PROJECT PART II- EXPERIMENTAL DATA COLLECTION AND DESCRIPTIVE STATISTICS

Submitted by

ARUN JIJU JOSEPH

(1002140653)

in partial fulfilment for the award of the degree of

M.S. IN DATA SCIENCE

THE UNIVERSITY OF TEXAS AT ARLINGTON

DASC 5302 INTRODUCTION TO PROBABILITY AND STATISTICS

INSTRUCTOR: DR. OBIAGELI LAWRENTIA NGWU

NOVEMBER 2023

Honor Code: I ARUN JIJU JOSEPH did not give or receive any assistance on this project, and the report submitted is wholly my own.

CONTENTS

Introduction

Analysing real-world data is the main objective of the study. For this study, I examined and analysed two different sets of real data. The total bill amount for a sample of 100 Starbucks patrons is included in Dataset 1. Dataset 2 was gathered at the University of Texas at Arlington Central Library and looked at the time different between two students who entered the building via the main entrance. The population growth pattern from the sample is better understood by means of statistical analysis and visualisation of data. The results are provided after being statistically aggregated to provide a more comprehensive view of the population trend depending on sample. The information from the descriptive statistics is used in the Chi-square Goodness of Fit-Test. When there are enough observations, data classes are combined. The number of observations, class probability, class anticipated value, and Chi-square component values are computed for each data class. Excel tools are used to run the goodness of fit test, which predicts the distribution of a population based on a sample. Additionally provided are the chi-square values, test outcomes, and degrees of freedom.

Goodness-of-Fit Test

A goodness-of-fit is a statistical test that tries to determine whether a set of observed values match those expected under the applicable model. They can show you whether your sample data fit an expected set of data from a population with normal distribution. There are multiple types of goodness-of-fit tests, but the most common is the chi-square test. The chi-square test determines if a relationship exists between categorical data.

➢Compare the data to a theoretical distribution.

• H0: the data follow this theoretical distribution

• H1: the data do not follow this distribution

Chi-Square Formula

$$\chi 2 = \sum (O_i - E_i)^2 / E_i$$

- n = Total number of observations or total frequency
- $\chi 2$ = The $\chi 2$ statistic derived from the chi-square goodness of fit table.
- O_i = It is the observed value from the frequency table for each class interval.
- E_i = It is calculated by multiplying the class probabilities of each class by n.
- Each value of (Oi - Ei)2/Ei is referred to as a " *χ2* Class Component."
- Degree of Freedom (v) = number of groups - 1

The Chi-Square test has a number of assumptions, one of which is that the groups should be mixed appropriately if the anticipated value of the observations is less than 5. The significance level for statistical analysis is set at 0.05, or $\alpha = 0.05$.

Using the significance level(α) and degree of freedom(v), the estimated χ2 statistic is compared with the Tabulated value that is derived from Table A.5.

- If χ2 statistic > Tabulated value - Conclusion: Reject H0

- If χ2 statistic < Tabulated value - Conclusion: Fail to Reject H0

Chi-Square Goodness-of-Fit Test for Data set 1

The chi-square goodness of-Fit Test is performed using the descriptive statistical analysis results of Data set 1, the sample of 100 customer's bill amount. We had calculated the sample mean, sample median, sample standard deviation, and sample variance, respectively.

Statistics Value Units	Values	Units
Sample Mean	6.65	$
Sample Median	6.42	$
Sample Variance	10.51	$
Sample Standard Deviation	3.24	$

Table 1.1 Descriptive Statistics for Evaluating the bill amount of 100 customers within UTA Starbucks restaurant

For the chi-square goodness of fit test, the sample mean and sample standard deviation are taken from **Table 1.1**.

Making a hypothesis regarding Sample Data Set 1 is the first stage of the test, which has a level of significance of 0.05.

Null Hypothesis (H0): The sample Data set 1 is sampled from Normal distribution with population mean equal to the sample mean and a population standard deviation equal to the sample standard deviation.

Alternative Hypothesis (H1): The sample Data set 1 is not sampled from Normal distribution with population mean equal to the sample mean and a population standard deviation equal to the sample standard deviation.

