumulative Relative Frequency
25000	65000	45000	33	0.2357	0.2357
65000	105000	85000	59	0.4214	0.6571
105000	145000	125000	24	0.1714	0.8286
145000	185000	165000	19	0.1357	0.9643
185000	225000	205000	4	0.0286	0.9929
225000	265000	245000	1	0.0071	1.0000
			140		

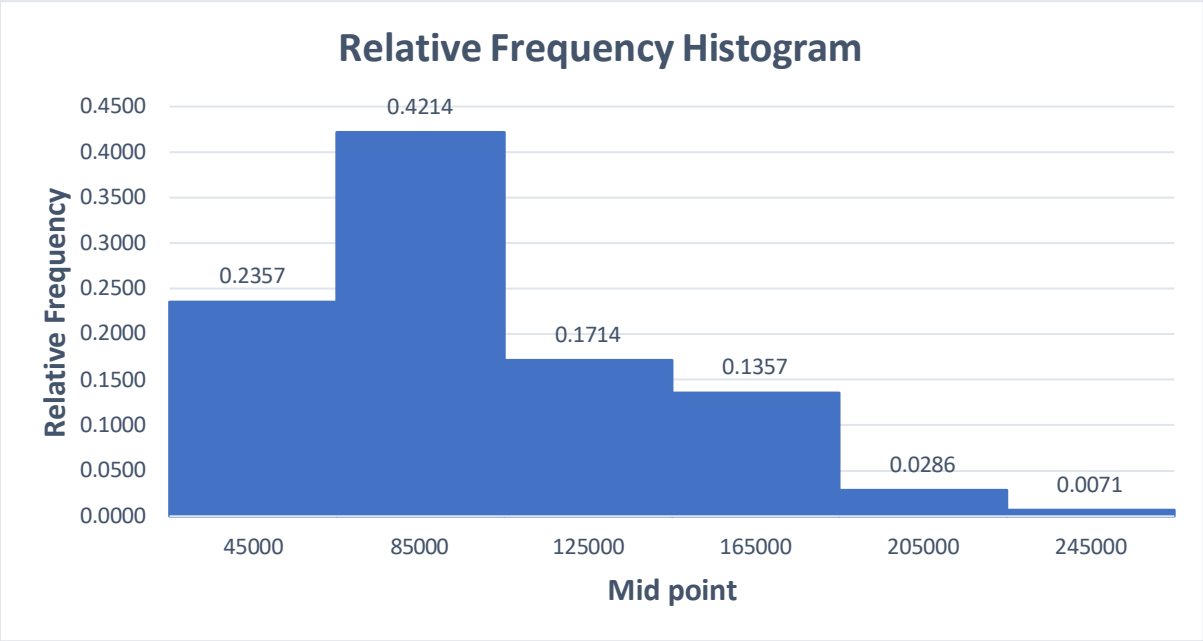
Table 1.4 – Tabular summary

Frequency Histogram:



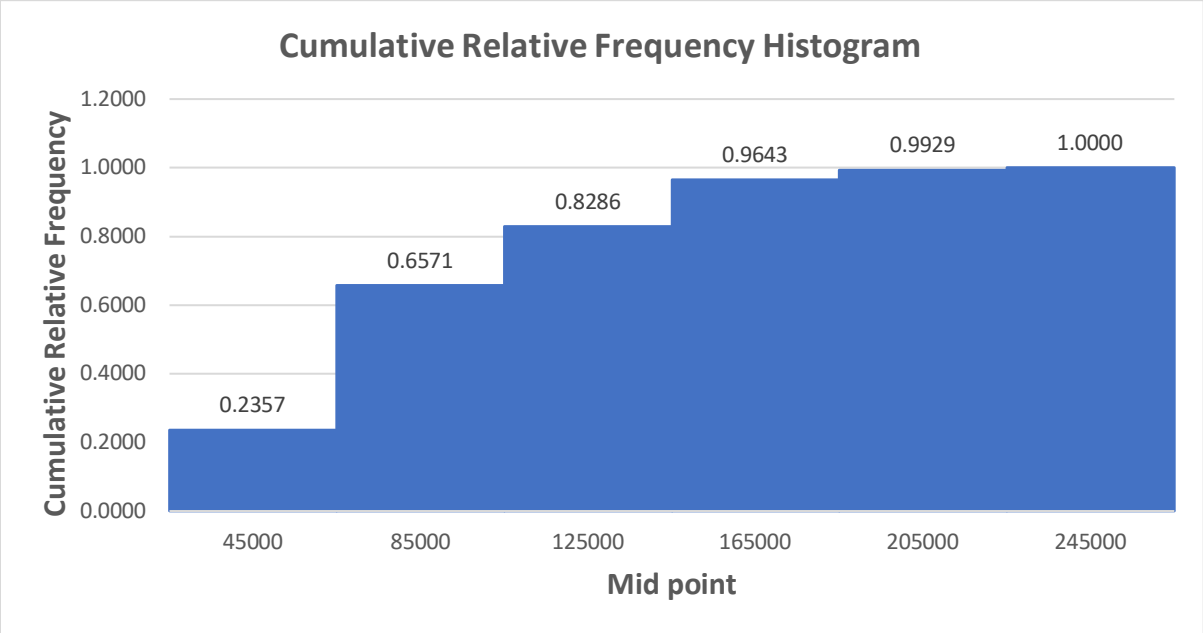
A histogram is a graphical depiction of a dataset's distribution. It is used to display the frequency or count of data points that fall into various intervals within a certain range. Histograms are very useful for determining the shape, central tendency, and distribution of data. In the above graph X-axis represents the midpoint of Salaries and the Y–axis represents Frequency. The number of people whose salary lies in the range of $65000 – $105000 is 59 which is the highest, followed by $25000 - $65000 with 33 people. Only 1 person receives pay in the range of $225000 – $265000. The above graph doesn't follow a Normal distribution curve.

Relative Frequency Histogram:



A relative frequency histogram is a type of histogram, that illustrates the relative frequencies or proportions of values within specified intervals. From the above relative frequency histogram and tabular summary, we can interpret that most of the salaries (around 83%) are between $25000 – $145000 and are slightly skewed towards the right thus making the above graph doesn't completely follow a normal distribution curve. Maximum number of employees salary lies in the interval of $65000-$105000. The number of employees who receive the highest salary is 5 comprising about 3.57%. The above graph doesn't follow a normal distribution curve.

Cumulative Relative Frequency Histogram:



Cumulative relative frequency, also known as cumulative distribution, is a statistical concept that describes the accumulation of data values in a dataset up to a specific point. It reveals how data points are dispersed in proportion to their total percentages. The cumulative relative frequency was calculated by adding the relative frequency values. The total value at the end sums up to 1. We can interpret that 65.71% of employees earn less than or equal to $100,000 and 82.86% earn less than $150,000. These are a few insights one could draw using histograms.

GOODNESS OF FIT TEST

For doing the Chi-square Goodness of Fit Test we require two important parameters. They are:

- Sample mean
- Sample Standard Deviation

Here, the sample mean salary is $100,392.85, which suggests that, on average, employees in this organisation earn around $100,392.85.

The sample Standard Deviation is 44114.25, which tells us that salaries are more spread out from the mean.

For dataset 1 sampled from a normal distribution, we test the hypothesis by performing a chi-square Goodness of Fit Test with a significance level of 0.05. We consider the population mean equal to the sample mean and the population standard deviation equal to the sample standard deviation.

