PROJECT PART 1&2: GOODNESS-OF-FIT TEST

PROJECT REPORT

Submitted by:
Vedashrith Goud Survi

Date: 11/27/2023

Course: Introduction to Probability and Statistics

Instructor: Obiageli Ngwu Ph.D.

STATEMENT OF AUTHENTICITY

"I, VEDASHRITH GOUD SURVI, HEREBY DECLARE THAT I DID NOT GIVE OR RECEIVE ANY ASSISTANCE ON THIS PROJECT, AND THE REPORT SUBMITTED IS WHOLLY MY OWN."

VEDASHRITH GOUD SURVI

Table of Contents

1. Introduction

This report presents a descriptive analysis and goodness-of-fit test of two distinct data sets, set 1 and Set 2, each obtained from different sources and designed to explore various aspects of statistical analysis. The primary goal of this analysis is to gain valuable insights into the distribution and characteristics of data by using a range of statistical and visualisation methods. The hypothesis testing is to see how good the data fits their respective distributions.

Set 1 consists of observations representing a continuous random variable expected to follow a Normal Distribution. The observations are the measurements of body weight of adult males. The sample size of set1 is 150.

Set 2 consists of inter-arrival times within a sequence of events. Here the actual clock time of each event is recorded for 101 consecutive events. The interval between occurrences is determined by taking the difference between successive event times. As a result, set 2 will comprise of 100 inter-arrival times. This data will enable us to explore patterns and characteristics of inter-arrival times of consecutive events.

The analysis includes application of various statistical tools and methods on both the sets, such as calculation of sample mean to get the average of the observations and sample standard deviation to understand the variability of the data. Additionally, quartiles Q1, Q2 & Q3 will be calculated for further understanding of the distribution of data. Visualization techniques such as box & whisker plot and frequency histograms are used to visualize the data, thereby getting better understanding of the data.

We aim to gain insights into the data, evaluate the applicability of statistical models, and derive significant inferences from the characteristics of Sets 1 and 2 by this thorough analysis. These findings

will deepen our understanding of data analysis and its practical applications, highlighting the value of statistical methods in decision-making.

The test is on the hypothesis that set 1 sample observations fits the normal distribution and set 2 sample observations fits the exponential distribution with a significance level of 0.05 using chi-square test.

Objectives:

To calculate,

- Class probability
- Class Expected value.
- Chi-square component values

2.Data

Data Collection Process

Set 1:

Set 1 data was collected from a sample of adult males. **This data represents the random variable of weights (in lbs) of 150 adult males.** The data collection process involved measuring the body weight of a group of students. The process included:

1. Identifying a representative sample of adult males.
2. Accurately measuring and recording the body weight of each individual.

Set 2:

Set 2 data represents the Random variable of inter-arrival times of students entering the University Library. The data collection process included:

1. Recording the actual clock time (to the nearest second) of each student's arrival.
2. Determining the inter-arrival time by taking the difference between successive arrival times.

Appendix I: Raw Data for Set 1

The raw data for Set 1 is provided in Appendix I. This data consists of the body weight measurements of adult males.

Appendix II: Raw Data for Set 2

The raw data for Set 2 is provided in Appendix II. This data includes recorded actual clock times of customer arrivals and the corresponding inter-arrival times.

3.Descriptive Statistics & Goodness-of-fit test

Sample Mean and Sample Standard Deviation

For both sets, we calculated the sample mean (average) and sample standard deviation to measure central tendency and dispersion, respectively.

- Sample mean is calculated using the formula,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$ [1]

  Where x_1, x_2, …, x_n are individual values, n is the total number of observations.

- Sample mean gives the estimated mean of total population.
- Excel function: AVERAGE(<data>)
- Sample standard deviation S is calculated using the formula,

$$s^2 = \frac{1}{n-1}\left[\sum x_i^2 - \left(\sum x_i\right)^2 / n\right]$$ [2]

- The above formula is to calculate variance. Standard deviation is simply the square root of the variance.
- Excel function: STDEV(<data>) = SQRT(VAR(<data>))

Quartiles (Q1, Q2, Q3)

Quartiles Q1, Q2 (median), and Q3 were computed to understand the data's distribution and variability.

- Q1 = First / Lower Quartile = 25th percentile.
- Q2 = Second / Middle Quartile = 50th percentile.

