Project Report Part 2

IE 3301-004

Muhammad Muawiz Farooqi

# Overview of Part I

The first part of this project consisted of using two sets of real-world data, obtained through secondary research, to observe trends. The first set of data was about the earth’s surface temperatures and did not appear to follow a normal distribution as previously expected but could contain underlying normal distribution curves [bimodal]. The second data set was for times between vehicle accidents in different boroughs of New York City, and it appeared to follow an exponential distribution and was skewed to the left.

# Overview of goal of Part II

The second part of this project aims to use the Chi-square Goodness-of-Fit Test to test the following hypotheses:

* Set 1 is sampled from a Normal Distribution with a population mean equal to the sample mean’s value and a population standard deviation equal to the sample standard deviation’s value.
* Set 2 is sampled from an Exponential Distribution with a population mean equal to the sample mean’s value.

Data sets from part 1 will be used and the number of observations, class probability, class expected value, and chi-square component values will be calculated for each data set.

# Goodness of Fit Tests

The analysis for both sets of data was done in Microsoft Excel.

**Data Set 1**

Chi-Square Table for set 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | |  |  |  |
| Lower  (minutes) | Upper  (minutes) | Observed Frequency | Class Probability | Expected Frequency |
| o(i) | p\_i= P(Lower < X < Upper) | e(i) = n\*p\_i |
| 1.746 | 3.506 | 7 | 0.037 | 5.55 |
| 3.506 | 5.266 | 23 | 0.052 | 7.8 |
| 5.266 | 7.026 | 15 | 0.068 | 10.2 |
| 7.026 | 8.786 | 12 | 0.084 | 12.6 |
| 8.786 | 10.546 | 13 | 0.096 | 14.4 |
| 10.546 | 12.306 | 10 | 0.103 | 15.45 |
| 12.306 | 14.066 | 11 | 0.103 | 15.45 |
| 14.066 | 15.826 | 6 | 0.096 | 14.4 |
| 15.826 | 17.586 | 9 | 0.084 | 12.6 |
| 17.586 | 19.346 | 10 | 0.069 | 10.35 |
| 19.346 | 21.106 | 11 | 0.053 | 7.95 |
| 21.106 | 22.866 | 16 | 0.038 | 5.7 |
| 22.866 | 24.626 | 7 | 0.025 | 3.75 |
|  |  | 150 |  | 136.2 |

Adjusted Chi-Square Table for set 1

|  |  |  |
| --- | --- | --- |
|  |  | Chi-square component |
| e(i) supercells | o(i) supercells | (oi-ei)^2/ei |
| 5.55 | 7 | 0.379 |
| 7.8 | 23 | 29.621 |
| 10.2 | 15 | 2.259 |
| 12.6 | 12 | 0.029 |
| 14.4 | 13 | 0.136 |
| 15.45 | 10 | 1.922 |
| 15.45 | 11 | 1.282 |
| 14.4 | 6 | 4.9 |
| 12.6 | 9 | 1.029 |
| 10.35 | 10 | 0.012 |
| 7.95 | 11 | 1.17 |
|  |  |  |
| 9.45 | 23 | 19.429 |

Chi-Square statistics for set 1

|  |  |  |  |
| --- | --- | --- | --- |
| Original (k): | 13 | Test Statistic: | 62.168 |
| Revised (k): | 12 | Critical value: | 19.675 |
| significance level: | 0.05 | Degree of freedom: | 11 |

This test uses a level of significance of 0.05 and the degree of freedom was found to be 11. The frequency distribution contains 13 bins which is approximately equal to the square root of the number of data points [150].

The null hypothesis in this test is that set 1 is assumed to be sampled from a normal distribution with a population mean equal to the sample mean’s value and a population standard deviation equal to the sample standard deviation’s value. The alternative hypothesis is that dataset 1 does not follow a normal distribution under the same conditions.

The classes and the observed frequency values in the Chi-Square Goodness-of-Fit Tests are the same ones used in the histogram in Part 1 of the project, obtained from the raw data tables. The class probabilities were calculated using normal distribution and used the upper and lower bounds of the classes. The expected frequency was then calculated using the total frequency and class probabilities. Supercells were then created for both expected and observed probabilities for values of expected probability that were lower than 5. Next, the chi-square component was calculated using both sets of supercells. Finally, the chi-square components were added up to calculate the test statistic which is the chi-square value of 62.168. A critical value was calculated using the degree of freedom and level of significance using chi-square distribution and was found to be 19.675.