Let the hypotheses be :

H_0: The dataset follows a normal distribution.

H_1: The dataset does not follow a normal distribution.

We calculate the following for fit test:

- Number of observations based on collected data (O_i)
- Class probability (P_i)
- Expected value (e_i)
- Chi-square (statistic value)

NORMAL DISTRIBUTION						
Class	Frequency (O_i)	Class Probability (P_i)	Expected Value (e_i)	Chi-Square		
X ≤ 65000	33	0.211189789	29.56657053	0.398708329		
65000 ≤ X ≤ 105000	59	0.330398771	46.25582795	3.511209903		
105000 ≤ X ≤ 145000	24	0.302444613	42.34224579	7.945681063		
145000 ≤ X ≤ 185000	19	0.128405572	17.97678008	0.05824063		
185000 ≤ X ≤ 225000	4	0.025194608	3.527245181	0.06336308		
X > 225000	1	0.002366646	0.331330468	1.349465219		
	140	1	140	13.32666822	(Chi-square test statistic value)	

Table 1.5 - Calculation of chi-square values of dataset 1

To calculate the class probability (P_i) we use the following formulas:

1. For the **first class interval** i.e. (X ≤ 65000) we use: =NORMDIST(upper-limit, **μ**, **σ**, cdf)

   Where upper limit is 65000, mean = 100392.8571, standard deviation = 44114.25591 and cdf=1, substituting these values in the formula

   =NORMDIST(65000,100392.8571,44114.25591,1) → 0.211189789

2. To calculate the **middle classes intervals** we use: =NORMDIST(upper-limit, **μ**, **σ**, cdf) - NORMDIST(lower-limit, **μ**, **σ**, cdf)

   Consider the class (185000 ≤ X ≤ 225000)
   Where the upper limit is 225000, the lower limit is 185000, mean = 100392.8571, standard deviation = 44114.25591 and cdf=1, substituting these values in the formula

   =NORMDIST(225000,100392.8571,44114.25591,1)-NORMDIST(185000,100392.8571,44114.25591,1)

3. For the **last class** i.e.(X > 225000) we use: = 1- NORMDIST(upper-limit, **μ**, **σ**, cdf)

   Where upper limit is 225000, mean= 100392.8571, standard deviation = 44114.25591 and cdf=1, substituting these values in the formula

   = 1-NORMDIST(225000,100392.8571,44114.25591,1) → 0.002366646

To calculate the expected value (e_i) we use the following formula: $n(P_i)$

To calculate the chi-square value we use the following formula: $\chi 2 = \sum(O_i - e_i)^2/e_i$

Combining the class Intervals:

From the above table, we can observe the **ei** values of classes (185000 ≤ X ≤ 225000) and (X > 225000) are less than 5, so we combine these classes accordingly which is shown in the below table.

Class	Frequency (O_i)	Class Probability (P_i)	Expected Value (e_i)	Chi-Square	
X ≤ 65000	33	0.211189789	29.56657053	0.398708329	
65000 ≤ X ≤ 105000	59	0.330398771	46.25582795	3.511209903	
105000 ≤ X ≤ 145000	24	0.302444613	42.34224579	7.945681063	
145000 ≤ X ≤ 185000	19	0.128405572	17.97678008	0.05824063	
X >185000	5	0.027561255	3.858575649	0.337650384	
	140	1	140	12.25149031	(Chi-square test statistic value)

Here the classes (185000 ≤ X ≤ 225000) and (X > 225000) are combined, but we can observe the **e_i** value still remains less than 5. So we further combine these two classes with (145000 ≤ X ≤ 185000). The below table shows us the final combined class intervals.

Class	Frequency (O_i)	Class Probability (P_i)	Expected Value (e_i)	Chi-Square	
X ≤ 65000	33	0.211189789	29.56657053	0.398708329	
65000 < X ≤ 105000	59	0.330398771	46.25582795	3.511209903	
105000 < X ≤ 145000	24	0.302444613	42.34224579	7.945681063	
X >145000	24	0.155966827	21.83535573	0.21459164	
	140	1	140	12.07019094	(Chi-square test statistic value)

For the **last class** i.e.(X > 145000) we use: = 1- NORMDIST(upper-limit, **μ**, **σ**, cdf)

Where upper limit is 145000, mean= 100392.8571, standard deviation = 44114.25591 and cdf=1, substituting these values in the formula

= 1-NORMDIST(225000,100392.8571,44114.25591,1) is 0.155966827 and the expected value is 21.83535573 which is greater than 5.

We perform the chi-square calculation for the (X > 145000) interval which is 0.21459164. Performing the summation of chi-square values $\sum(O_i - e_i)^2/e_i$ we get the final value as **12.07019094**

Calculating degree of freedom:

In Chi-square statistics degree of freedom is the number of class intervals – 1. It is denoted by v. where v = (k-1) and k is the number of class intervals. In this case, v = (4-1) = 3.

Calculating the Chi-Square value from the table:

- We use Table A.5 to find the values.
- $\chi 2_{(\alpha,v)}$ where $\alpha = 0.05$ and $v = 3$.
- Therefore $\chi 2_{(0.05,3)}$ is 7.815

Table A.5 (continued) Critical Values of the Chi-Squared Distribution

	α							
v	**0.30**	**0.25**	**0.20**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**
1	1.074	1.323	1.642	2.706	3.841	5.024	5.412	6.635
2	2.408	2.773	3.219	4.605	5.991	7.378	7.824	9.210
3	3.665	4.108	4.642	6.251	7.815	9.348	9.837	11.345
4	4.878	5.385	5.989	7.779	9.488	11.143	11.668	13.277
5	6.064	6.626	7.289	9.236	11.070	12.832	13.388	15.086

Decision Rule and Conclusion:

a decision rule is to

Reject H_0 when $\chi^2 > \chi^2_{\alpha,k-1}$

$\chi 2_{(Statistic)} > \chi 2_{(0.05,3)}$ i.e. 12.07 > 7.815.

Hence we Reject the Null hypothesis. The given dataset does not follow Normal distribution.

We can conclude with 95% confidence that the given distribution of dataset 1 does not follow the Normal distribution as previously mentioned.

Data Collection - Dataset 2

100 time intervals of people exiting the Chase building at 500 East Border.

In order to compile an accurate representation of people exiting the Chase building, a systematic and comprehensive data collection process was meticulously executed**.**
The data for Dataset 2 is primary data which was collected on 19th September 2023 from 15:50 pm to 17:12 pm by directly visiting the building. The Chase building consisted of a Chase Bank, Texas Health Resources Office, and ALL IN Sports & Entertainment. This experimental data was collected by observing and noting down the values to the nearest second. An online tool called 'Clockface' was used to record the time intervals to the nearest second. Primarily the data was recorded on paper form and later uploaded to Excel. Ethical considerations were paramount, with Chase staff providing informed consent.

The approach of the Data Set 2 collection :

1. Approached the Chase staff to get their consent for my dataset collection upholding Ethical considerations.
2. Recorded time interval of people exiting the Chase building on 19th September 2023 from 15:50 pm to 17:12 pm
3. I used an online tool called 'Clockface' to record my accurate measurements.
4. The data collected was then stored in an Excel sheet with the column name '**Clock Timings'.**
5. The cells were further formatted to consider timing format.
6. A new column B was created named '**Interval**' with units seconds, and the difference between two successive time intervals was taken and recorded in the number format in column B.
7. Now dataset 2 with 99 time intervals was ready for the descriptive analysis.