By extending the columns into six group intervals, an expanded frequency table is created in order to run the chi-square goodness of test on Data Set 1. From the extended frequency table, the class probability, class expected value, and class chi-square components were found.

Class	Observed Frequency(oi)	Class Probability	Expected value(ei)	χ2 Class Component
X ≤ 4	25	0.206706793	20.67	0.9070585389
4 < X ≤ 7	34	0.336305154	33.63	0.0040707701
7< X ≤ 10	27	0.306408428	30.64	0.4324281984
10< X ≤ 13	10	0.125574514	12.56	0.5217834395
13< X ≤ 16	3	0.023053025	2.30	0.2130434783
X >16	1	0.001952086	0.2	3.2
SUM	100	1.00	100.00	5.278384425

Table 1.2 Expanded Frequency table for performing Goodness of Fit Test on Data set 1

The Class probability of each class intervals is calculated by:
The class probabilities are:
P [X ≤ 4] = 0.206706793
P [4 < X ≤ 7]=P [X ≤ 7]-P [X ≤ 4] =0.336305154

$P [7 < X \leq 10] = P [X \leq 10] - P [X \leq 7] = 0.306408428$
$P [10 < X \leq 13] = P [X \leq 13] - P [X \leq 10] = 0.125574514$
$P [13 < X \leq 16] = P [X \leq 16] - P [X \leq 13] = 0.023053025$
$P [X > 16] = 1 - P [X \leq 16] = 0.001952086$

For the Normal distribution, the sample mean (μ) and the sample standard deviation (σ) are the two parameters taken to calculate the class probabilities. The excel formula NORMDIST (x, μ, σ, 1) is used to calculate $P [X \leq x]$.

Expected values are found by multiplying each class probabilities by total number of observations(n), here n = 100. But here the expected value in the class $13 < X \leq 16$ AND $X > 16$ is less than 5, so the class is regrouped as below.

Class	**Observed Frequency(oi)**	**Class Probability**	**Expected value(ei)**	**χ2 Class Component**
$X \leq 4$	25	0.206706793	20.67	0.9070585389
$4 < X \leq 7$	34	0.336305154	33.63	0.0040707701
$7 < X \leq 10$	27	0.306408428	30.64	0.4324281984
$X > 10$	14	0.150579625	15.06	0.0746082337
SUM	100	1.00	100.00	1.418165741

Table 1.3 Expanded frequency table with all the values with intervals combined

$P [X > 10] = 1 - P [X \leq 10] = 15.06$

Chi-square class components are found by the chi-square formula $(Oi – Ei)^2/ Ei$ for each class and added up to get χ2 test statistic, here the χ2 test statistic is 1.418165741.

The Tabulated value for the Dataset 1 is found from Table A.5, Critical values of the Chi-Squared Distribution.

Significance level(α) = 0.05
Degree of freedom(v) = No of groups – 1 = 4-1 = 3

Therefore $\chi2(\alpha, v) = \chi2(0.05, 3) = 7.815$
The Tabulated value is compared with χ2 test statistic to test our hypothesis.

Here the χ2 test statistic < Tabulated value i.e., 1.418165741< 7.815
Conclusion: Fail to Reject Null Hypothesis(H0)

Thus, a weak conclusion is obtained that Dataset 1 is sampled from Normal distribution with population mean equal to the sample mean and a population standard deviation equal to the sample standard deviation.

<u>Conclusion</u>

After descriptive statistical analysis of Data Set 1, it was discovered that the average bill amount is 6.65. Given the mean shoe size of 6.65, the mode of 6.12, and the median of 6.42, it can be concluded that the distribution is right skewed and not symmetric or normal. Since the result fail to reject the null hypothesis (H0), the Data Set 1 is derived from a population that is presumed to have a normal distribution, with sample mean and standard deviation equal to population mean and standard deviation, respectively. Despite the sample data's lacking uniformity with respect to the mean centre, the Goodness of Fit-Test suggested that the sample adheres to a normal distribution.

Chi-Square Goodness-of-Fit Test for Data set 2

The chi-square goodness of-Fit Test is performed using the descriptive statistical analysis results of Data set 2, the samples of the time differences between two students entering the library through the main door. We had calculated the sample mean, sample median, sample standard deviation, and sample variance, respectively.