- Q3 = Third / Upper Quartile = 75th percentile.
- Median is the second quartile (Q2).
- Excel function: PERCENTILE(<data>, f), where 0 < f < 1.

Box-and-Whisker Plots

The Box-and-Whisker Plot also known as Box plot is helpful for reflecting properties of a sample. The plot encloses the interquartile range of the data in a box that has median. The interquartile range has the Q3(upper quartile) and Q1(Lower quartile) as its boundaries. There are “whiskers” on both sides, that represent the outliers in the sample. If the sample is sufficiently large, the plot shows the centre of location, variability, and the degree of symmetry.

The “Whiskers” represent the outliers in the sample data. Outliers are the observations that are unusually far from the total observations. There are many statistical tests designed to find outliers in a dataset. The Box Plot is viewed as a diagnostic tool for visualisation of the data.

Box-and-Whisker plots were constructed to visualize the distribution and identify outliers in Sets 1 and 2.

Frequency Tables and Histograms

Frequency tables and histograms were created to explore the frequency distribution of values in each data set. Frequency tables provide a summary of the data set distribution by classifying data points into intervals called class intervals. Histograms provide a visual representation of these frequency distributions.

Goodness-of-fit Test:

Hypothesis:

- H_0: The data follow this theoretical distribution.
- H_1: The data do not follow this theoretical distribution.

Procedure:

- Designate a set of disjoint classes (i =1,2,…,k)
- Obtain observed frequencies (o_i) from the data and expected frequencies (e_i) using the distribution.
- Calculate chi-square value:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

The Chi-square test statistic:

- Approximate chi-square distribution with $k - 1$ d.f.
- Must have each e_i >= 5. May need to combine classes to satisfy this.

Decision Rule:

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

Set 1:

Mean: 126.76

Sample Standard Deviation: 11.78

Quartiles:

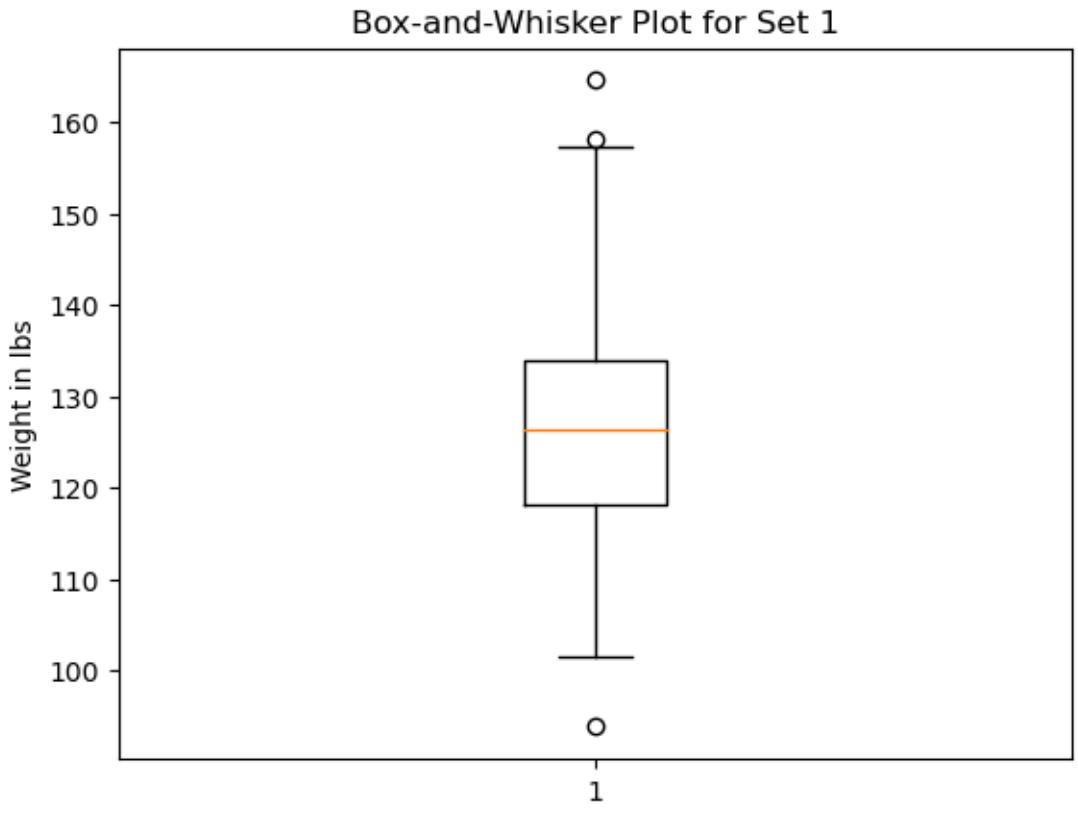
Q1: 118.18

Q2 (Median): 126.43

Q3: 134.02

Box-and-Whisker plot for set 1:

In the below figure, the Box-and-Whisker plot represents the body weight of the adult males in the sample.