The main idea behind taking this dataset is to gather and analyse data related to the exit times of individuals leaving the building.

Descriptive Analysis – Dataset 2

Data Set 2: 100 time intervals of people exiting the Chase building at 500 East Border.

Measurements of central tendency:

Measurement	Value
Sample Mean	**47.76**
Sample Median	**36**
Sample Mode	**38**

Table 2.1 - Measurements of central tendency

Sample Mean: The sample mean, is a fundamental measure used to describe the central tendency of a dataset. It is calculated by summing up all the values in a sample and then dividing that sum by the total number of observations in the sample. Mathematically, it can be expressed as: $\bar{x} = (\Sigma x) / n$

n is the total number of observations in the dataset i.e. 99

In Excel, we use the = AVERAGE(B2:B100) formula to calculate the sample mean. Here the sample mean time interval is 47.76 seconds, which suggests that, on average, it takes 47.76 seconds for individuals to exit the Chase building during observed intervals.

SAMPLE MEAN =	47.76
Formula used =AVERAGE(B2:B100)	

Sample Median: When a dataset is arranged in ascending or descending order, the sample median is a measure that reflects the middle value. It is a measure of central tendency that is less sensitive to extreme outliers compared to the mean.
First, we arrange the datasets in ascending order. This step is crucial to identify the middle value. Then we use the Excel function = MEDIAN(B2:B100) to calculate the sample median. Here the sample median is 36.

SAMPLE MEDIAN =	36.00
Formula used =MEDIAN(B2:B100)	

Sample Mode: The sample mode is a measure that represents the value or values that occur with the highest frequency in a dataset. Unlike the mean and median, which focus on central tendencies, the mode identifies the most prevalent or frequently occurring value(s) in the dataset.

In Excel, we use the = MODE(B2:B100) formula to calculate the sample mode. The sample mode is 38 which tells us that this is the most common interval of people exiting the building.

SAMPLE MODE =	38.00
Formula used =MODE(B2:B100)	

Measurements of Variability :

Measurement	Value
Sample Range	**382**
Sample Variance	**2878.14**
Sample Standard Variation	**53.65**

Table 2.2 - Measurements of Variability

Sample Range: The sample range is a basic measure that quantifies the spread or variability of a dataset. It is calculated as the difference between the maximum and minimum values in the sample.

In Excel, we use the = MAX(B2:B100)-MIN(B2:B100) formula to calculate the sample range. The sample range is 382, Where the maximum value is 383 and the minimum value is 1.

SAMPLE RANGE =	382.00
Formula used =MAX(B2:B100)-MIN(B2:B100)	

Sample Variance: Sample variance is a measure that quantifies the spread or dispersion of data points in a sample. It provides insight into how individual data points deviate from the sample mean.

In Excel, we use the = VAR.S(B2:B100) formula to calculate the sample variance. The sample variance is 2878.14, which tells us that there is lesser variability among individuals exiting the building.

SAMPLE VARIANCE =	2878.14
Formula used =VAR.S(B2:B100)	

Sample Standard Deviation: The sample standard deviation is a statistical measure that quantifies the amount of variation or dispersion in a dataset. It is particularly useful for understanding how individual data points deviate from the sample mean. The sample standard deviation is calculated as the square root of the sample variance.

In Excel, we use the = STDDEV.S(B2:B100) formula to calculate the sample Standard Deviation The sample Standard Deviation is 53.65, which tells us that exit times are a little clustered towards the mean.

SAMPLE STANDARD DEVIATION =	53.65
Formula used =STDDEV.S(B2:B100)	

Percentiles:

Measurement	Value
QUARTILE 1 (25th Percentile)	**17**
QUARTILE 2 (50th Percentile)	**36**
QUARTILE 3 (75th Percentile)	**62**

Table 2.3 - Measurements of Percentiles

Quartile 1: The first quartile, often denoted as Q1 is a measure that represents the 25th percentile of a dataset when it is ordered in increasing or decreasing order. In other words, it divides the dataset into 4 equal parts, with Q1 marking the boundary between the lowest 25% of the data.

In Excel, we use = QUARTILE.INC(B2:B100, 1) formula to calculate the Q1, where 1 in the formula represents the 25th percentile. The Q1 value is 17, which indicates the exit time level below which a quarter of the recorded exit intervals are found.

QUARTILE 1 =	17.00
Formula used =QUARTILE.INC(B2:B100, 1)	

Quartile 2: The second quartile, often denoted as Q2 and commonly referred to as the median is a measure that represents the 50th percentile of a dataset when it is ordered in ascending or descending order. It divides the dataset into two equal halves, with Q2 representing the midpoint or middle value.

In Excel, we use = QUARTILE.INC(B2:B100, 2) formula to calculate the Q2, where 2 in the formula represents the 50th percentile. The Q2 value is 36, this means that approximately 50% of the exit times are less than or equal to Q2, and the other 50% are greater than or equal to Q2.

QUARTILE 2 =	36.00
Formula used =QUARTILE.INC(B2:B100, 2)	

Quartile 3: The third quartile, often denoted as Q3 is a measure that represents the 75th percentile of a dataset when it is ordered in ascending or descending order. In other words, it divides the dataset into four equal parts, with Q3 marking the boundary between the upper 25% of the data and the lower 75%.

In Excel, we use = QUARTILE.INC(B2:B100, 3) formula to calculate the Q3, where 3 in the formula represents the 75^{th} percentile. The Q3 value is 62, which is the departure time level at which the vast majority (75%) of the reported exit intervals are found.

QUARTILE 3 =	62.00
Formula used =QUARTILE.INC(B2:B100, 3)	

Box-and-Whisker Plot:

A box-and-whisker plot, commonly referred to as a box plot, is a graphical depiction of the distribution and important summary statistics of a dataset. It shows the data's central tendency, spread, and any probable outliers. Box and whisker charts are effective for displaying data shape and variability.

To calculate boxplot :

Q1= 17

Q2= 36

Q3= 62

IQR= Q3-Q1 = 45

Calculating Lower Whisker :