Statistics Value Units	Values	Units
Sample Mean	7.18	Seconds
Sample Median	6	Seconds
Sample Variance	26.45	Seconds
Sample Standard Deviation	5.14	Seconds

Table 2.1 Descriptive Statistics Exploring the time interval between students entering the UTA Central Library

Table 2.1 displays the descriptive statistics values for the intervals between students entering the library. The 7.18 second obtained mean is used as **β** parameter for determining the class probabilities for each class.

The test is done with a level of significance of 0.05 and the first step is to make the hypothesis about the sample data set 2.

Null Hypothesis (H0): The sample data set 2 is sampled from an Exponential Distribution with a population mean equal to the sample mean.

Alternative Hypothesis (H1): The sample data set 2 is not sampled from an Exponential Distribution with a population mean equal to the sample mean.

By extending the columns into six group intervals, an expanded frequency table is created in order to run the chi-square goodness of test on Data Set 2. From the extended frequency table, the class probability, class expected value, and class chi-square components were found.

Class	Observed Frequency(oi)	Class Probability	Expected value(ei)	χ2 Class Component
$X \leq 3$	23	0.341524061	34.15	3.6404831625
$3 < X \leq 6$	28	0.224885377	22.49	1.3499377501
$6 < X \leq 9$	26	0.14808161	14.80	8.4756756757
$9 < X \leq 12$	11	0.097508177	9.76	0.1575409836
$12 < X \leq 15$	4	0.064206788	6.42	0.912211838
$X > 15$	8	0.123793987	12.38	1.549628433
SUM	100	1.00	100.00	16.08547784

Table 2.2 Expanded Frequency table for performing Goodness of Fit Test on Data set 2

The observed Frequency is found by adding up the number of observations in the interval. The Class probability of each class intervals is calculated by: The class probabilities are:

$P[X \leq 3] = 0.341524061$

$P[3 < X \leq 6] = P[X \leq 6] - P[X \leq 3] = 0.224885377$

$P[6 < X \leq 9] = P[X \leq 9] - P[X \leq 6] = 0.14808161$

$P[9 < X \leq 12] = P[X \leq 12] - P[X \leq 9] = 0.097508177$

$P[12 < X \leq 15] = P[X \leq 15] - P[X \leq 12] = 0.064206788$

$P[X > 15] = 1 - P[X \leq 15] = 0.123793987$

For the Exponential distribution, the sample mean (μ) is taken as β parameter to calculate the class probabilities. The excel formula GAMMADIST (x, α, β, 1) is used to calculate $P[X \leq x]$. α =1 as we are using the function for exponential distribution.

Chi-square class components are found by the chi-square formula (Oi – Ei)^2/ Ei for each class and added up to get χ2 test statistic, here the χ2 test statistic is 16.08547784.

The Tabulated value for the Dataset 2 is found from Table A.5, Critical values of the Chi-Squared Distribution.

Significance level(α) = 0.05
Degree of freedom(v) = No of groups - 1 = 6-1 = 5

Therefore χ2(α, v) = χ2(0.05, 5) = 11.070

The Tabulated value is compared with χ2 test statistic to test our hypothesis.

Here the χ2 test statistic >Tabulated value i.e., 16.08547784> 11.070

Conclusion: Reject Null Hypothesis(H0)

Thus, a make a strong conclusion is obtained that the sample Data set 2 is not sampled from an Exponential Distribution with a population mean equal to the sample mean.

Conclusion

According to Data Set 2's descriptive statistics, students need an average of 7.18 seconds to enter the library. According to the earlier study based on the histogram, it appears that the sample data has an exponential distribution. When the sample data was first split up into only four classes, the frequency histogram showed an exponential decline. Class interval counts varied from 3 to 6, resulting in a non-exponential trend in the frequencies of each class. The more recent class interval observations benefited from the Chisquare Goodness of Fit-Test results. The goodness of fit test led to the rejection of the null hypothesis (H0), which states that the sample data is taken from a population with an exponential distribution. The goodness of fit test yielded a solid conclusion that the sample Data Set 2 did not follow an exponential distribution.