The test statistic is determined to be larger than the critical value which shows that the null hypothesis is rejected. The large difference in the two values shows almost no correlation between set 1 and a normally distributed dataset. Therefore, dataset 1 does not follow a normal distribution.

**Data Set 2**

Chi-Square Table for set 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | |  |  |  |
| Lower (minutes) | Upper (minutes) | Observed Frequency | Class Probability | Expected Frequency |
| o(i) | p\_i= P(Lower < X < Upper) | e(i) = n\*p\_i |
| 0 | 4 | 40 | 0.359 | 52.055 |
| 4 | 8 | 33 | 0.23 | 33.35 |
| 8 | 12 | 35 | 0.148 | 21.46 |
| 12 | 16 | 14 | 0.095 | 13.775 |
| 16 | 20 | 4 | 0.061 | 8.845 |
| 20 | 24 | 5 | 0.039 | 5.655 |
| 24 | 28 | 6 | 0.025 | 3.625 |
| 28 | 32 | 1 | 0.016 | 2.32 |
| 32 | 36 | 3 | 0.01 | 1.45 |
| 36 | 40 | 0 | 0.007 | 1.015 |
| 40 | 44 | 3 | 0.004 | 0.58 |
| 44 | 48 | 1 | 0.003 | 0.435 |
|  |  | 145 |  | 144.565 |

Adjusted Chi-Square Table for set 2

|  |  |  |
| --- | --- | --- |
|  |  | Chi-square component |
| e(i) supercells | o(i) supercells | (oi-ei)^2/ei |
| 52.055 | 40 | 2.792 |
| 33.35 | 33 | 0.004 |
| 21.46 | 35 | 8.543 |
| 13.775 | 14 | 0.004 |
| 8.845 | 4 | 2.654 |
| 5.655 | 5 | 0.076 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
| 9.425 | 14 | 2.221 |

Chi-Square statistics for set 2

|  |  |  |  |
| --- | --- | --- | --- |
| Original (k): | 12 | Test Statistic: | 16.294 |
| Revised (k): | 7 | Critical value: | 12.592 |
| significance level: | 0.05 | Degree of freedom: | 6 |

This test uses a level of significance of 0.05 and the degree of freedom was found to be 6. The frequency distribution contains 12 bins which is approximately equal to the square root of the number of data points [145]. The null hypothesis in this test is that set 2 is assumed to be sampled from an exponential distribution with a population mean equal to the sample mean’s value. The alternative hypothesis is that dataset 2 does not follow an exponential distribution under the same conditions.

The classes and the observed frequency values in the Chi-Square Goodness-of-Fit Tests are the same ones used in the histogram in Part 1 of the project, obtained from the raw data tables. The class probabilities were calculated using exponential distribution and used the upper and lower bounds of the classes. The expected frequency was then calculated using the total frequency and class probabilities. Supercells were then created for both expected and observed probabilities for values of expected probability that were lower than 5. Next, the chi-square component was calculated using both sets of supercells. Finally, the chi-square components were added up to calculate the test statistic which is the chi-square value of 16.294. A critical value was calculated using the degree of freedom and level of significance using chi-square distribution and was found to be 12.592.

The test statistic is determined to be larger than the critical value which shows that the null hypothesis is rejected. There is a smaller difference between the two values which shows that there may be some correlation between set 1 and an exponentially distributed dataset, but it is not enough to be accepted with the level of significance of 0.05. Therefore, dataset 2 does not follow an exponential distribution.

# Manager’s Summary

Set 1:

The assumed notion that the data in set 1 follows a normal distribution is rejected as it is found to not be normally distributed.

Set 2:

The assumed notion that the data in set 2 follows an exponential distribution is rejected as it is found to not be exponentially distributed.

# Conclusion

In conclusion, the second part of this project used the Chi-square Goodness-of-Fit Test to test the following hypotheses:

* Set 1 is sampled from a Normal Distribution with a population mean equal to the sample mean’s value and a population standard deviation equal to the sample standard deviation’s value.
* Set 2 is sampled from an Exponential Distribution with a population mean equal to the sample mean’s value.

Both hypotheses were concluded to be rejected, with set 1 not following a normal distribution at all and set 2 not following an exponential distribution under the level of significance of 0.05 but could be accepted under a lower level of significance (0.01).

