



School of Informatics & IT  
TEMASEK POLYTECHNIC

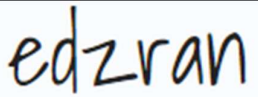
**AY2021/2022 April Semester**

**Quantitative Analysis (CIA2C12)**

**Project Report**

**Individual**

Submitted by: Edzran Hisham

Class	P03	Signature
Admin No	2004986B	
Name	Edzran Hisham	
Video Presentation link	<a href="https://youtu.be/1b_JuneCh_w">https://youtu.be/1b_JuneCh_w</a>	

“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:



**Project (60%)**

Attached Files:

- QUAS project 2021.pdf (331.729 KB)
- QUAS Project Cover page (group report).docx (76.624 KB)
- QUAS Project Cover page (individual report).docx (71.875 KB)
- QUAS project data 2021.xlsx (50.203 KB)

Project Reports (both Individual and group) due on 15 August 2359 hours.

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.

1998	1998	$\geq 2.3 < 2.7$	2
1999	1999	$\geq 2.7 < 3.1$	1
2000	2000	$\geq 1.5 < 1.9$	2

### **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.

Random Number Generation

Number of Variables:  OK

Number of Random Numbers:  Cancel

Distribution: Uniform Help

Parameters

Between  and

Random Seed:

Output options

☒ Output Range:

☐ New Worksheet Ply:

☐ New Workbook

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.

Sort

+ Add Level - Delete Level Copy Level ^ v Options... ☐ My data has headers

Column	Sort On	Order
Sort by Column D	Cell Values	Smallest to Largest

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

1	607	>=3.1 <3.5	2	0.000427
2	1971	>=1.9 <2.3	2	0.000793
3	540	>=2.7 <3.1	1	0.00116
4	102	>=1.5 <1.9	2	0.001343
5	34	>=4.3 <4.7	2	0.002625
6	1957	>=1.9 <2.3	2	0.002899
7	946	>=0.7 <1.1	1	0.003296
8	968	>=2.7 <3.1	2	0.003571
9	676	>=2.3 <2.7	1	0.003601
10	1826	>=3.1 <3.5	2	0.003601

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

Firstly, I have identified 2 variables for analysis, 1<sup>st</sup> variable – Newborn Weight Range, 2<sup>nd</sup> 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.

=UNIQUE(B1:B400)

Here is the following results:

>=3.1<3.5
>=1.9<2.3
>=2.7<3.1
>=1.5<1.9
>=4.3<4.7
>=0.7<1.1
>=2.3<2.7
>=1.1<1.5
>=3.5<3.9
>=3.9<4.3
>=0.3<0.7
>=4.7

### 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:

Newborn Weight Range	Observed Frequency
>=0.3<0.7	
>=0.7<1.1	
>=1.1<1.5	
>=1.5<1.9	
>=1.9<2.3	
>=2.3<2.7	
>=2.7<3.1	
>=3.1<3.5	
>=3.5<3.9	
>=3.9<4.3	
>=4.3<4.7	
>=4.7	

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:

Observed Frequency

=SUM(IF(\$B\$1:\$B\$400=">=0.3<0.7",1,0))

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:

Newborn Weight Range	Observed Frequency
$\geq 0.3 < 0.7$	4
$\geq 0.7 < 1.1$	9
$\geq 1.1 < 1.5$	21
$\geq 1.5 < 1.9$	46
$\geq 1.9 < 2.3$	69
$\geq 2.3 < 2.7$	86
$\geq 2.7 < 3.1$	80
$\geq 3.1 < 3.5$	44
$\geq 3.5 < 3.9$	25
$\geq 3.9 < 4.3$	9
$\geq 4.3 < 4.7$	4
$\geq 4.7$	3
Total	400

**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  $\geq 0.3 < 0.7$  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.