Appendix A

Data set 1

The raw data from Data Set 1 that was collected with different values and used for the descriptive statistical analysis are shown in the table below.

No	NAME	Gender	BILL AMOUNT ($)
1	Evan	Male	6.12
2	John	Male	3.54
3	Jose	Male	7.01
4	Sam	Male	12.09
5	Kate	Female	6.12
6	Marie	Female	9.01
7	Shaun	Male	10.33
8	Jr	Male	3.45
9	Erik	Male	2.07
10	Erica	Female	8.76
11	Briana	Female	11.43
12	Lewis	Male	10.87
13	Sai	Male	6.78
14	Ram	Male	5.67
15	Jessica	Female	6.12
16	Evana	Female	1.3
17	EKR	Male	2.41
18	Angel	Female	11.32
19	Fathima	Female	6.12
20	Nikita	Female	1.3
21	Laya	Female	2.36
22	Dj	Male	4.32
23	Lillian	Female	8.76
24	Lilly	Female	14.54
25	Sona	Female	3.32
26	Danny	Male	2.68
27	Yesha	Female	7.48
28	Gabriel	Male	5.53
29	Derrick	Male	3.01
30	John	Male	2.98
31	Isaac	Male	3.78
32	Ben	Male	4.66
33	Nora	Female	7.91
34	Lesley	Female	3.21
35	Erin	Female	4.34
36	Mathew	Male	5.32
37	Anna	Female	4.88
38	Ashley	Female	2.56
39	Jenna	Female	3.78
40	Lily	Female	6.7
41	Jason	Male	8.99
42	Juan	Male	5.45

43	Son	Male	6.75
44	Mellany	Female	3.45
45	Jenna	Female	6.32
46	Nora	Female	3.32
47	Sara	Female	4.19
48	Monica	Female	8.44
49	Michael	Male	7.31
50	Jake	Male	9.48
51	Jetty	Male	11.43
52	Ken	Male	6.51
53	Cassey	Female	3.32
54	Tj	Male	15.67
55	Edwin	Male	8.66
56	Chang	Male	10.41
57	Joanna	Female	2.58
58	Alen	Male	6.67
59	Dk	Male	7.66
60	Lee	Male	5.41
61	Mona	Female	6.14
62	Kendra	Female	8.16
63	Alisa	Male	9.88
64	Sona	Female	10.19
65	Katelyn	Female	8.94
66	Robert	Male	7.89
67	Charles	Male	18.95
68	Elizabeth	Female	10.42
69	Suffiya	Female	3.45
70	Bhavya	Female	6.12
71	Anita	Female	7.17
72	Sharon	Female	9.01
73	Edward	Male	8.76
74	Amy	Female	9.12
75	Maria	Female	10.12
76	Diane	Female	2.67
77	Jacqueline	Female	8.9
78	Sean	Male	5.11
79	Kathryn	Female	6.11
80	Lawrence	Male	2.91
81	Russell	Male	6.88
82	Doris	Female	3.51
83	Scott	Male	4.31
84	Stephen	Male	5.51
85	Gregory	Male	6.71
86	Omar	Male	8.78
87	Janet	Female	7.71
88	Jerry	Male	1.89
89	Kelly	Female	7.93
90	Peter	Male	7.51
91	Gerald	Male	4.87

92	Jordan	Male	13.45
93	Mary	Female	8.97
94	Claire	Female	4.54
95	Rohan	Male	6.88
96	Shine	Male	4.64
97	Stephen	Male	6.78
98	Kevin	Male	8.93
99	Ben	Male	4.81
100	Joseph	Male	3.93

Appendix B

Data set 2

The raw data from Data Set 2 that was collected with different values and used for the descriptive statistical analysis are shown in the table below.