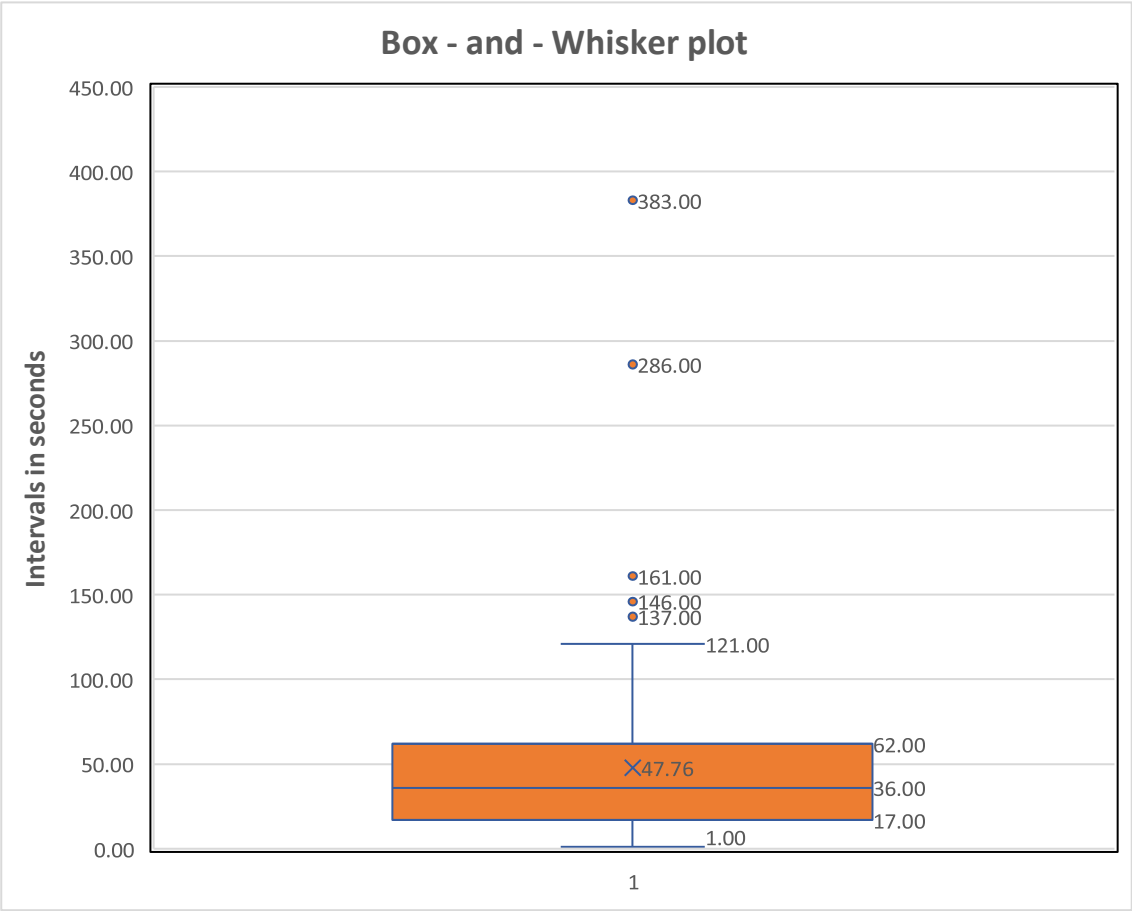
Draw the whisker down to the smallest value > Q1 – (1.5*IQR)

The lower whisker for our dataset is 1.

Calculating Upper Whisker :

Draw the whisker up to the largest value < Q3 + (1.5*IQR)

The upper whisker for our dataset is 121.



1. The interquartile range (IQR) is represented by the box in a box and whisker plot, which encompasses the middle 50% of the data.
2. A vertical line or "whisker" at the median (Q2 or the second quartile) divides it into two portions. (Q2 = 36)
3. The box's bottom border indicates the first quartile 17 (Q1 or the 25%ile), while its top edge represents the third quartile 62 (Q3 or the 75%ile).
4. The whiskers extend from the box's borders to display the data range. The values above and below the upper whisker and lower whisker are called outliers and are commonly shown as isolated points or dots.
5. A line or vertical mark within the box represents the median (Q2= 36).

To construct a box plot in Excel, select column B data (Interval (s)) -> go to Insert -> select the statistical graph and choose Box and Whisker. Your box plot will be constructed.

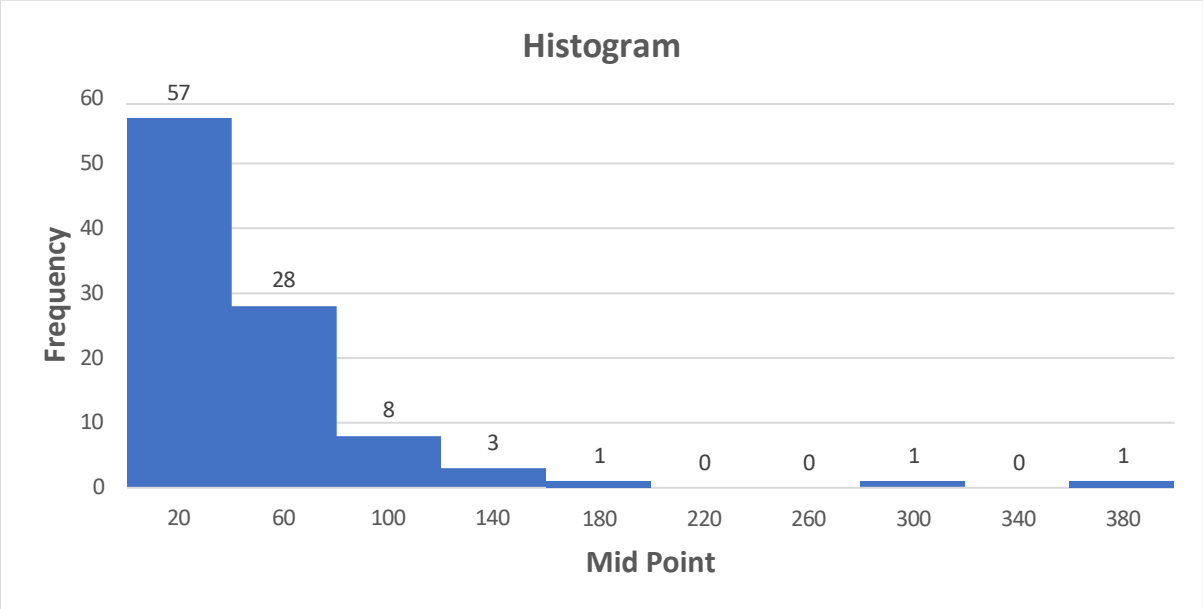
In the above box plot, the outliers are 137.00, 146.00, 161.00, 286.00, 383.00.

Tabular Summary :

LOWER BOUND	UPPER BOUND	MID POINT	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE FREQUENCY
0	40	20	57	0.575757576	0.575757576
40	80	60	28	0.282828283	0.858585859
80	120	100	8	0.080808081	0.939393939
120	160	140	3	0.03030303	0.96969697
160	200	180	1	0.01010101	0.97979798
200	240	220	0	0	0.97979798
240	280	260	0	0	0.97979798
280	320	300	1	0.01010101	0.98989899
320	360	340	0	0	0.98989899
360	400	380	1	0.01010101	1
			99		