Newborn Weight Range	Observed Frequency(f)	x = Mid Point
$\geq 0.3 < 0.7$	4	0.45
$\geq 0.7 < 1.1$	9	0.85
$\geq 1.1 < 1.5$	21	1.25
$\geq 1.5 < 1.9$	46	1.65
$\geq 1.9 < 2.3$	69	2.05
$\geq 2.3 < 2.7$	86	2.45
$\geq 2.7 < 3.1$	80	2.85
$\geq 3.1 < 3.5$	44	3.25
$\geq 3.5 < 3.9$	25	3.65
$\geq 3.9 < 4.3$	9	4.05
$\geq 4.3 < 4.7$	4	4.45
$\geq 4.7$	3	4.7

Step 3: Calculate fx

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

Newborn Weight Range	Observed Frequency(f)	x = Mid Point	fx
>=0.3<0.7	4	0.45	1.8
>=0.7<1.1	9	0.85	7.65
>=1.1<1.5	21	1.25	26.25
>=1.5<1.9	46	1.65	75.9
>=1.9<2.3	69	2.05	141.45
>=2.3<2.7	86	2.45	210.7
>=2.7<3.1	80	2.85	228
>=3.1<3.5	44	3.25	143
>=3.5<3.9	25	3.65	91.25
>=3.9<4.3	9	4.05	36.45
>=4.3<4.7	4	4.45	17.8
>=4.7	3	4.7	14.1

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**

$$\text{Mean} = \frac{\text{sum of } fx}{\text{sum of } f}$$

n(sum of f)	400
sum of fx	994.35
mean	=146/145

The mean is = **2.485875**

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

607	>=3.1<3.5	3.25	2	0.000427	=STDEV.S(C:C
1971	>=1.9<2.3	2.05	2	0.000793	STDEV.S(number1, [
540	>=2.7<3.1	2.85	1	0.00116	
102	>=1.5<1.9	1.65	2	0.001343	
34	>=4.3<4.7	4.45	2	0.002625	
1957	>=1.9<2.3	2.05	2	0.002899	

Result:

**0.77626555**

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:

Newborn	Observed Frequency(f)	Range Start	Range End
>=0.3<0.7	4	0.3	0.6
>=0.7<1.1	9	0.7	1
>=1.1<1.5	21	1.1	1.4
>=1.5<1.9	46	1.5	1.8
>=1.9<2.3	69	1.9	2.2
>=2.3<2.7	86	2.3	2.6
>=2.7<3.1	80	2.7	3
>=3.1<3.5	44	3.1	3.4
>=3.5<3.9	25	3.5	3.8
>=3.9<4.3	9	3.9	4.2
>=4.3<4.7	4	4.3	4.6
>=4.7	3	4.7	

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.

Newborn	Observed Frequency(f)	range start	range end	normal probabilitiy	expected
>=0.3<0.7	4	0.3	0.6	=NORM.DIST(M11,\$K\$26,\$K\$27,TRUE)-NORM.DIST(L11,\$K\$26,\$K\$27,TRUE)	

Result:

Newborn	Observed Frequency(f)	range start	range end	normal probabilitiy
>=0.3<0.7	4	0.3	0.6	0.005129353
>=0.7<1.1	9	0.7	1	0.017094443
>=1.1<1.5	21	1.1	1.4	0.043824283
>=1.5<1.9	46	1.5	1.8	0.086429409
>=1.9<2.3	69	1.9	2.2	0.131132195
>=2.3<2.7	86	2.3	2.6	0.153061866
>=2.7<3.1	80	2.7	3	0.137447821
>=3.1<3.5	44	3.1	3.4	0.094955386
>=3.5<3.9	25	3.5	3.8	0.05046664
>=3.9<4.3	9	3.9	4.2	0.020633788
>=4.3<4.7	4	4.3	4.6	0.006489704
>=4.7	3	4.7		0.002170357

Step 6: Calculating expected frequency

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

normal probabilitiy	Expected Frequencies
0.005129353	=N11*400

Result:



Newborn	Observed Frequency(f)	range start	range end	normal probabilitiy	Expected Frequencies
>=0.3<0.7	4	0.3	0.6	0.005129353	2.051741347
>=0.7<1.1	9	0.7	1	0.017094443	6.837777025
>=1.1<1.5	21	1.1	1.4	0.043824283	17.52971302
>=1.5<1.9	46	1.5	1.8	0.086429409	34.57176368
>=1.9<2.3	69	1.9	2.2	0.131132195	52.45287795
>=2.3<2.7	86	2.3	2.6	0.153061866	61.22474629
>=2.7<3.1	80	2.7	3	0.137447821	54.97912839
>=3.1<3.5	44	3.1	3.4	0.094955386	37.9821544
>=3.5<3.9	25	3.5	3.8	0.05046664	20.18665594
>=3.9<4.3	9	3.9	4.2	0.020633788	8.253515224
>=4.3<4.7	4	4.3	4.6	0.006489704	2.595881752
>=4.7	3	4.7		0.002170357	0.868142942

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**.

Newborn Weight Range
>=0.3<1.1
>=1.1<1.5
>=1.5<1.9
>=1.9<2.3
>=2.3<2.7
>=2.7<3.1
>=3.1<3.5
>=3.5<3.9
>=3.9

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

Newborn Weight Range	Observed Frequency(f)	range start	range end	normal probabilitiy	expected	chi square test stats
>=0.3<1.1	9	0.3	1	=NORM.DIST(M30,\$K\$26,\$K\$27,TRUE)-NORM.DIST(L30,\$K\$26,\$K\$27,TRUE)		

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”

Newborn Weight Range	Observed Frequency(f)	range start	range end	normal probabilitiy
>=0.3<1.1	9	0.3	1	0.025369475
>=1.1<1.5	21	1.1	1.4	0.043824283
>=1.5<1.9	46	1.5	1.8	0.086429409
>=1.9<2.3	69	1.9	2.2	0.131132195
>=2.3<2.7	86	2.3	2.6	0.153061866
>=2.7<3.1	80	2.7	3	0.137447821
>=3.1<3.5	44	3.1	3.4	0.094955386
>=3.5<3.9	25	3.5	3.8	0.05046664
>=3.9	16	3.9		=1-SUM(N30:N37)



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.

Newborn Weight Range	Observed Frequency(f)	range start	range end	normal probability	expected
>=0.3<1.1	9	0.3	1	0.025369475	10.14778984
>=1.1<1.5	21	1.1	1.4	0.043824283	17.52971302
>=1.5<1.9	46	1.5	1.8	0.086429409	34.57176368
>=1.9<2.3	69	1.9	2.2	0.131132195	52.45287795
>=2.3<2.7	86	2.3	2.6	0.153061866	61.22474629
>=2.7<3.1	80	2.7	3	0.137447821	54.97912839
>=3.1<3.5	44	3.1	3.4	0.094955386	37.9821544
>=3.5<3.9	25	3.5	3.8	0.05046664	20.18665594
>=3.9	16	3.9		0.277312926	110.9251705

#### 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:

$$\frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

$$(O-E)^2/E$$

Newborn Weight Range	Observed Frequency(f)	range start	range end	normal probability	expected	chi square test stats
>=0.3<1.1	9	0.3	1	0.025369475	10.14778984	$(K30-O30)^2/O30$

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

chi square test stats
0.129823492
0.686998796
3.777781967
5.220061491
10.02557354
11.38693964
0.953460019
1.147702775
81.23303262
=SUM(P30:P38)
114.5613743

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  $\chi^2(k-m-1)$ .

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  $df \neq$  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%.

degree of freedom	5
alpha	0.05
chisq crit	=CHISQ.INV.RT(K44,K43)

Result:

chisq crit	11.07049769
------------	-------------

#### 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:

Newborn Weight Range	Gender	
	Male(1)	Female(2)
>=0.3<0.7		
>=0.7<1.1		
>=1.1<1.5		
>=1.5<1.9		
>=1.9<2.3		
>=2.3<2.7		
>=2.7<3.1		
>=3.1 <3.5		
>=3.5<3.9		
>=3.9<4.3		
>=4.3<4.7		
>=4.7		

### 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:

Vlookup table	
>=0.3<0.7	0.45
>=0.7<1.1	0.85
>=1.1<1.5	1.25
>=1.5<1.9	1.65
>=1.9<2.3	2.05
>=2.3<2.7	2.45
>=2.7<3.1	2.85
>=3.1 <3.5	3.25
>=3.5<3.9	3.65
>=3.9<4.3	4.05
>=4.3<4.7	4.45
>=4.7	4.7

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.