[3]

The interquartile range and the middle 50% (25% - 75%) of the data are shown in the box in Figure 1. The median (Q2), or the line within the box, is the point at which half of the data falls above and half below. Within the data range, the whiskers reach the minimum and maximum values. we can observe that the body weight of adult males in Set 1 is symmetrically distributed with the median value (Q2). This data set comprises of few outliers that are identified as separate data points outside of the whiskers. The existence of three outliers indicates that this data set contains some extreme values. These outliers need to be examined to understand the impact on the data analysis.

Frequency Table

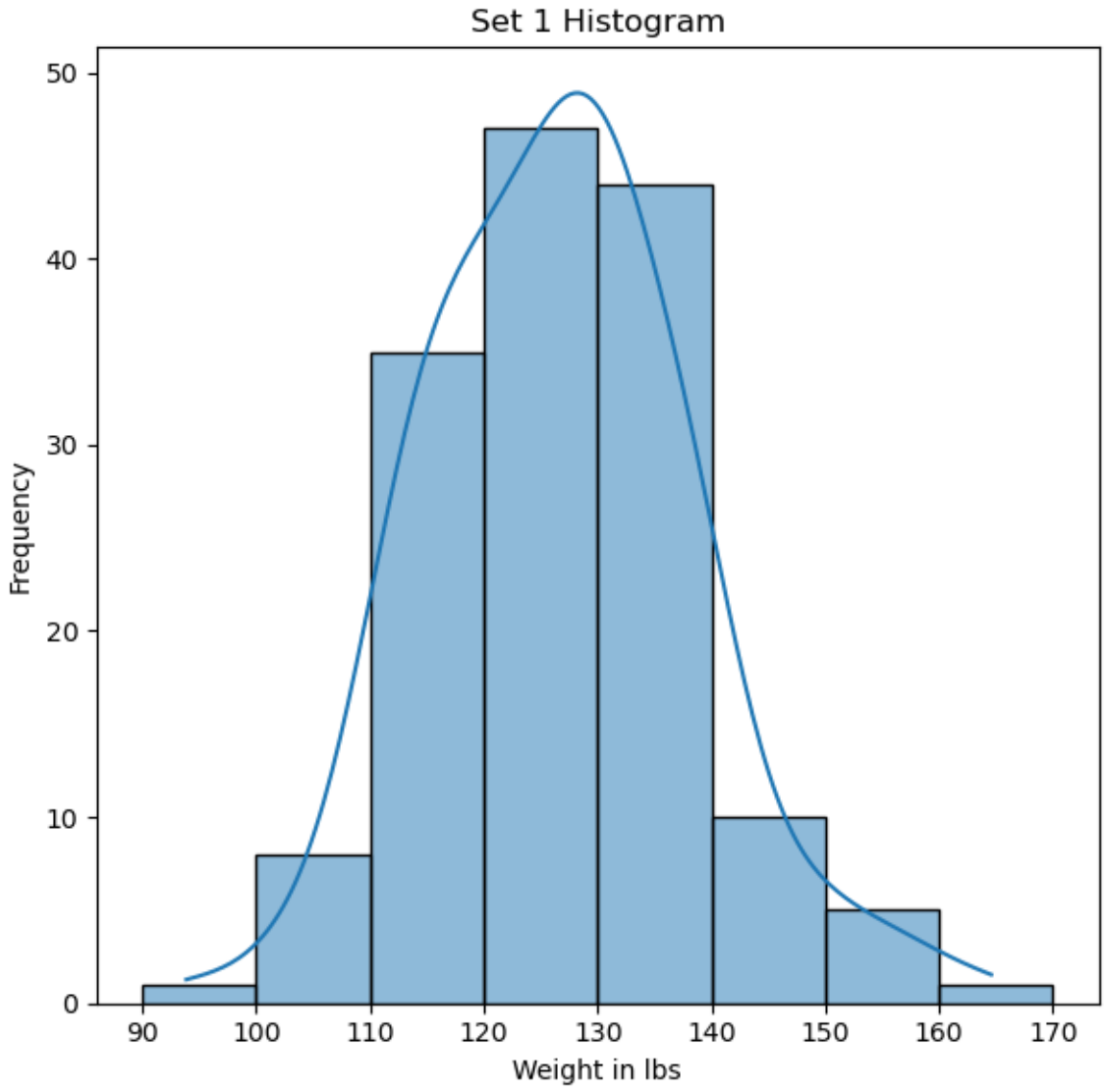
The following frequency table shows the distribution of body weights (in lbs) of adult males from set 1. The data has categorized into intervals of class size 10.