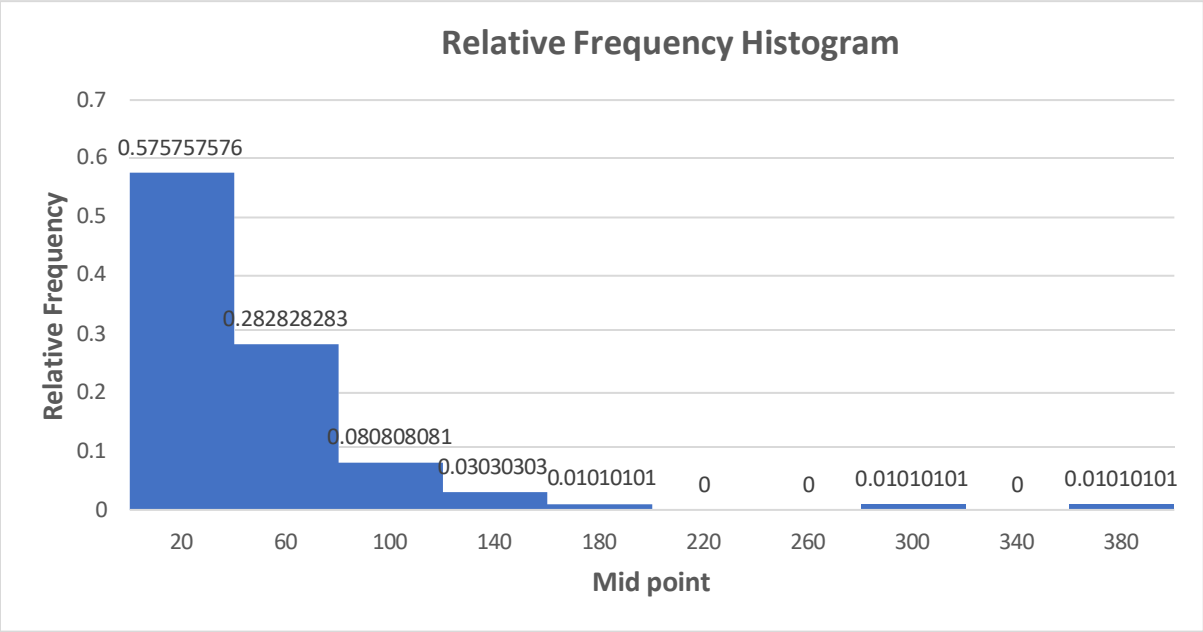
Table 2.4 – Tabular Summary

Frequency Histogram:



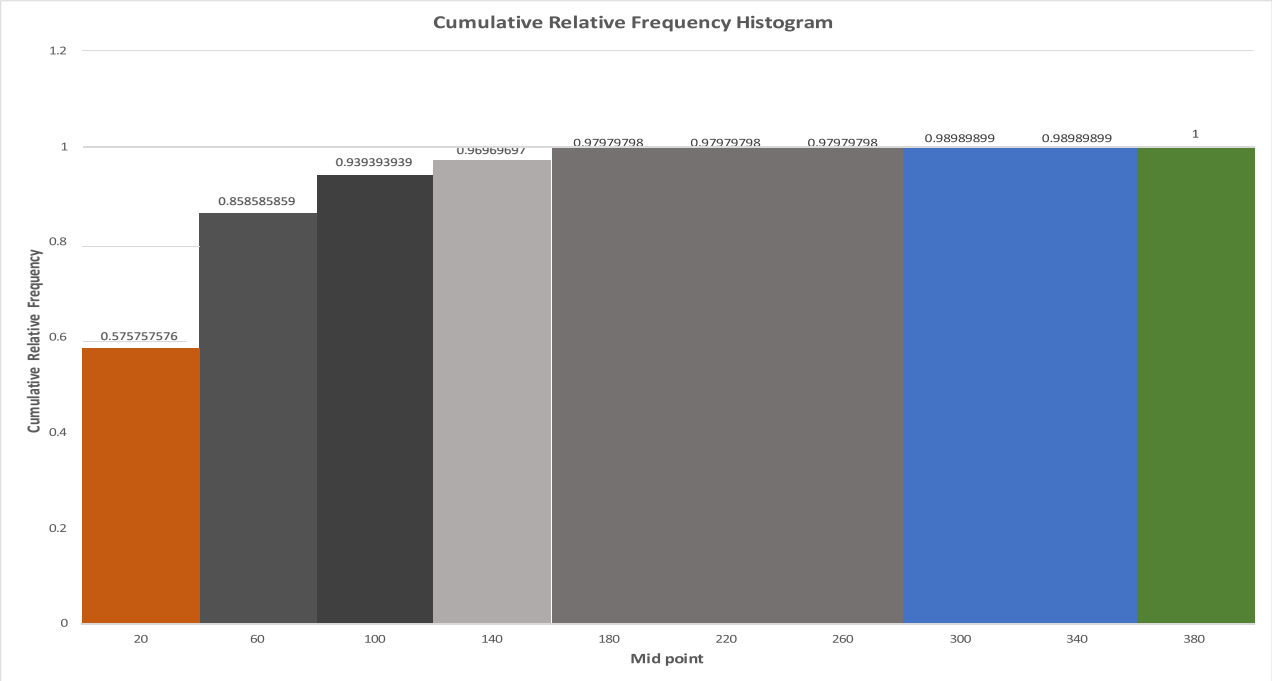
A histogram is a graphical depiction of a dataset's distribution. It is used to display the frequency or count of data points that fall into various intervals within a certain range. Histograms are very useful for determining data's shape, central tendency, and distribution. From the above table, we can infer that people exiting the Chase Bank was highest during the interval 0 - 40 seconds. Followed by 40 – 80 seconds. The graph follows an exponential distribution curve.

Relative Frequency Histogram:



A relative frequency histogram is a type of histogram, that illustrates the relative frequencies or proportions of values within specified intervals. According to the following relative frequency histogram and tabular summary, most of the time intervals of individuals exiting the building (around 85%) are between 0 – 80 seconds. The maximum number of time interval where people exited the building was recorded between 0 – 40 seconds i.e., 57%. The minimum number of individuals exiting was recorded across various time intervals. The graph follows an exponential curve.

Cumulative Relative Frequency Histogram:



Cumulative relative frequency, also known as cumulative distribution, is a statistical concept that describes the accumulation of data values in a dataset up to a specific point. It reveals how data points are dispersed in proportion to their total percentages. The cumulative relative frequency was calculated by adding the relative frequency values. The total value at the end sums up to 1. We can interpret that 97% of individuals exited the building before or equal to 160 seconds. The least number of individuals exited was after 160 seconds. These are a few insights one could draw using histograms.

GOODNESS OF FIT TEST

For doing the Chi-square Goodness of Fit Test we require two important parameters. They are:

- Sample mean
- Sample Standard Deviation

Here the sample mean time interval is 47.76 seconds, which suggests that, on average, it takes 47.76 seconds for individuals to exit the Chase building during observed intervals.

The sample Standard Deviation is 53.65, which tells us that exit times are a little clustered towards the mean.

For Dataset 2 sampled from an Exponential distribution, we test the hypothesis by performing a chi-square Goodness of Fit Test with a significance level of 0.05. We consider the population mean equal to the sample mean and the population standard deviation equal to the sample standard deviation.

Let the hypotheses be :

H_0: The dataset follows an Exponential distribution.

H_1: The dataset does not follow an Exponential distribution.

We calculate the following for the fit test:

- Number of observations based on collected data (O_i)
- Class probability (P_i)
- Expected value (e_i)
- Chi-square (statistic value)

			EXPONENTIAL DISTRIBUTION		
Class	**Frequency (O_i)**	**Class Probability (P_i)**	**Expected Value (e_i)**	**Chi-Square**	
X ≤ 40	57	0.567217912	56.15457326	0.012728195	
40 < X ≤ 80	28	0.245481752	24.30269348	0.562492199	
80 < X ≤ 120	8	0.106240105	10.51777044	0.602710242	
120 < X ≤ 160	3	0.045978815	4.551902653	0.529097836	
160 < X ≤ 200	1	0.019898807	1.969981936	0.477600804	
200 < X ≤ 240	0	0.008611847	0.852572896	0.852572896	
240 < X ≤ 280	0	0.003727053	0.368978278	0.368978278	
280 < X ≤ 320	1	0.001613002	0.15968719	4.421930274	
320 < X ≤ 360	0	0.000698078	0.069109755	0.069109755	
X > 360	1	0.000532627	0.052730112	17.01722597	
	99	**1**	**99**	**24.91444645**	(Chi-square test statistic value)