No	Time	Time Difference (Seconds)	Seconds
1	11:35:29	0	0
2	11:35:36	00:00:07	7
3	11:36:08	00:00:32	32
4	11:36:20	00:00:12	12
5	11:36:25	00:00:05	5
6	11:36:34	00:00:09	9
7	11:36:50	00:00:16	16
8	11:37:00	00:00:10	10
9	11:37:21	00:00:21	21
10	11:37:32	00:00:11	11
11	11:37:48	00:00:16	16
12	11:37:53	00:00:05	5
13	11:38:00	00:00:07	7
14	11:38:05	00:00:05	5
15	11:38:09	00:00:04	4
16	11:38:16	00:00:07	7
17	11:38:21	00:00:05	5
18	11:38:24	00:00:03	3
19	11:38:30	00:00:06	6
20	11:38:50	00:00:20	20
21	11:39:05	00:00:15	15
22	11:39:07	00:00:02	2
23	11:39:15	00:00:08	8
24	11:39:25	00:00:10	10
25	11:39:28	00:00:03	3
26	11:39:30	00:00:02	2
27	11:39:37	00:00:07	7
28	11:39:50	00:00:13	13
29	11:39:58	00:00:08	8
30	11:40:03	00:00:05	5
31	11:40:07	00:00:04	4
32	11:40:25	00:00:18	18
33	11:40:32	00:00:07	7
34	11:40:35	00:00:03	3
35	11:40:40	00:00:05	5
36	11:40:48	00:00:08	8
37	11:40:55	00:00:07	7
38	11:41:07	00:00:12	12
39	11:41:25	00:00:18	18
40	11:41:30	00:00:05	5
41	11:41:35	00:00:05	5

42	11:41:40	00:00:05	5
43	11:41:48	00:00:08	8
44	11:41:50	00:00:02	2
45	11:41:57	00:00:07	7
46	11:41:59	00:00:02	2
47	11:42:02	00:00:03	3
48	11:42:10	00:00:08	8
49	11:42:21	00:00:11	11
50	11:42:24	00:00:03	3
51	11:42:29	00:00:05	5
52	11:42:33	00:00:04	4
53	11:42:35	00:00:02	2
54	11:42:39	00:00:04	4
55	11:42:43	00:00:04	4
56	11:42:47	00:00:04	4
57	11:42:48	00:00:01	1
58	11:42:56	00:00:08	8
59	11:42:58	00:00:02	2
60	11:43:00	00:00:02	2
61	11:43:12	00:00:12	12
62	11:43:15	00:00:03	3
63	11:43:20	00:00:05	5
64	11:43:38	00:00:18	18
65	11:43:43	00:00:05	5
66	11:43:47	00:00:04	4
67	11:43:49	00:00:02	2
68	11:43:50	00:00:01	1
69	11:43:53	00:00:03	3
70	11:44:02	00:00:09	9
71	11:44:08	00:00:06	6
72	11:44:09	00:00:01	1
73	11:44:17	00:00:08	8
74	11:44:25	00:00:08	8
75	11:44:28	00:00:03	3
76	11:44:30	00:00:02	2
77	11:44:40	00:00:10	10
78	11:44:44	00:00:04	4
79	11:44:48	00:00:04	4
80	11:44:57	00:00:09	9
81	11:45:05	00:00:08	8
82	11:45:12	00:00:07	7
83	11:45:25	00:00:13	13
84	11:45:30	00:00:05	5
85	11:45:38	00:00:08	8
86	11:45:39	00:00:01	1
87	11:45:41	00:00:02	2
88	11:45:48	00:00:07	7
89	11:45:58	00:00:10	10
90	11:46:05	00:00:07	7
91	11:46:20	00:00:15	15

92	11:46:28	00:00:08	8
93	11:46:33	00:00:05	5
94	11:46:37	00:00:04	4
95	11:46:45	00:00:08	8
96	11:46:52	00:00:07	7
97	11:46:55	00:00:03	3
98	11:47:06	00:00:11	11
99	11:47:18	00:00:12	12
100	11:47:23	00:00:05	5
101	11:47:27	00:00:04	4