Frequency Table for Set 1

Class Interval	Frequency
$90 \leq x < 100$	1
$100 \leq x < 110$	8
$110 \leq x < 120$	35
$120 \leq x < 130$	47
$130 \leq x < 140$	44
$140 \leq x < 150$	10
$150 \leq x < 160$	5
$160 \leq x < 170$	1

Histograms

The figure below represents the histogram for the data in set 1. The histogram is created using the frequency table. The x-axis represents class intervals of weights, y-axis represents the frequency of observations in each class interval.



[5]

Distribution analysis

The histogram above for set 1 displays the visual representation of data distribution from set 1. This plot appears to follow a normal distribution as it is approximately symmetric. We can observe slight negligible skewness in the data distribution which is quite reasonable in real world data sets. In conclusion, the set 1 data follows **Normal Distribution**.

Goodness-of-fit test:

Hypothesis:

- H_0: The data follows Normal distribution.
- H_1: The data do not follow Normal distribution.

Procedure:

- No. of classes k = 5 after adjusting the e_i to be greater than 5.
- The respective expected values and chi-square components are calculated using excel formulae.
- For the normal distribution, we need 2 parameters: use the sample mean for μ = 126.76 and the sample standard deviation for σ = 11.78.
- Significance level (α) = 0.05.

Class	**Frequency (f_i)**	**Class Probability (pi)**	**Expected value (e_i)**	**Chi-square**
x ≤ 110	9	0.077404	11.688	0.618
110 < x ≤ 120	34	0.205629	31.049	0.280
120 < x ≤ 130	47	0.325325	48.124	0.091
130 < x ≤ 140	44	0.261122	39.429	0.529
x > 140	16	0.130519	19.708	0.697
Total	**150**	**1**	**150**	**2.218**

Test:

Critical value of Chi-square distribution with 4 d.f. and Significance level (α) = 0.05 is **9.488**. (From table A5)

Obtained Chi-square value: 2.218.

Decision Rule: Since 2.218 < 9.488, we ***fail to reject*** Null Hypothesis H_0.

Conclusion: ***We are 95% confident that Set 1 follows Normal Distribution.*** This is a Weak Conclusion.

Set 2:

Mean: 22.2

Sample Standard Deviation: 20.73

Quartiles:

Q1: 5.75

Q2 (Median): 16.0

Q3: 32.25

Box-and-Whisker plot for set 2:

In the below figure, the Box-and-Whisker plot represents the inter-arrival times of students entering a library.

