

**AY2021/2022 April Semester**

**Quantitative Analysis (CIA2C12)**

**Project Report**

**Individual**

Submitted by: Edzran Hisham

|  |  |  |
| --- | --- | --- |
| Class | P03 | Signature |
| Admin No | 2004986B |  |
| Name | Edzran Hisham |
| Video Presentation link | <https://youtu.be/1b_JuneCh_w> |

“By submitting this work, I am declaring that I am the originator of this work and that all other original sources used in this work have been appropriately acknowledged.

I understand that plagiarism is the act of taking and using the whole or any part of another person’s work and presenting it as my own without proper acknowledgement.

I also understand that plagiarism is an academic offence, and that disciplinary action will be taken for plagiarism.”

Hi everyone, I’ll be working on my individual report. Let’s begin (:

**Task 1 – Chi-Square Test of goodness of fit**

In this individual project, I would be performing Chi Square tests on the given dataset. I shall start by downloading the dataset from LMS blackboard as shown below:

Graphical user interface, text, application, chat or text message

Description automatically generated

Opening the dataset I can verify that there are 3 columns and 2000 rows:

Column 1 – Running number from 1 to 2000

Column 2 – Weight range of the newborn

Column 3 – Coded to indicate baby gender, 1=male, 2=female.

Table

Description automatically generated with medium confidence

**Random number generator to randomly sample.**

I will now use the Random Number Generator in the Data Analysis toolpak to generate random numbers from the 2000 records. Below are the settings I used. For the output range, I selected 2000 rows in column D to be chosen to populate the random numbers.

Graphical user interface, application

Description automatically generated

After generating random numbers, I will now sort them in ascending order to randomize the sample data and then only take rows up to 400 as my final randomized dataset.

Graphical user interface, text, application

Description automatically generated

Finished. Here is the preview of the first 10 rows:

A picture containing text, electronics

Description automatically generated

**Transforming the dataset to prepare for Chi-Square tests**

Firstly, I have identified 2 variables for analysis, 1st variable – Newborn Weight Range, 2nd variable – Gender. There are 2 columns which are not useful in my analysis: column A – Running number, column D – Randomly generated numbers. These columns I will therefore drop.

After dropping the columns, I noticed that there are different ranges of weights. To identify easily identify the different ranges, I will use the method UNIQUE() in excel and highlight the entire column.



Here is the following results:

Text

Description automatically generated with medium confidence

**Creating frequency table of newborn weights from sampled data**

Before creating the frequency table, I will arrange the newborn weight range in ascending order and create a new column called “Observed Frequency” as such:

Table

Description automatically generated

After creating the new column, I will now populate the “Observed Frequency” column with the count of every unique range from the original data. Here is the formula I used:



This formula is repeated 12 times by changing the value of the logical test to the unique range I want to find for e.g. “>=0.3<0.7” to “>=0.7<1.1” and so on and so forth.

To verify that it is correct, I did a SUM() from the first row to the last row of values in “Observed Frequency”. The result is 400 which is correct as it tallies with the number of rows in the sampled dataset.

Here are the results:

Table

Description automatically generated

**Evaluating the claim that newborn weights follow a normal distribution using Chi Square goodness of fit test**

Step 1: Formulating null and alternative hypothesis

H0: The newborn weights **follow** a normal distribution

H1: The newborn weights **do not follow** a normal distribution

Step 2: Calculate mid-point of intervals

As my data values are in a range/intervals, and the mean and standard deviation is unknown, I would first have to calculate the midpoints first before calculating mean and standard deviation.

A new column is created called “Mid Point” which would be my x in this case. I then insert values which are the mid points of the respective range. For e.g. The range >=0.3<0.7 is 0.3 to 0.6. Therefore, the midpoint would be 0.45.

For the last range, I need to select an acceptable representative value and thus in my opinion I will take 4.7.

Table

Description automatically generated

Step 3: Calculate fx

Once I have my mid-points, I will multiply the observed frequency and mid-point into a new column “fx”:

Table

Description automatically generated

Step 4: Calculate mean and standard deviation

Finally, I can start to calculate my mean and standard deviation. I will calculate the sum of f which is n(400) and the sum of fx which is 994.35.

Now to get the mean I will take **sum of fx / sum of f**

Table

Description automatically generated

Timeline

Description automatically generated

The mean is = **2.485875**

To get the standard deviation, I will use the function =STDEV.S() and highlight the entire rows in column.

Table

Description automatically generated

Result:



The standard deviation is **0.77626555**

Step 5: Calculating normal probability

Firstly, I will create 2 columns “Range Start” and “Range End” that will be populated by the value at the start and end of each range category:

Table

Description automatically generated

I will then calculate Normal Probability by using the function =NORM.DIST() to calculate the probability of “Range End” minus the probability of “Range Start” value.



Result:

Table

Description automatically generated

Step 6: Calculating expected frequency

To calculate expected frequency, I take the Probability \* Sample Size(n=400):



Result:

Table

Description automatically generated

From the results, I can see that there are expected frequencies which are less than 5. Having expected frequencies that are too small would negate the validity of the chi-squared goodness of fit test.

Therefore, I will try to **combine** the **last 3 classes** and **first 2 classes**.

Table

Description automatically generated

Classes **“>=0.3<0.7**“was combined with **“>=0.7<1.1”** to make: **“>=0.3<1.1”**



The normal probability that was used to calculate “>=0.3<1.1” was similar by subtracting the probability of the new “Range Start” and “Range End” values of the class.

