**PROJECT 2: STATISTICAL ANALYSIS**

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

TABLE OF CONTENTS:

|  |  |  |
| --- | --- | --- |
| Section Number | Section Name | Page Number |
| I | Cover Page | 2 |
| II | Chi Square goodness of fit-tests for both data sets | 3 - 5 |
| III | Appendices | 6 - 7 |

**COVER PAGE:**

NAME*:* GORREPATI SAI SURYA

UTA ID: 1002028664

EMAIL: [sxg8664@mavs.uta.edu](mailto:sxg8664@mavs.uta.edu)

Section: 002

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

**CHI SQUARE GOODNESS-OF-FIT TESTS:**

In this analysis Chi Square goodness of fit test will be performed on the data that was used in project part 1. The data set 1 is a sample of sepal width corresponding to iris flowers and a sample of time intervals between accidents occurring in the United States is used for data set 2.

The main aim of this project is to determine whether data set 1 follows normal distribution and data set 2 follows exponential distribution.

The data that was collected initially for project 1 is used for chi square analysis. Initially we have set up our classes in which our observed data falls for both the collected data sets.

For Dataset I: class intervals which were used for the first part of project were used in this analysis as the classes in which our random variable X lies. The observed frequency is the same as the sample frequency. The class probabilities are calculated by using the Norm.Dist command in Microsoft excel; it takes the values of current class, mean and standard deviation of sample data. Furthermore, we have to calculate the class frequency by multiplying class probabilities with the number of sample (n=130). Lastly the chi square class component is calculated by squaring the difference between class and expected frequency and this value is divided by expected frequency.

For Dataset 2: The initial two class intervals used in the first part of project were used as two classes and remaining all class intervals are merged into a single class here to ensure it has sufficient number of observations. The class probabilities are calculated by using the Gamma.Dist command in Microsoft excel; it takes the values of current class, alpha and beta of sample data. Furthermore, we have to calculate the class frequency by multiplying class probabilities with the number of samples (n=102). Lastly the chi square class component is calculated by squaring the difference between class and expected frequency and this value is divided by expected frequency.

To complete the goodness of fit test we have to compare the values of chi square test statistic which is a summation of chi square class components to the chi squared tabulated value with the given significance level of 0.05% and k-1 degrees of freedom (k is the number of classes used).

We reject the null hypothesis Ho when the chi square test statistic computed is greater than the value of chi square from table and conclude that the collected data sets does not follow the assumed distribution. Similarly, if chi square test statistic is less than the tabulated value we fail to reject to null hypothesis H0 Failing to reject the null hypothesis does not provide enough evidence to conclude the data sets follows assumed distribution.

**Data set I:**

Sample mean and sample standard deviation

Estimate for = = 3.057692308

Estimate for Y = Y = 0.440553807

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Class | Observed Frequency  (oi) | Class Probabilities (P) | Expected frequency  (ei)= P\*n | (oi-ei)^2 | Chi square class Components X^2= [(oi-ei)^2]/ei |
| X<= 2.25 | 4 | 0.033374877 | 4.33873402 | 0.114740736 | 0.026445672 |
| 2.25<X <=2.5 | 5 | 0.069401723 | 9.022224049 | 16.1782863 | 1.793159449 |
| 2.5<X <=2.75 | 19 | 0.13968049 | 18.15846374 | 0.708183284 | 0.039000176 |
| 2.75<X <=3.0 | 22 | 0.205449034 | 26.70837448 | 22.16879029 | 0.830031431 |
| 3.0<X <=3.25 | 42 | 0.220861962 | 28.71205506 | 176.5694807 | 6.14966363 |
| 3.25<X <=3.5 | 17 | 0.173538111 | 22.55995447 | 30.91309376 | 1.370264013 |
| 3.5<X <=3.75 | 10 | 0.099654568 | 12.95509382 | 8.732579467 | 0.674065321 |
| 3.75<X <=4 | 7 | 0.041817998 | 5.436339686 | 2.445033576 | 0.449757322 |
| 4.<X <=4.25 | 3 | 0.012820212 | 1.666627556 | 1.777882075 | 1.066754278 |
| X>4.25 | 1 | 0.003401024 | 0.442133117 | 0.311215459 | 0.703895381 |
| Total |  | =1 | 130 |  | 13.10303667 |

|  |  |
| --- | --- |
| Chi-Square Test Statistic | 13.10303667 |
| Chi Square (0.05,9) | 16.919 |

The required parameters for fit test are:

µ = 3.057692308

σ = 0.440553807

k = Number of classes =10

α = 0.05

degrees of freedom = k-1 = 9

Hypothesis testing of Data Set I:

We have to test whether Data set I follows a normal distribution. Therefore, we have to use the following null and alternative hypothesis

Let H0 Data set I follows normal distribution

H1 Data set I do not follow normal distribution

According to chi square goodness of fit test we reach the following conclusion: we fail to reject H0 because our chi square test statistic is less than tabular chi square value that has 5 % significance level and 9 degrees of freedoms.