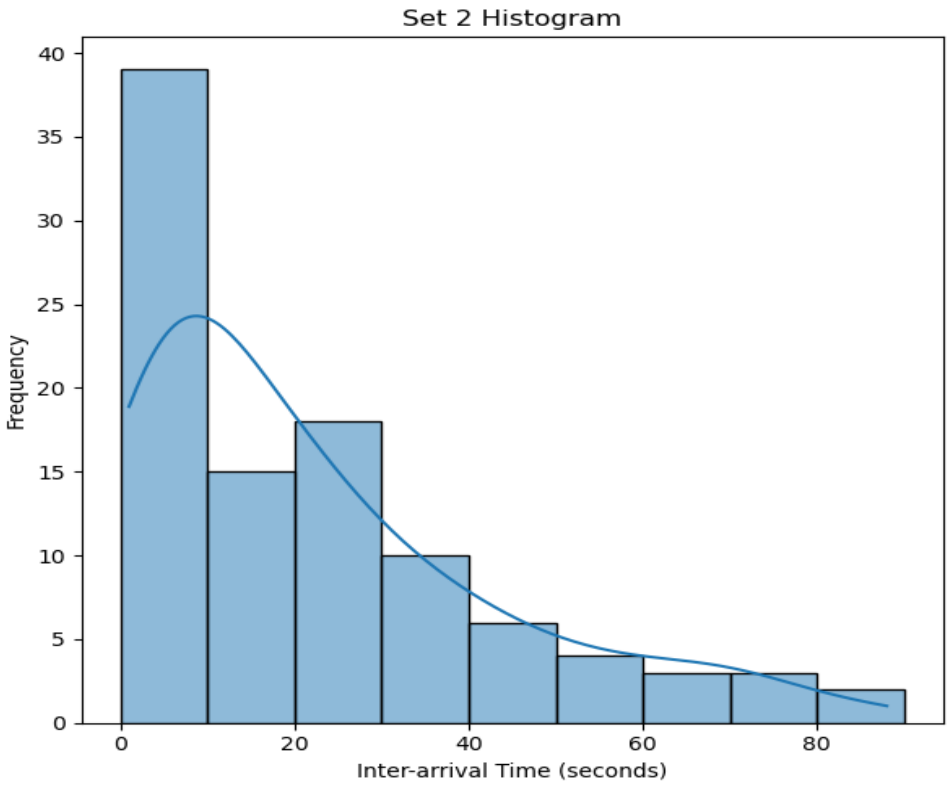
[4]

The interquartile range and the middle 50% (25% - 75%) of the data are shown in the box in Figure 2. We can observe that set 2 clearly shows positively skewed distribution. There are outliers beyond the Q3 point, indicates that there are few extreme observations. These outliers need to be examined to understand the impact on the data analysis.

By comparing the two plots, we can see that set 1 has a narrower spectrum of data whereas set 2 has wide spectrum of data. Both the datasets have a couple of outliers. This demonstrates that set 2 has more variability and longer time gaps between students compared to body weight of adult males in set1.

Histograms

The figure below represents the histogram for the data in set 2. The histogram is created using the frequency table. The x-axis represents class intervals of inter-arrival times, y-axis represents the frequency of observations in each class interval.



[6]

Distribution analysis

The histogram above for set 2 displays the visual representation of data distribution from set 2. This plot appears to follow an exponential distribution clearly. We can observe the right skewed nature of the data distribution which means there is a higher probability for shorter inter-arrival times. There is a slight variability in the distribution which is practical in real world scenarios. In conclusion, the set 2 data follow an **Exponential Distribution**.

Goodness-of-fit test:

Hypothesis:

- H_0: The data follows Exponential distribution.
- H_1: The data do not follow Exponential distribution.

Procedure:

- No. of classes k = 6 after adjusting the e_i to be greater than 5.
- The respective expected values and chi-square components are calculated using excel formulae.
- For the exponential, we need β: use the sample mean for β = 22.1.
- Significance level (α) = 0.05.

Class	**Frequency (f_i)**	**Class Probability (p_i)**	**Expected value (e_i)**	**Chi-square**
X ≤ 10	39	0.3626	36.26	0.206
10 < x ≤ 20	15	0.2311	23.11	2.848
20< x ≤ 30	18	0.1473	14.73	0.725
30 < x ≤ 40	10	0.0938	9.38	0.039
40 < x ≤ 50	6	0.0598	5.98	0.00006
X > 50	12	0.1051	10.51	0.209
Total	**100**	**1**	**100**	**4.028**

Test:

Critical value of Approximate Chi-square distribution with 5 d.f. and Significance level (α) = 0.05 is **11.070**. (From table A5).

Obtained Chi- square value: 4.028.

Decision Rule: Since 4.028 < 11.070, we ***fail to reject*** Null Hypothesis H_0.

Conclusion: ***We are 95% confident that Set 2 follows Exponential Distribution.*** This is a weak Conclusion.

4.Data application analysis

In this section, we explore the real-world applications based on the analysis performed on set 1 and set 2.

Set 1:

Set 1 consists of weight data which is used in many fields to make decisions and draw conclusions on the population. Types of applications:

- Health and Nutrition studies: The Weight data can be used to assess overall health and nutritional levels of students.
- Dietary Guidance: Nutritionists can utilize this data to provide nutritional diet guidance to students.
- Physical education curriculum: Weight data can help design and develop curriculum of physical education.

There are many other applications which help to promote overall health and well-being of students.