Table 2.5 - Calculation of chi-square values of dataset 2

To calculate the class probability (P_i) we use the following formulas:

1. For the **first class interval** i.e. (X ≤ 40) we use: =GAMMADIST(upper limit, **α**, **β**, 1)

   Where upper limit is 40, **α** =1, **β**(mean) = 47.76, substituting these values in the formula

   =GAMMADIST(40,1,47.76,1) → 0.567217912

2. To calculate the **middle classes intervals** we use: =GAMMADIST(upper limit, **α**, **β**, 1)-GAMMADIST(lower limit, **α**, **β**, 1)

   Consider the class (80 ≤ X ≤ 120)
   Where the upper limit is 120, the lower limit is 80, **α** =1, **β** (mean) = 47.76, substituting these values in the formula

   =GAMMADIST(120,1,47.76,1)-GAMMADIST(80,1,47.76,1) →0.106240105

3. For the **last class** i.e.(X > 360) we use: = 1- GAMMADIST(upper limit, **α**, **β**, 1)

   Where upper limit is 360, **α** =1, **β**(mean) = 47.76, substituting these values in the formula

   =1-GAMMADIST(360,1,47.76,1) → 0.000532627

To calculate the expected value (e_i) we use the following formula: $n(P_i)$

To calculate the chi-square value we use the following formula: $\chi 2 = \sum(O_i - e_i)^2/e_i$

Combining the classes:

From the above table, we can observe the e_i values of classes after ($80 \le X \le 120$) are less than 5. so we combine all those classes accordingly which is shown in the below table.

Class	Frequency (O_i)	Class Probability (P_i)	Expected Value (e_i)	Chi-Square	
X ≤ 40	57	0.567217912	56.15457326	0.012728195	
40 < X ≤ 80	28	0.245481752	24.30269348	0.562492199	
80 < X ≤ 120	8	0.106240105	10.51777044	0.602710242	
X > 120	6	0.081060231	8.024962821	0.510964912	
	99	**1**	**99**	**1.688895548**	(Chi-square test statistic value)

For the **last class** i.e.(X > 120) we use: = 1- GAMMADIST(upper limit, **α**, **β**, 1)

Where upper limit is 120, **α** =1, **β**(mean) = 47.76, substituting these values in the formula =1-GAMMADIST(120,1,47.76,1) is 0.081060231 and the expected value is 8.024962821 which is greater than 5.

We perform the chi-square calculation for the (X > 120) interval which is 0.510964912. Performing the summation of chi-square values $\sum(O_i - e_i)^2/e_i$ we get the final value as 1.688895548

Calculating degree of freedom:

In Chi-square statistics degree of freedom is the number of class intervals – 1. It is denoted by v. where v = (k-1) and k is the number of class intervals. In this case, v = (4-1) = 3.

Calculating the Chi-Square value from the table:

- We use **Table A.5** to find the values.
- **$\chi^2(\alpha,v)$** where **α = 0.05** and **v = 3.**
- Therefore **$\chi^2(0.05,3)$ is 7.815**

Table A.5 (continued) Critical Values of the Chi-Squared Distribution

	α							
v	0.30	0.25	0.20	0.10	0.05	0.025	0.02	0.01
1	1.074	1.323	1.642	2.706	3.841	5.024	5.412	6.635
2	2.408	2.773	3.219	4.605	5.991	7.378	7.824	9.210
3	3.665	4.108	4.642	6.251	7.815	9.348	9.837	11.345
4	4.878	5.385	5.989	7.779	9.488	11.143	11.668	13.277
5	6.064	6.626	7.289	9.236	11.070	12.832	13.388	15.086

Decision Rule and Conclusion:

a decision rule is to

Reject H_0 when $\chi^2 > \chi^2_{\alpha,k-1}$

$\chi 2_{(Statistic)} < \chi 2_{(0.05,3)}$ i.e. 1.688 < 7.815.

Hence we fail to Reject the Null hypothesis. The given dataset follows an Exponential distribution.

We can conclude with 95% confidence that the given distribution of dataset 2 follows an Exponential distribution as previously mentioned.

REFERENCES

1. https://www.kaggle.com/datasets/rkiattisak/salaly-prediction-for-beginer.

Appendices I

Raw Data Set 1 – Salaries of employees at a company

	JOB TITLE	SALARY
2	Sales Representative	30000
3	Junior Accountant	35000
4	Junior Sales Representative	35000
5	Junior Customer Support Specialist	35000
6	Junior Designer	40000
7	Sales Associate	40000
8	Junior Accountant	40000
9	Junior Marketing Coordinator	40000
10	Customer Service Representative	45000
11	Junior Account Manager	45000
12	Junior HR Generalist	45000
13	Junior Recruiter	45000
14	Junior Copywriter	45000
15	Junior Business Development Associate	45000
16	Junior Data Scientist	45000
17	Marketing Analyst	50000
18	Technical Support Specialist	50000
19	Digital Content Producer	50000
20	Junior Operations Analyst	50000
21	Junior Financial Analyst	50000
22	Junior HR Coordinator	50000
23	Junior Accountant	50000
24	Marketing Coordinator	55000
25	Junior Software Developer	55000
26	Event Coordinator	55000
27	Junior Web Designer	55000
28	Administrative Assistant	55000
29	Junior Marketing Specialist	55000
30	Junior Data Analyst	60000
31	Junior Business Analyst	60000
32	Junior Project Manager	60000
33	Training Specialist	65000
34	Marketing Coordinator	65000
35	Recruiter	70000
36	Technical Recruiter	70000
37	Junior Marketing Manager	70000
38	Junior Software Developer	70000
39	Junior Software Engineer	70000
40	Business Analyst	75000
41	Customer Success Manager	75000
42	HR Manager	80000
43	Content Marketing Manager	80000
44	Data Analyst	80000
45	HR Generalist	80000
46	Junior Web Developer	80000
47	Customer Service Manager	80000

48	Senior Accountant	80000
49	Senior Marketing Coordinator	80000
50	Senior Human Resources Coordinator	80000
51	Web Developer	85000
52	Office Manager	85000
53	Senior Financial Advisor	85000
54	Business Intelligence Analyst	85000
55	Senior Operations Analyst	85000
56	Senior Marketing Specialist	85000
57	Senior Marketing Analyst	85000
58	Senior Operations Analyst	85000
59	Senior Marketing Analyst	85000
60	Senior Operations Analyst	85000
61	Digital Marketing Manager	90000
62	Business Development Manager	90000
63	Public Relations Manager	90000
64	IT Support Specialist	90000
65	Senior Business Analyst	90000
66	Senior Business Analyst	90000
67	Senior Sales Representative	90000
68	Social Media Manager	95000
69	Financial Advisor	95000
70	Senior Financial Analyst	95000
71	Senior HR Generalist	95000
72	Senior Marketing Analyst	95000
73	Senior Financial Analyst	95000
74	Senior Account Executive	95000
75	Senior Product Manager	95000
76	Senior Project Coordinator	95000
77	Senior Operations Manager	95000
78	Product Marketing Manager	95000
79	Software Project Manager	95000
80	Senior Software Engineer	100000
81	Senior Marketing Analyst	100000
82	Sales Manager	100000
83	Senior Quality Assurance Analyst	100000
84	Senior Financial Advisor	100000
85	Senior Sales Representative	100000
86	Senior Product Development Manager	100000
87	Senior Financial Analyst	100000
88	Senior Financial Advisor	100000
89	Senior Training Specialist	100000
90	Product Manager	105000
91	Supply Chain Manager	105000
92	Senior Software Developer	105000
93	Senior Software Engineer	105000
94	Operations Analyst	110000