# Appendices

Appendix 1

|  |  |
| --- | --- |
| **Date**  **(MM/DD/YYYY)** | **Average Temperature** |
| 1/1/2001 | 5.441 |
| 2/1/2001 | 5.755 |
| 3/1/2001 | 10.06 |
| 4/1/2001 | 10.074 |
| 5/1/2001 | 14.17 |
| 6/1/2001 | 20.379 |
| 7/1/2001 | 21.01 |
| 8/1/2001 | 22.62 |
| 9/1/2001 | 17.434 |
| 10/1/2001 | 13.859 |
| 11/1/2001 | 5.409 |
| 12/1/2001 | 1.746 |
| 1/1/2002 | 5.265 |
| 2/1/2002 | 6.35 |
| 3/1/2002 | 8.581 |
| 4/1/2002 | 10.432 |
| 5/1/2002 | 12.783 |
| 6/1/2002 | 20.186 |
| 7/1/2002 | 21.523 |
| 8/1/2002 | 20.655 |
| 9/1/2002 | 17.008 |
| 10/1/2002 | 12.711 |
| 11/1/2002 | 8.125 |
| 12/1/2002 | 6.571 |
| 1/1/2003 | 3.564 |
| 2/1/2003 | 3.927 |
| 3/1/2003 | 9.624 |
| 4/1/2003 | 10.377 |
| 5/1/2003 | 15.1 |
| 6/1/2003 | 22.305 |
| 7/1/2003 | 22.528 |
| 8/1/2003 | 24.524 |
| 9/1/2003 | 18.952 |
| 10/1/2003 | 11.325 |
| 11/1/2003 | 8.229 |
| 12/1/2003 | 4.634 |
| 1/1/2004 | 5.265 |
| 2/1/2004 | 5.253 |
| 3/1/2004 | 6.593 |
| 4/1/2004 | 8.991 |
| 5/1/2004 | 12.617 |
| 6/1/2004 | 21.125 |
| 7/1/2004 | 22.177 |
| 8/1/2004 | 21.352 |
| 9/1/2004 | 19.14 |
| 10/1/2004 | 13.078 |
| 11/1/2004 | 6.235 |
| 12/1/2004 | 4.243 |
| 1/1/2005 | 2.646 |
| 2/1/2005 | 2.125 |
| 3/1/2005 | 7.878 |
| 4/1/2005 | 10.763 |
| 5/1/2005 | 15.782 |
| 6/1/2005 | 21.884 |
| 7/1/2005 | 22.923 |
| 8/1/2005 | 22.415 |
| 9/1/2005 | 17.637 |
| 10/1/2005 | 13.101 |
| 11/1/2005 | 6.315 |
| 12/1/2005 | 3.683 |
| 1/1/2006 | 2.656 |
| 2/1/2006 | 3.8 |
| 3/1/2006 | 8.566 |
| 4/1/2006 | 11.388 |
| 5/1/2006 | 16.416 |
| 6/1/2006 | 20.639 |
| 7/1/2006 | 24.298 |
| 8/1/2006 | 21.052 |
| 9/1/2006 | 19.334 |
| 10/1/2006 | 14.448 |
| 11/1/2006 | 10.054 |
| 12/1/2006 | 4.032 |
| 1/1/2007 | 4.315 |
| 2/1/2007 | 6.868 |
| 3/1/2007 | 7.149 |
| 4/1/2007 | 10.558 |
| 5/1/2007 | 13.661 |
| 6/1/2007 | 17.417 |
| 7/1/2007 | 21.366 |
| 8/1/2007 | 20.619 |
| 9/1/2007 | 18.215 |
| 10/1/2007 | 12.238 |
| 11/1/2007 | 6.443 |
| 12/1/2007 | 3.816 |
| 1/1/2008 | 5.708 |
| 2/1/2008 | 7.133 |
| 3/1/2008 | 7.585 |
| 4/1/2008 | 10.576 |
| 5/1/2008 | 12.996 |
| 6/1/2008 | 18.154 |
| 7/1/2008 | 21.296 |
| 8/1/2008 | 21.843 |
| 9/1/2008 | 16.876 |
| 10/1/2008 | 11.633 |
| 11/1/2008 | 5.491 |
| 12/1/2008 | 3.709 |
| 1/1/2009 | 3.005 |
| 2/1/2009 | 5.23 |
| 3/1/2009 | 9.026 |
| 4/1/2009 | 9.18 |
| 5/1/2009 | 15.945 |
| 6/1/2009 | 20.272 |
| 7/1/2009 | 22.57 |
| 8/1/2009 | 23.71 |
| 9/1/2009 | 18.501 |
| 10/1/2009 | 14.611 |
| 11/1/2009 | 9.194 |
| 12/1/2009 | 4.264 |
| 1/1/2010 | 3.443 |
| 2/1/2010 | 4.061 |
| 3/1/2010 | 6.672 |
| 4/1/2010 | 11.455 |
| 5/1/2010 | 12.979 |
| 6/1/2010 | 18.203 |
| 7/1/2010 | 23.974 |
| 8/1/2010 | 22.934 |
| 9/1/2010 | 18.323 |
| 10/1/2010 | 11.684 |
| 11/1/2010 | 6.327 |
| 12/1/2010 | 4.022 |
| 1/1/2011 | 4.232 |
| 2/1/2011 | 5.804 |
| 3/1/2011 | 7.579 |
| 4/1/2011 | 13.499 |
| 5/1/2011 | 16.44 |
| 6/1/2011 | 19.625 |
| 7/1/2011 | 21.49 |
| 8/1/2011 | 22.852 |
| 9/1/2011 | 19.865 |
| 10/1/2011 | 14.487 |
| 11/1/2011 | 8.841 |
| 12/1/2011 | 4.604 |
| 1/1/2012 | 3.753 |
| 2/1/2012 | 2.947 |
| 3/1/2012 | 8.808 |
| 4/1/2012 | 8.126 |
| 5/1/2012 | 15.943 |
| 6/1/2012 | 20.234 |
| 7/1/2012 | 21.831 |
| 8/1/2012 | 23.021 |
| 9/1/2012 | 18.204 |
| 10/1/2012 | 12.509 |
| 11/1/2012 | 7.61 |
| 12/1/2012 | 4.972 |
| 1/1/2013 | 4.691 |
| 2/1/2013 | 4.403 |
| 3/1/2013 | 7.504 |
| 4/1/2013 | 9.483 |
| 5/1/2013 | 11.629 |
| 6/1/2013 | 17.565 |