$\geq 3.1 < 3.5$	=VLOOKUP(B1,\$H\$17:\$I\$28,2,)
$\geq 1.9 < 2.3$	VLOOKUP(lookup_value, table_array, col_index_num, [range_lookup])

Here is the end result of the first 12 rows:

$\geq 3.1 < 3.5$	3.25
$\geq 1.9 < 2.3$	2.05
$\geq 2.7 < 3.1$	2.85
$\geq 1.5 < 1.9$	1.65
$\geq 4.3 < 4.7$	4.45
$\geq 1.9 < 2.3$	2.05
$\geq 0.7 < 1.1$	0.85
$\geq 2.7 < 3.1$	2.85
$\geq 2.3 < 2.7$	2.45
$\geq 3.1 < 3.5$	3.25
$\geq 2.7 < 3.1$	2.85
$\geq 2.7 < 3.1$	2.85

#### 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.

						Gender	
						Male(1)	Female(2)
607	$\geq 3.1 < 3.5$	3.25	2	0.000427			
1971	$\geq 1.9 < 2.3$	2.05	2	0.000793			
540	$\geq 2.7 < 3.1$	2.85	1	0.00116			
102	$\geq 1.5 < 1.9$	1.65	2	0.001343			
34	$\geq 4.3 < 4.7$	4.45	2	0.002625			
1957	$\geq 1.9 < 2.3$	2.05	2	0.002899			
946	$\geq 0.7 < 1.1$	0.85	1	0.003296			
968	$\geq 2.7 < 3.1$	2.85	2	0.003571			
676	$\geq 2.3 < 2.7$	2.45	1	0.003601			
1826	$\geq 3.1 < 3.5$	3.25	2	0.003601			
351	$\geq 2.7 < 3.1$	2.85	2	0.00528			
1125	$\geq 2.7 < 3.1$	2.85	1	0.006928			
846	$\geq 3.1 < 3.5$	3.25	2	0.007385			
890	$\geq 1.9 < 2.3$	2.05	2	0.007874			
1121	$\geq 2.7 < 3.1$	2.85	1	0.008026			
532	$\geq 2.7 < 3.1$	2.85	1	0.008484			
1019	$\geq 2.3 < 2.7$	2.45	1	0.0094			
882	$\geq 2.3 < 2.7$	2.45	2	0.009491			
694	$\geq 2.3 < 2.7$	2.45	1	0.010865			
957	$\geq 1.1 < 1.5$	1.25	1	0.010987			
972	$\geq 1.9 < 2.3$	2.05	2	0.012329			
704	$\geq 2.3 < 2.7$	2.45	1	0.013916			
1001	$\geq 2.7 < 3.1$	2.85	1	0.014374			
1211	$\geq 1.5 < 1.9$	1.65	2	0.014801			
1797	$\geq 3.5 < 3.9$	3.65	2	0.015168			
1598	$\geq 2.7 < 3.1$	2.85	1	0.017273			
962	$\geq 1.9 < 2.3$	2.05	2	0.018311			
28	$\geq 2.3 < 2.7$	2.45	2	0.019105			
						Gender	
						Male(1)	Female(2)
						=COUNTIFS(\$C\$1:\$C\$400,17,\$D\$1:\$D\$400,1)	
						COUNTIFS(criteria_range1, criteria1, [criteria_range2, criteria2], [criteria_range3, ...])	
						8	13
						18	28
						33	36
						40	46
						38	42
						23	21
						13	12
						5	4
						2	2
						3	0
						Total	400
						Vlookup table	
						$\geq 0.3 < 0.7$	0.45
						$\geq 0.7 < 1.1$	0.85
						$\geq 1.1 < 1.5$	1.25
						$\geq 1.5 < 1.9$	1.65
						$\geq 1.9 < 2.3$	2.05
						$\geq 2.3 < 2.7$	2.45
						$\geq 2.7 < 3.1$	2.85
						$\geq 3.1 < 3.5$	3.25
						$\geq 3.5 < 3.9$	3.65
						$\geq 3.9 < 4.3$	4.05
						$\geq 4.3 < 4.7$	4.45
						$\geq 4.7$	4.7