95	Senior Graphic Designer	110000
96	Sales Operations Manager	110000
97	Senior Marketing Manager	110000
98	Senior IT Support Specialist	110000
99	Senior Account Manager	110000
100	Senior Operations Manager	115000
101	IT Manager	120000
102	Creative Director	120000
103	Principal Scientist	120000
104	Senior Product Manager	120000
105	Senior Product Marketing Manager	120000
106	Senior Business Development Manager	120000
107	Senior Human Resources Manager	120000
108	Senior Project Coordinator	130000
109	Supply Chain Analyst	130000
110	Project Manager	135000
111	Senior Project Manager	135000
112	Senior Sales Manager	135000
113	Senior Scientist	140000
114	Senior Project Manager	140000
115	Senior Product Designer	140000
116	Senior Project Manager	140000
117	Senior Marketing Manager	140000
118	Senior Financial Analyst	150000
119	Senior HR Manager	150000
120	Senior Manager	150000
121	Senior Data Scientist	150000
122	Senior Researcher	150000
123	Operations Manager	160000
124	Research Scientist	160000
125	Director of Operations	160000
126	Senior Research Scientist	160000
127	Director of Operations	170000
128	Senior Marketing Manager	170000
129	Principal Engineer	170000
130	Director of Sales	175000
131	Director of Product Management	175000
132	Senior Data Scientist	180000
133	Human Resources Director	180000
134	Director of Marketing	180000
135	Director of Finance	180000
136	Director of Human Resources	185000
137	Research Director	190000
138	Director	200000
139	VP of Finance	200000
140	Chief Data Officer	220000
141	Chief Technology Officer	250000

Excel formulas for table calculations (Dataset 1) :

Class probability: We use the Normal distribution formula to calculate the values.

Before combining Intervals :

1st Interval (X ≤ 65000) =NORMDIST(65000,100392.8571,44114.25591,1)

2nd Interval (65000 < X ≤ 105000) =NORMDIST(105000,100392.8571,44114.25591,1)-NORMDIST(65000,100392.8571,44114.25591,1)

3rd Interval(105000 < X ≤ 145000)=NORMDIST(145000,100392.8571,44114.25591,1)-NORMDIST(105000,100392.8571,44114.25591,1)

4th Interval(145000 < X ≤ 185000)=NORMDIST(185000,100392.8571,44114.25591,1)-NORMDIST(145000,100392.8571,44114.25591,1)

5th Interval(185000 < X ≤ 225000)=NORMDIST(225000,100392.8571,44114.25591,1)-NORMDIST(185000,100392.8571,44114.25591,1)

6th Interval(X > 225000) = 1 -NORMDIST(225000,100392.8571,44114.25591,1)

After combining Intervals:

1st Interval (X ≤ 65000) =NORMDIST(65000,100392.8571,44114.25591,1)

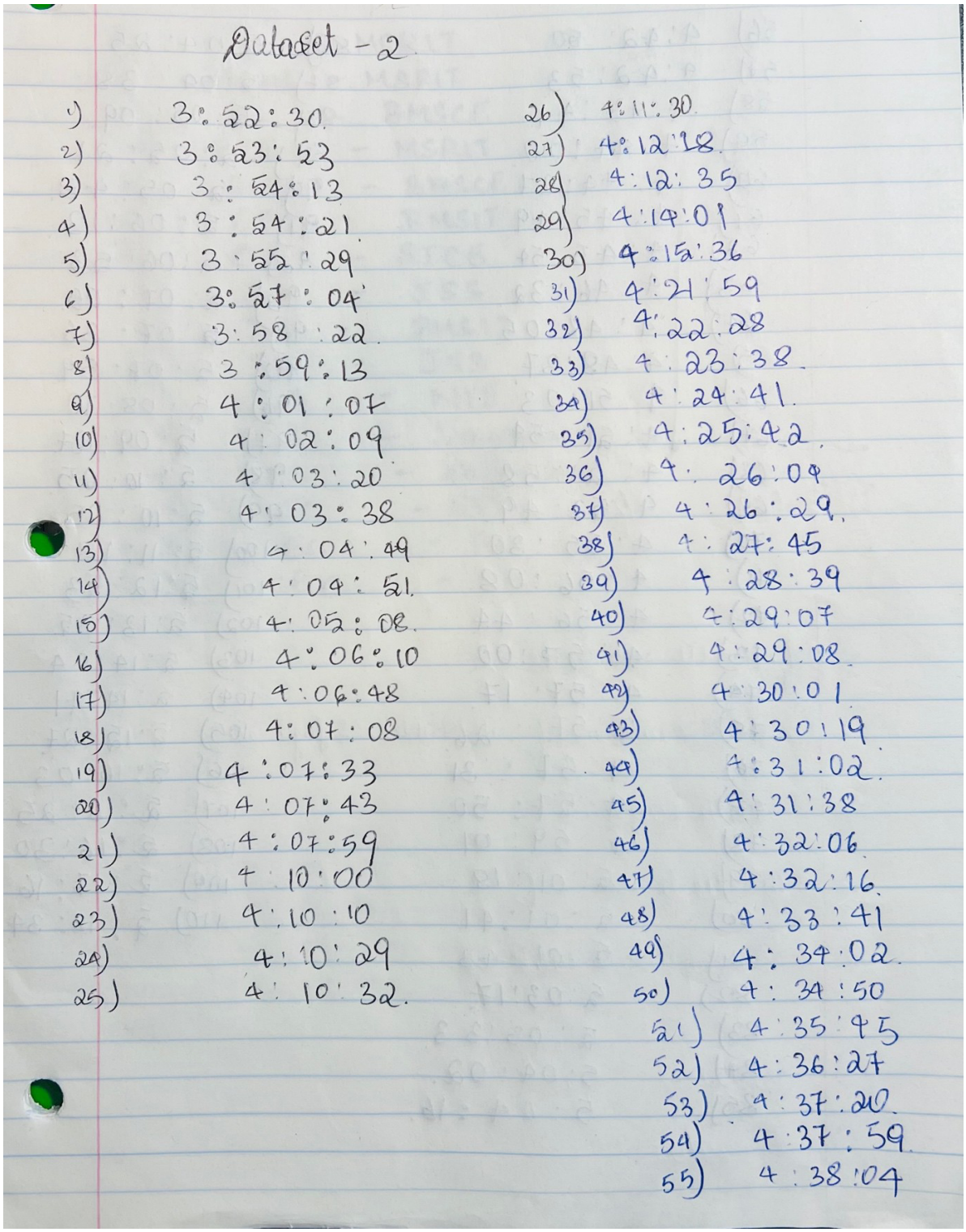
2nd Interval (65000 < X ≤ 105000) =NORMDIST(105000,100392.8571,44114.25591,1)-NORMDIST(65000,100392.8571,44114.25591,1)

3rd Interval(105000 < X ≤ 145000)=NORMDIST(145000,100392.8571,44114.25591,1)-NORMDIST(105000,100392.8571,44114.25591,1)

4th Interval(X >145000) =1-NORMDIST(145000,100392.8571,44114.25591,1)

Appendices II

Raw Data Set 2 – Time intervals of individuals exiting the Chase building at 500 East Border.