Set 2:

Set 2 consists of inter-arrival times of students entering the library. This data can be utilized for library management operations such as:

- Management: Based on the arrival times, library staff can allocate study spaces, staff members, books, and other resources more efficiently during peak hours. This gives the best experience for students and faculty.
- Staff scheduling: By understanding the patterns in the data, staff can be scheduled according to the needs of students for better assistance.
- Queue management: Inter-arrival times help to organize queues for library services such as book and device checkouts, computer access and queries.

5.Summary and Conclusion

In this report, we conducted descriptive analysis and goodness-of-fit test on sets 1 & 2, which represent body weight data and inter-arrival times of students entering a library. The descriptive statistics analysis of Sets 1 and 2 provides valuable insights into the characteristics of the data. We have calculated various parameters such as sample mean and standard deviation and quartiles to get the summary statistics of the data from set 1 and set 2. Box-and-whisker plots were plotted to understand the spread of data and, we were able to detect the outliers from the data sets. There are few outliers in set 1 which deviates from achieving a perfectly normal distribution for set1. Also, for set 2 there is a slight variability in the data distribution and the data distribution follows an exponential distribution.

In the goodness-of-fit test of set1 and set2, the Null and Alternate hypothesis is stated. We have calculated various parameters such as class probabilities, expected values and chi-square components. Almost all the expected values are close to the actual frequency for their respective classes. Based on the obtained Chi-square value and Critical value, we ***fail to reject Null hypothesis*** H_0 for both the sets which indicates that both the datasets follow their assumed distributions, and it is a weak conclusion.

The true nature of the data is like the way it was visualized. The plot was Normal for set1 and Exponential for set2 as seen in their respective histograms. This Test doesn't provide any statistically significant evidence that the data does not follow their assumed distributions.

6.Appendix I: Raw data of set 1

Index	Weight	Index	Weight	Index	Weight	Index	Weight
1	113.9851	41	111.7532	81	113.1916	127	112.2939
2	114.5627	42	137.1098	82	126.8993	128	120.8841
3	147.1207	43	123.4111	83	131.1289	129	164.5865
4	114.7406	44	116.3094	84	138.466	130	150.9348
5	120.7329	45	129.7406	85	120.1186	131	127.4484
6	135.6964	46	106.6076	86	127.9894	132	119.2318
7	128.7526	47	138.0277	87	136.1519	133	139.8852
8	130.4452	48	118.1795	88	123.4697	134	117.7473
9	131.4881	49	137.9495	89	112.9789	135	112.7753
10	127.5423	50	133.6875	90	123.9453	136	117.2326
11	112.9442	51	130.6206	91	131.5604	137	108.4964
12	113.787	52	133.2398	92	131.2806	138	125.317
13	117.4293	53	116.7499	93	135.1037	139	122.3296
14	119.1248	54	110.589	94	120.699	140	125.4864
15	130.6601	55	135.987	95	109.6276	141	141.4465
16	127.8872	56	130.0924	96	120.7825	142	140.1776
17	154.1866	57	139.991	97	126.4223	143	115.7894
18	115.0219	58	124.9344	98	134.3626	144	124.8506
19	118.1788	59	112.8224	99	123.5052	145	118.673
20	133.457	60	124.7065	100	115.3473	146	140.5952
21	126.4289	61	93.85997	101	137.818	147	118.2042
22	132.5879	62	132.68	102	142.4697	148	130.9092
23	130.9031	63	136.9751	103	121.5589	149	139.2995
24	111.4887	64	130.0858	104	130.488	150	135.7782
25	145.3706	65	139.7374	105	157.4164		
26	127.2747	66	151.6632	106	126.0415		
27	130.5737	67	128.6134	107	147.1265		
28	141.1106	68	124.6756	108	135.505		
29	129.5885	69	121.506	109	129.3979		
30	128.725	70	131.7102	110	136.8933		
31	124.4	71	116.534	111	108.6386		
32	134.4078	72	116.31	112	126.0584		
33	128.8091	73	121.4863	113	114.9174		
34	118.8537	74	124.9753	114	125.9758		
35	102.3968	75	140.744	115	109.7412		
36	123.15	76	135.6258	116	121.506		
37	130.5961	77	106.5245	117	139.0893		
38	112.8459	78	120.4203	118	111.1782		
39	117.4122	79	147.6976	119	128.6381		
40	158.2312	80	136.3169	120	126.1899		