Set 1 raw data

Table

Description automatically generated

Chi-Square Table from Excel for set 1

Appendix 2

|  |  |
| --- | --- |
| **Crash Date**  **(MM/DD/YYYY)** | **Crash Time** |
| 3/1/2020 | 12:00 AM |
| 3/1/2020 | 12:00 AM |
| 3/1/2020 | 12:05 AM |
| 3/1/2020 | 12:05 AM |
| 3/1/2020 | 12:30 AM |
| 3/1/2020 | 12:42 AM |
| 3/1/2020 | 1:28 AM |
| 3/1/2020 | 1:30 AM |
| 3/1/2020 | 1:50 AM |
| 3/1/2020 | 2:05 AM |
| 3/1/2020 | 2:29 AM |
| 3/1/2020 | 2:40 AM |
| 3/1/2020 | 2:45 AM |
| 3/1/2020 | 3:05 AM |
| 3/1/2020 | 3:15 AM |
| 3/1/2020 | 3:20 AM |
| 3/1/2020 | 3:46 AM |
| 3/1/2020 | 3:50 AM |
| 3/1/2020 | 3:50 AM |
| 3/1/2020 | 4:30 AM |
| 3/1/2020 | 4:40 AM |
| 3/1/2020 | 4:48 AM |
| 3/1/2020 | 5:00 AM |
| 3/1/2020 | 5:15 AM |
| 3/1/2020 | 5:25 AM |
| 3/1/2020 | 5:35 AM |
| 3/1/2020 | 5:50 AM |
| 3/1/2020 | 6:25 AM |
| 3/1/2020 | 6:30 AM |
| 3/1/2020 | 6:39 AM |
| 3/1/2020 | 6:50 AM |
| 3/1/2020 | 6:54 AM |
| 3/1/2020 | 7:35 AM |
| 3/1/2020 | 7:45 AM |
| 3/1/2020 | 8:20 AM |
| 3/1/2020 | 8:30 AM |
| 3/1/2020 | 8:30 AM |
| 3/1/2020 | 9:00 AM |
| 3/1/2020 | 9:10 AM |
| 3/1/2020 | 9:30 AM |
| 3/1/2020 | 9:40 AM |
| 3/1/2020 | 9:49 AM |
| 3/1/2020 | 9:50 AM |
| 3/1/2020 | 9:50 AM |
| 3/1/2020 | 10:00 AM |
| 3/1/2020 | 10:00 AM |
| 3/1/2020 | 10:00 AM |
| 3/1/2020 | 10:20 AM |
| 3/1/2020 | 10:40 AM |
| 3/1/2020 | 10:54 AM |
| 3/1/2020 | 11:00 AM |
| 3/1/2020 | 11:05 AM |
| 3/1/2020 | 11:29 AM |
| 3/1/2020 | 11:45 AM |
| 3/1/2020 | 11:58 AM |
| 3/1/2020 | 12:00 PM |
| 3/1/2020 | 12:00 PM |
| 3/1/2020 | 12:08 PM |
| 3/1/2020 | 12:22 PM |
| 3/1/2020 | 12:30 PM |
| 3/1/2020 | 12:30 PM |
| 3/1/2020 | 12:45 PM |
| 3/1/2020 | 12:50 PM |
| 3/1/2020 | 12:52 PM |
| 3/1/2020 | 12:56 PM |
| 3/1/2020 | 1:00 PM |
| 3/1/2020 | 1:10 PM |
| 3/1/2020 | 1:10 PM |
| 3/1/2020 | 1:20 PM |
| 3/1/2020 | 1:30 PM |
| 3/1/2020 | 1:41 PM |