Appendix C

Goodness of fit Calculations of Data set 1

Class	Observed Frequency(oi)	Class Probability Calculations	Expected Frequency (ei)	χ2 Class Component
X ≤ 4	Count the observations of each class from the frequency table	NORMDIST (4,6.65,3.24,1)	Take the probability for each class and multiply it by n	(oi-ei) ^2/ei
4 < X ≤ 7		NORMDIST (7,6.65,3.24,1)-NORMDIST (4,6.65,3.24,1)		
7 < X ≤ 10		NORMDIST (10,6.65,3.24,1)-NORMDIST (7,6.65,3.24,1)		
10 < X ≤ 13		NORMDIST (13,6.65,3.24,1)-NORMDIST (10,6.65,3.24,1)		
13 < X ≤ 16		NORMDIST (16,6.65,3.24,1)-NORMDIST (13,6.65,3.24,1)		
X >16		1-NORMDIST (16,6.65,3.24,1)		
Total	**n**	**1.00**	**n**	**χ2 statistic**

Class	Observed Frequency(oi)	Class Probability Calculations	Expected Frequency (ei)	χ2 Class Component
X ≤ 4	Count the observations of each class from the frequency table	NORMDIST (4,6.65,3.24,1)	Take the probability for each class and multiply it by n	(oi-ei) ^2/ei
4 < X ≤ 7		NORMDIST (7,6.65,3.24,1)-NORMDIST (4,6.65,3.24,1)		
7< X ≤ 10		NORMDIST (10,6.65,3.24,1)-NORMDIST (7,6.65,3.24,1)		
X > 10		1-NORMDIST (10,6.65,3.24,1)		
Total	**n**	**1.00**	**n**	**χ2 statistic**

The class probabilities are calculated using the excel functions.

P [X ≤ 4] = NORMDIST (4,6.65,3.24,1)
P [4< X ≤ 7] = P [X ≤ 7] − P [X ≤ 4] = NORMDIST (7,6.65,3.24,1) - NORMDIST (4,6.65,3.24,1)
P [7 < X ≤ 10] = P [X ≤ 10] − P [X ≤ 7] = NORMDIST (10,6.65,3.24,1) - NORMDIST (7,6.65,3.24,1)
P [X > 10] = 1 - P [X ≤ 10] = 1-NORMDIST (10,6.65,3.24,1)

Goodness of fit Calculations of Data set 2

Class	Observed Frequency(oi)	Class Probability Calculations	Expected Frequency (ei)	χ2 Class Component
X ≤ 3	Count the observations of each class from the frequency table	GAMMADIST (3,1,7.18,1)	Take the probability for each class and multiply it by n	(oi-ei) ^2/ei
3 < X ≤ 6		GAMMADIST (6,1,7.18,1)-GAMMADIST (3,1,7.18,1)		
6 < X ≤ 9		GAMMADIST (9,1,7.18,1)-GAMMADIST (6,1,7.18,1)		
9 < X ≤ 12		GAMMADIST (9,1,7.18,1)-GAMMADIST (6,1,7.18,1)		
12 < X ≤ 15		GAMMADIST (15,1,7.18,1)-GAMMADIST (15,1,7.18,1)		
X >15		1-GAMMADIST (15,1,7.18,1)		
Total	**n**	**1.00**	**n**	**χ2 statistic**

P [X ≤ 3] = GAMMADIST (3,1,7.18,1)
P [3 < X ≤ 6] = P [X ≤ 6] - P [X ≤ 3] = GAMMADIST (6,1,7.18,1)-GAMMADIST (3,1,7.18,1)
P [6 < X ≤ 9] = P [X ≤ 9] - P [X ≤ 6] = GAMMADIST (9,1,7.18,1)-GAMMADIST (6,1,7.18,1)
P [9 < X ≤ 12] = P [X ≤ 12] - P [X ≤ 9] = GAMMADIST (12,1,7.18,1)-GAMMADIST (9,1,7.18,1)
P [12 < X ≤ 15] = P [X ≤ 15] - P [X ≤ 12] = GAMMADIST (15,1,7.18,1)-GAMMADIST (12,1,7.18,1)
P [X > 15] = 1 - P [X ≤ 15] = 1-GAMMADIST (15,1,7.18,1)

References

- https://www.analyticsvidhya.com/blog/2019/11/what-is-chi-square-test-how-it-works/
- https://www.statisticshowto.com/probability-and-statistics/chi-square/
- https://medium.com/intuitionmath/chi-square-test-intuition-examples-and-step-by-step-calculation-de45c873abd1