Classes **“>=4.3<4.7”** and **“>=4.7”** was combined with **“>=3.9<4.3”** to make: **“>=3.9”**

Table

Description automatically generated

The normal probability for this class was calculated by taking 1 subtracting the sum of the rest of the probability. This would allow us to give a total probability of = 1.

After collapsing our data, there are no expected frequencies that fall below 5. Thus, we are able to continue with our chi-square goodness of fit test.

Table

Description automatically generated

Step 7: Calculating Chi-Square test statistics

I can calculate the Chi-Square test statistics by first using the formula: (O-E)^2/E. Where O is observed frequency and E is expected frequency.

The formula we can use is:

(Observed– Expected)^2/Expected

(O-E)^/E



Applying the same formula to all classes and calculating the sum will give us the Chi-Square Test Statistics value as show:

Table

Description automatically generated



The Chi-Square Test Statistics is: **114.5613743**

Step 8: Calculating degree of freedom

Where there are m unknown parameters in the distribution or curve being fitted, the test statistic has approximately the chi-square distribution χ2(k–m–1).

Thus, when fitting data to a Poisson distribution m=1 (the mean parameter), while if fitting data to a normal distribution m=2 (the mean and standard deviation parameters).

I cannot use the function =CHISQ.TEST() to calculate my chi-square test statistics as df ≠ the sample size minus 1.

Since there were 2 estimated parameters(m) which are mean and standard deviation, my degree of freedom(df) formula would be:

**df = k – m – 1 = 9 – 2 – 1 = 5**

Step 9: Calculating Chi-Square Crit

Using the function =CHISQ.INV.RT(), I can calculate the Chi-Square Crit by using the previously calculated df and given significance level of 5%.

Timeline

Description automatically generated with medium confidence

Result:



Step 10: Conclusion

Since **114.5613743 (chisq test statistics) > 11.07049769 (chisq crit)**. I have evidence to **reject** the Null hypothesis at a significance level of 5%.

The newborn weights **do not** fit into an expected Normal Distribution.

Reference:

<https://www.real-statistics.com/chi-square-and-f-distributions/goodness-of-fit/>

<https://www.real-statistics.com/tests-normality-and-symmetry/statistical-tests-normality-symmetry/chi-square-test-for-normality/>

<https://www.real-statistics.com/descriptive-statistics/frequency-tables/>

<https://www.onlinemathlearning.com/mean-frequency-table-interval.html>

**Task 2 – Chi-Square Test of independence**

Step 1: Formulate Null and Alternative hypothesis

H0: Weight of newborn **is independent** of gender of the newborn

H1: Weight of newborn **is not independent** of gender of the newborn

Step 2: Create contingency table of Gender and Weight

I will first create an empty contingency table of Gender and Weight range:

Table

Description automatically generated

Step 3: Create new midpoint column in sample dataset

My method may be a little unorthodox but I was having difficulty using various functions and methods and this worked for me.

The method is just another way of distinguishing weight ranges by assigning specific values to specific weight ranges. As I was not able to find the count of both male and female for a specific weight range using the weight range table itself.

I first created a Vlookup table with the weight range and respective mid-point value to be assigned, essentially this midpoint value is a value to help act as a referral to the respective weight ranges:

A picture containing text, receipt, screenshot

Description automatically generated

Then inserting a new empty column into the sample dataset beside the weight range column, I used the function =VLOOKUP() to populate assign midpoint values next to the respective weight range column.

Text

Description automatically generated with medium confidence

Here is the end result of the first 12 rows:

Table

Description automatically generated

Step 4: Calculate count of Male and Female by specific weight range

Now that each weight range in the column is assigned a value to be distinguished,

using the =COUNTIFS() function, I pass in:

(1)Criteria range: C1:C400 (mid point column)

(1)Criteria: 0.45(respective mid point value)

(2)Criteria range: D1:D400

(2)Criteria: Either Male or Female

I can verify that the results is correct by doing a =SUM() to values in both Male and Female columns which adds up to 400.

Table

Description automatically generated

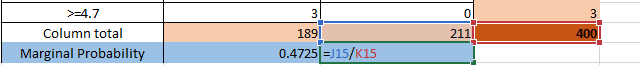
Step 5: Sum up totals for rows and columns

**Table

Description automatically generated**

Step 6: Calculate (marginal) probabilities

The formula to calculate marginal probabilities is Column Total/ Sample Total



The results:

Table

Description automatically generated

Step 7: Calculate expected observations. Ei = nP

The formula to calculate expected observations is Marignial Probability \* Row total:

The results are:

Table

Description automatically generated

Step 8: Calculating Chi-Square test statistics

The formula is: (Observed Frequencies – Expected Frequencies) ^2/Expected frequencies

Table

Description automatically generated

Step 9: Calculate test statistics

To calculate the test statistics, I will take the sum of all the test statistics in the table above

**Test statistics:** 7.278577678

**Degree of freedom:**

(r-1)\*(c-1) = (12-1)\*(2-1) = 11

**Level of significance:** 5% = 0.05

**Critical Value:**

CHISQ.INV.RT(0.05,11**) =** 19.67513757

Step 10: Conclusion

H0: Weight of newborn **is independent** of gender of the newborn

H1: Weight of newborn **is not independent** of gender of the newborn

Since 7.278577678(test stats) < 19.67513757(crit value), I will accept the null hypothesis at 5% level of significance. Thus, the weight of newborn is independent of gender of the newborn.