**Data set 2:**

Sample mean

Estimate for = = 2944.068627

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Class | Observed Frequency  (oi) | Class Probabilities (P) | Expected frequency  (ei)= P\*n | (oi-ei)^2 | Chi square class Components X^2= [(oi-ei)^2]/ei |
| T<=2000 | 69 | 0.493044543 | 50.29054337 | 350.0437673 | 6.960429215 |
| 2000<T<=4000 | 20 | 0.249951622 | 25.4950654 | 30.1957438 | 1.184376009 |
| T>4000 | 13 | 0.257003907 | 26.21439848 | 174.6203272 | 6.661237234 |
| Total |  | =1 | 102 |  | 14.80604246 |

|  |  |
| --- | --- |
| Chi-Square Test Statistic | 14.80604246 |
| Chi Square (0.05,2) | 5.991 |

The required parameters for fit test are:

µ =2944.068627

k = Number of classes =3

α = 0.05

degrees of freedom = k-1 = 2

Hypothesis testing of Data Set II:

We have to test whether Data set II follows exponential distribution. Therefore, we have to use the following null and alternative hypothesis

Let H0 Data set I follows exponential distribution

H1 Data set I do not follow exponential distribution

According to chi square goodness of fit test we reach the following conclusion: reject the initial hypothesis H0 because our chi square test statistic is greater than tabular chi square value (14.806 > 5.991) that has 5 % significance level and 2 degrees of freedoms.

**APPENDIX:**

**Data set I:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Class | Observed Frequency  (oi) | Class Probabilities (P) | Expected frequency  (ei)= P\*n | (oi-ei)^2 | Chi square class Components X^2= [(oi-ei)^2]/ei |
| X<= 2.25 | 4 | 0.033374877 | 4.33873402 | 0.114740736 | 0.026445672 |
| 2.25<X <=2.5 | 5 | 0.069401723 | 9.022224049 | 16.1782863 | 1.793159449 |
| 2.5<X <=2.75 | 19 | 0.13968049 | 18.15846374 | 0.708183284 | 0.039000176 |
| 2.75<X <=3.0 | 22 | 0.205449034 | 26.70837448 | 22.16879029 | 0.830031431 |
| 3.0<X <=3.25 | 42 | 0.220861962 | 28.71205506 | 176.5694807 | 6.14966363 |
| 3.25<X <=3.5 | 17 | 0.173538111 | 22.55995447 | 30.91309376 | 1.370264013 |
| 3.5<X <=3.75 | 10 | 0.099654568 | 12.95509382 | 8.732579467 | 0.674065321 |
| 3.75<X <=4 | 7 | 0.041817998 | 5.436339686 | 2.445033576 | 0.449757322 |
| 4<X <=4.25 | 3 | 0.012820212 | 1.666627556 | 1.777882075 | 1.066754278 |
| X>4.25 | 1 | 0.003401024 | 0.442133117 | 0.311215459 | 0.703895381 |
| Total |  | =1 | 130 |  | 13.10303667 |

Class probability:

P (x<=2.25) = NORM.DIST(2.25,3.057692,0.440554,1)

P (2.25<X <=2.5) =NORM.DIST(2.5,3.057692,0.440554,1)- NORM.DIST(2.25,3.057692,0.440554,1)

P (2.5<X <=2.75) =NORM.DIST(2.75,3.057692,0.440554,1)- NORM.DIST(2.5,3.057692,0.440554,1)

P (2.75<X <=3.0) =NORM.DIST(3,3.057692,0.440554,1)- NORM.DIST(2.75,3.057692,0.440554,1)

P (3.0<X <=3.25) =NORM.DIST(3.25,3.057692,0.440554,1)- NORM.DIST(3,3.057692,0.440554,1)

P (3.25<X <=3.5) =NORM.DIST(3.5,3.057692,0.440554,1)- NORM.DIST(3.25,3.057692,0.440554,1)

P (3.5<X <=3.75) =NORM.DIST(3.75,3.057692,0.440554,1)- NORM.DIST(3.5,3.057692,0.440554,1)

P (3.75<X <=4) =NORM.DIST(4,3.057692,0.440554,1)- NORM.DIST(3.75,3.057692,0.440554,1)

P (4<X <=4.25) =NORM.DIST(4.25,3.057692,0.440554,1)- NORM.DIST(4,3.057692,0.440554,1)

P (X>4.25) =1- NORM.DIST(4.25,3.057692,0.440554,1)

Expected Frequency:

Ei = Pi \*130

For Class x<= 2.25 Ei = 0.033374877 \*130= 4.33873402

For Class 2.25 <X<=2.5 Ei =0.069401723 \*130=9.022224049

Similarly, the remaining values are obtained

Chi square class component:

For x<= 2.25 the value of chi square class component is

= = 0.026445672

For 2.25<X <=2.5 the value of chi square class component is

= = 1.793159449

Similarly, the remaining values are obtained

**Data set II:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Class | Observed Frequency  (oi) | Class Probabilities (P) | Expected frequency  (ei)= P\*n | (oi-ei)^2 | Chi square class Components X^2= [(oi-ei)^2]/ei |
| T<=2000 | 69 | 0.493044543 | 50.29054337 | 350.0437673 | 6.960429215 |
| 2000<T<=4000 | 20 | 0.249951622 | 25.4950654 | 30.1957438 | 1.184376009 |
| T>4000 | 13 | 0.257003907 | 26.21439848 | 174.6203272 | 6.661237234 |
| Total |  | =1 | 102 |  | 14.80604246 |

Class probability calculations:

P (T<=2000) = GAMMA.DIST(2000,1,2944.068,1)

P (2000< T<= 4000) =GAMMA.DIST(4000,1,2944.068,1)-GAMMA.DIST(2000,1,2944.068,1)

P (T>4000) = 1-GAMMA.DIST(4000,1,2944.0686,1)

Expected Frequency:

Ei = Pi \*n = Pi \*102

For Class T<= 2000; Ei = 0.493044543\*102= 50.29054337

For Class 2000 <X<=4000; Ei= 0.249951622 \* 102 = 25.4950654

For Class T >4000; Ei = 26.21439848 \*102 =26.21439848

Chi square class component:

For T<= 2000 the value of chi square class component is

== 6.960429215

For 2000<T <=4000 the value of chi square class component is

= = 1.184376009

For T> 4000 the value of chi square class component is

=26.21439848 = 6.661237234