Dataset - 2

1) 3:52:30.
2) 3:53:53
3) 3:54:13
4) 3:54:21
5) 3:55:29
6) 3:57:04
7) 3:58:22.
8) 3:59:13
9) 4:01:07
10) 4:02:09
11) 4:03:20
12) 4:03:38
13) 4:04:49
14) 4:04:51.
15) 4:05:08.
16) 4:06:10
17) 4:06:48
18) 4:07:08
19) 4:07:33
20) 4:07:43
21) 4:07:59
22) 4:10:00
23) 4:10:10
24) 4:10:29
25) 4:10:32.
26) 4:11:30.
27) 4:12:18.
28) 4:12:35
29) 4:14:01
30) 4:15:36
31) 4:21:59
32) 4:22:28
33) 4:23:38.
34) 4:24:41.
35) 4:25:42.
36) 4:26:04
37) 4:26:29.
38) 4:27:45
39) 4:28:39
40) 4:29:07
41) 4:29:08.
42) 4:30:01
43) 4:30:19
44) 4:31:02.
45) 4:31:38
46) 4:32:06.
47) 4:32:16.
48) 4:33:41
49) 4:34:02.
50) 4:34:50
51) 4:35:45
52) 4:36:27
53) 4:37:20.
54) 4:37:59.
55) 4:38:04

56) 4:42:50
57) 4:42:53
58) 4:43:41
59) 4:44:00
60) 4:44:37
61) 4:45:49
62) 4:45:54
63) 4:46:32
64) 4:48:05
65) 4:48:37
66) 4:51:03
67) 4:51:54
68) 4:51:58
69) 4:52:49
70) 4:55:30
71) 4:56:08
72) 4:56:44
73) 4:57:00
74) 4:57:17
75) 4:57:26
76) 4:57:31
77) 4:57:50
78) 4 59:01
79) 5:01:18
80) 5:01:41
81) 5:01:43
82) 5:03:17
83) 5:03:33
84) 5:04:02
85) 5:04:16
86) 5:04:25
87) 5:04:38
88) 5:05:09
89) 5:05:21
90) 5:05:48
91) 5:06:13
92) 5:06:53
93) 5:07:05
94) 5:07:36
95) 5:07:59
96) 5:08:10
97) 5:09:27
98) 5:10:05
99) 5:10:25
100) 5:11:18
101) 5:12:53
102) 5:13:35
103) 5:14:04
104) 5:14:41
105) 5:15:27
106) 5:16:03
107) 5:16:25
108) 5:16:30
109) 5:17:16
110) 5:18:34

1	Clock Timings	Interval (s)
2	15:52:30	83.00
3	15:53:53	20.00
4	15:54:13	8.00
5	15:54:21	68.00
6	15:55:29	95.00
7	15:57:04	78.00
8	15:58:22	51.00
9	15:59:13	114.00
10	16:01:07	62.00
11	16:02:09	71.00
12	16:03:20	18.00
13	16:03:38	71.00
14	16:04:49	2.00
15	16:04:51	17.00
16	16:05:08	62.00
17	16:06:10	38.00
18	16:06:48	20.00
19	16:07:08	25.00
20	16:07:33	10.00
21	16:07:43	16.00
22	16:07:59	121.00
23	16:10:00	10.00
24	16:10:10	19.00
25	16:10:29	3.00
26	16:10:32	58.00
27	16:11:30	48.00
28	16:12:18	17.00
29	16:12:35	86.00
30	16:14:01	95.00
31	16:15:36	383.00
32	16:21:59	29.00
33	16:22:28	70.00
34	16:23:38	63.00
35	16:24:41	61.00
36	16:25:42	22.00
37	16:26:04	25.00
38	16:26:29	76.00
39	16:27:45	54.00
40	16:28:39	28.00
41	16:29:07	1.00
42	16:29:08	53.00
43	16:30:01	18.00
44	16:30:19	43.00
45	16:31:02	36.00
46	16:31:38	28.00
47	16:32:06	10.00
48	16:32:16	85.00
49	16:33:41	21.00
50	16:34:02	48.00

51	16:34:50	55.00
52	16:35:45	42.00
53	16:36:27	53.00
54	16:37:20	39.00
55	16:37:59	5.00
56	16:38:04	286.00
57	16:42:50	3.00
58	16:42:53	48.00
59	16:43:41	19.00
60	16:44:00	37.00
61	16:44:37	72.00
62	16:45:49	5.00
63	16:45:54	38.00
64	16:46:32	93.00
65	16:48:05	32.00
66	16:48:37	146.00
67	16:51:03	51.00
68	16:51:54	4.00
69	16:51:58	51.00
70	16:52:49	161.00
71	16:55:30	38.00
72	16:56:08	36.00
73	16:56:44	16.00
74	16:57:00	17.00
75	16:57:17	9.00
76	16:57:26	5.00
77	16:57:31	19.00
78	16:57:50	71.00
79	16:59:01	137.00
80	17:01:18	23.00
81	17:01:41	2.00
82	17:01:43	94.00
83	17:03:17	16.00
84	17:03:33	29.00
85	17:04:02	14.00
86	17:04:16	9.00
87	17:04:25	13.00
88	17:04:38	31.00
89	17:05:09	12.00
90	17:05:21	27.00
91	17:05:48	25.00
92	17:06:13	40.00
93	17:06:53	12.00
94	17:07:05	31.00
95	17:07:36	23.00
96	17:07:59	11.00
97	17:08:10	77.00
98	17:09:27	38.00
99	17:10:05	20.00
100	17:10:25	53.00
101	17:11:18	

Excel formulas for table calculations (Dataset 2) :

Class probability: We use the Gamma Distribution formula to calculate the values.

Before combining Intervals :

1st Interval (X ≤ 40) =GAMMADIST(40,1,47.76,1)

2nd Interval(40 < X ≤ 80) =GAMMADIST(80,1,47.76,1)-GAMMADIST(40,1,47.76,1)

3rd Interval(80 < X ≤ 120) =GAMMADIST(120,1,47.76,1)-GAMMADIST(80,1,47.76,1)

4th Interval(120 < X ≤ 160) =GAMMADIST(160,1,47.76,1)-GAMMADIST(120,1,47.76,1)

5th Interval(160 < X ≤ 200) =GAMMADIST(200,1,47.76,1)-GAMMADIST(160,1,47.76,1)

6th Interval(200 < X ≤ 240) =GAMMADIST(240,1,47.76,1)-GAMMADIST(200,1,47.76,1)

7th Interval(240 < X ≤ 280) =GAMMADIST(280,1,47.76,1)-GAMMADIST(240,1,47.76,1)

8th Interval(280 < X ≤ 320) =GAMMADIST(320,1,47.76,1)-GAMMADIST(280,1,47.76,1)

9th Interval(320 < X ≤ 360) =GAMMADIST(360,1,47.76,1)-GAMMADIST(320,1,47.76,1)

10th Interval(X > 360) =1-GAMMADIST(360,1,47.76,1)

After combining Intervals:

1st Interval (X ≤ 40) =GAMMADIST(40,1,47.76,1)

2nd Interval(40 < X ≤ 80) =GAMMADIST(80,1,47.76,1)-GAMMADIST(40,1,47.76,1)

3rd Interval(80 < X ≤ 120) =GAMMADIST(120,1,47.76,1)-GAMMADIST(80,1,47.76,1)

4th Interval(X > 120) =1-GAMMADIST(120,1,47.76,1)