| 3/1/2020 | 1:41 PM |
| 3/1/2020 | 1:45 PM |
| 3/1/2020 | 2:00 PM |
| 3/1/2020 | 2:15 PM |
| 3/1/2020 | 2:15 PM |
| 3/1/2020 | 2:17 PM |
| 3/1/2020 | 2:21 PM |
| 3/1/2020 | 2:23 PM |
| 3/1/2020 | 2:24 PM |
| 3/1/2020 | 2:26 PM |
| 3/1/2020 | 2:30 PM |
| 3/1/2020 | 2:33 PM |
| 3/1/2020 | 2:43 PM |
| 3/1/2020 | 2:45 PM |
| 3/1/2020 | 2:50 PM |
| 3/1/2020 | 2:54 PM |
| 3/1/2020 | 2:56 PM |
| 3/1/2020 | 3:00 PM |
| 3/1/2020 | 3:03 PM |
| 3/1/2020 | 3:05 PM |
| 3/1/2020 | 3:10 PM |
| 3/1/2020 | 3:18 PM |
| 3/1/2020 | 3:35 PM |
| 3/1/2020 | 3:53 PM |
| 3/1/2020 | 3:54 PM |
| 3/1/2020 | 4:00 PM |
| 3/1/2020 | 4:00 PM |
| 3/1/2020 | 4:09 PM |
| 3/1/2020 | 4:20 PM |
| 3/1/2020 | 4:25 PM |
| 3/1/2020 | 4:30 PM |
| 3/1/2020 | 4:30 PM |
| 3/1/2020 | 4:40 PM |
| 3/1/2020 | 4:43 PM |
| 3/1/2020 | 4:50 PM |
| 3/1/2020 | 4:50 PM |
| 3/1/2020 | 4:58 PM |
| 3/1/2020 | 5:00 PM |
| 3/1/2020 | 5:12 PM |
| 3/1/2020 | 5:15 PM |
| 3/1/2020 | 5:25 PM |
| 3/1/2020 | 5:29 PM |
| 3/1/2020 | 5:30 PM |
| 3/1/2020 | 5:34 PM |
| 3/1/2020 | 6:00 PM |
| 3/1/2020 | 6:02 PM |
| 3/1/2020 | 6:04 PM |
| 3/1/2020 | 6:10 PM |
| 3/1/2020 | 6:10 PM |
| 3/1/2020 | 6:11 PM |
| 3/1/2020 | 6:20 PM |
| 3/1/2020 | 6:30 PM |
| 3/1/2020 | 6:39 PM |
| 3/1/2020 | 6:46 PM |
| 3/1/2020 | 7:00 PM |
| 3/1/2020 | 7:08 PM |
| 3/1/2020 | 7:10 PM |
| 3/1/2020 | 7:15 PM |
| 3/1/2020 | 7:20 PM |
| 3/1/2020 | 7:30 PM |
| 3/1/2020 | 7:35 PM |
| 3/1/2020 | 7:53 PM |
| 3/1/2020 | 8:00 PM |
| 3/1/2020 | 8:10 PM |
| 3/1/2020 | 8:16 PM |
| 3/1/2020 | 8:30 PM |
| 3/1/2020 | 8:30 PM |
| 3/1/2020 | 8:40 PM |
| 3/1/2020 | 9:13 PM |
| 3/1/2020 | 9:20 PM |
| 3/1/2020 | 10:00 PM |
| 3/1/2020 | 10:24 PM |
| 3/1/2020 | 10:35 PM |
| 3/1/2020 | 10:40 PM |
| 3/1/2020 | 10:40 PM |

Set 2 raw data

A picture containing table

Description automatically generated

Chi-Square Table from Excel for set 2