Appendix II: Raw data of set 2

Index	Inter-arrival time (in seconds)	Actual recorded time	Index	Inter-arrival time (in seconds	Actual recorded time	Index	Inter-arrival time (in seconds	Actual recorded time
1	34	17:00:34	29	32	17:11:18	57	5	17:21:40
1	2	17:00:36	30	3	17:11:21	58	38	17:22:18
1	27	17:01:03	31	4	17:11:25	59	22	17:22:40
1	24	17:01:27	32	4	17:11:29	60	4	17:22:44
1	67	17:02:34	33	18	17:11:47	61	9	17:22:53
1	1	17:02:35	34	71	17:12:58	62	8	17:23:01
1	7	17:02:42	35	5	17:13:03	63	18	17:23:19
1	2	17:02:44	36	2	17:13:05	64	6	17:23:25
1	9	17:02:53	37	16	17:13:21	65	12	17:23:37
10	6	17:02:59	38	1	17:13:22	66	12	17:23:49
11	40	17:03:39	39	20	17:13:42	67	4	17:23:53
12	7	17:03:46	40	10	17:13:52	68	5	17:23:58
13	20	17:04:06	41	5	17:13:57	69	56	17:24:54
14	24	17:04:30	42	36	17:14:33	70	4	17:24:58
15	57	17:05:27	43	38	17:15:11	71	72	17:26:10
16	32	17:05:59	44	80	17:16:31	72	18	17:26:28
17	4	17:06:03	45	16	17:16:47	73	24	17:26:52
18	27	17:06:30	46	20	17:17:07	74	8	17:27:00
19	4	17:06:34	47	1	17:17:08	75	56	17:27:56
20	43	17:07:17	48	20	17:17:28	76	14	17:28:10
21	3	17:07:20	49	47	17:18:15	77	5	17:28:15
22	8	17:07:28	50	45	17:19:00	78	88	17:29:43
23	20	17:07:48	51	35	17:19:35	79	12	17:29:55
24	66	17:08:54	52	48	17:20:23	80	4	17:29:59
25	8	17:09:02	53	21	17:20:44	81	61	17:31:00
26	33	17:09:35	54	12	17:20:56	82	1	17:31:01

Index	Inter-arrival time (in seconds)	Actual recorded time
83	8	17:31:09
84	22	17:31:31
85	9	17:31:40
86	15	17:31:55
87	32	17:32:27
88	51	17:33:18
89	20	17:33:38
90	71	17:34:49
91	26	17:35:15
92	22	17:35:37
93	11	17:35:48
94	8	17:35:56
95	25	17:36:21
96	2	17:36:23
97	2	17:36:25
98	9	17:36:34
99	14	17:36:48
100	12	17:37:00

Appendix III: Excel Formulae

In Excel, to calculate c.d.f. values P[X ≤ x]: use NORMDIST(x, μ, σ, 1) for the normal distribution and GAMMADIST(x, 1, β, 1) for the exponential distribution.

Class probability calculations for set1:

- **x ≤ 110:** NORMDIST(110,126.76,11.78,1)
- **110 < x ≤ 120:** NORMDIST(120,126.76,11.78,1) NORMDIST(110,126.76,11.78,1)
- **120 < x ≤ 130:** NORMDIST(130,126.76,11.78,1) - NORMDIST(120,126.76,11.78,1)
- **130 < x ≤ 140:** NORMDIST(140,126.76,11.78,1) - NORMDIST(130,126.76,11.78,1)
- **x > 140:** 1 - NORMDIST(140,126.76,11.78,1)

Class probability calculations for set2:

- **x ≤ 10:** GAMMADIST(10,1,22.2,1)
- **10 < x ≤ 20:** =GAMMADIST(20,1,22.2,1) - GAMMADIST(10,1,22.2,1)
- **20 < x ≤ 30:** =GAMMADIST(30,1,22.2,1) - GAMMADIST(30,1,22.2,1)
- **30 < x ≤ 40:** =GAMMADIST(40,1,22.2,1) - GAMMADIST(30,1,22.2,1)
- **40 < x ≤ 50:** =GAMMADIST(50,1,22.2,1) - GAMMADIST(40,1,22.2,1)
- **X > 50:** 1 - GAMMADIST(50,1,22.2,1)

Expected value: $e_i = p_i \sum f_i$.

7.References:

[1],[2]: Formula pictures taken from class slides.

[3],[4]: Plots generated using software.

[5],[6]: Plots generated using software.