#### Step 5: Sum up totals for rows and columns

H	I	J	K
	Gender		
Newborn Weight Range	Male(1)	Female(2)	Row Total
>=0.3<0.7	1	3	4
>=0.7<1.1	5	4	9
>=1.1<1.5	8	13	21
>=1.5<1.9	18	28	46
>=1.9<2.3	33	36	69
>=2.3<2.7	40	46	86
>=2.7<3.1	38	42	80
>=3.1<3.5	23	21	44
>=3.5<3.9	13	12	25
>=3.9<4.3	5	4	9
>=4.3<4.7	2	2	4
>=4.7	3	0	3
Column total	189	211	400

Step 6: Calculate (marginal) probabilities

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

>=4.7	3	0	3
Column total	189	211	400
Marginal Probability	0.4725	=J15/K15	

The results:

>=4.7	3	0	3
Column total	189	211	400
Marginal Probability	0.4725	0.5275	

Step 7: Calculate expected observations.  $E_i = nP$

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

The results are:

	Gender		
Newborn Weight Range	Male(1)	Female(2)	Row Total
>=0.3<0.7	1	3	4
>=0.7<1.1	5	4	9
>=1.1<1.5	8	13	21
>=1.5<1.9	18	28	46
>=1.9<2.3	33	36	69
>=2.3<2.7	40	46	86
>=2.7<3.1	38	42	80
>=3.1<3.5	23	21	44
>=3.5<3.9	13	12	25
>=3.9<4.3	5	4	9
>=4.3<4.7	2	2	4
>=4.7	3	0	3
Column total	189	211	400
Marginal Probability	0.4725	0.5275	
Gender, Expected Observations (E)			
Newborn Weight Range	Male(1)	Female(2)	
>=0.3<0.7	1.89	2.11	
>=0.7<1.1	4.2525	4.7475	
>=1.1<1.5	9.9225	11.0775	
>=1.5<1.9	21.735	24.265	
>=1.9<2.3	32.6025	36.3975	
>=2.3<2.7	40.635	45.365	
>=2.7<3.1	37.8	42.2	
>=3.1<3.5	20.79	23.21	
>=3.5<3.9	11.8125	13.1875	
>=3.9<4.3	4.2525	4.7475	
>=4.3<4.7	1.89	2.11	
>=4.7	1.4175	1.5825	

#### Step 8: Calculating Chi-Square test statistics

The formula is:  $(\text{Observed Frequencies} - \text{Expected Frequencies})^2 / \text{Expected frequencies}$

	Calculate $(O-E)^2/E$	
	Gender	
Newborn Weight Range	Male(1)	Female(2)
>=0.3<0.7	0.419100529	0.375402844
>=0.7<1.1	0.131394768	0.117694839
>=1.1<1.5	0.372487402	0.333649853
>=1.5<1.9	0.641832298	0.574911395
>=1.9<2.3	0.004846446	0.004341129
>=2.3<2.7	0.009923096	0.00888846
>=2.7<3.1	0.001058201	0.000947867
>=3.1<3.5	0.234925445	0.210430849
>=3.5<3.9	0.119378307	0.10693128
>=3.9<4.3	0.131394768	0.117694839
>=4.3<4.7	0.006402116	0.005734597
>=4.7	1.766706349	1.5825

#### 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:**

$$\text{CHISQ.INV.RT}(0.05,11) = 19.67513757$$

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.