CHI-SQUARED ANALYSIS

ISMAIL SEVER

Table of Content

- Description of Analysis
- Summary of Insights
- Two Combinations with Explanation
- R script

Description of Analysis

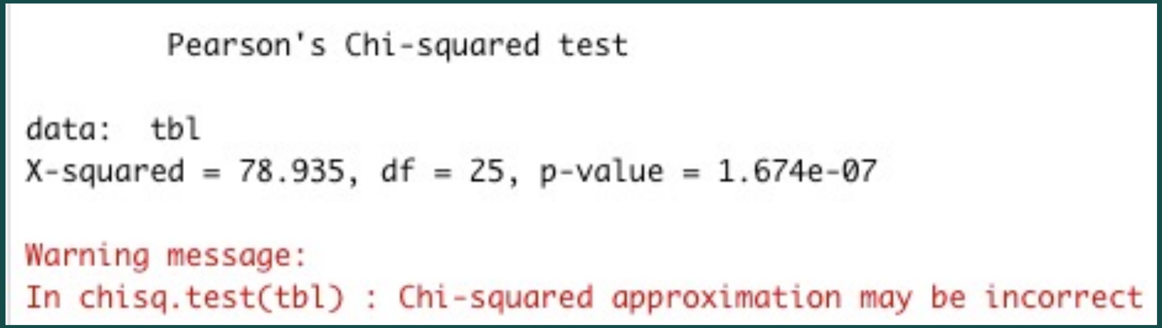
- Chi-square test is a statistical method to determine if two **categorical variables** have a significant **correlation** between them.
- We have a dataset called Carsdatabase. We will import this dataset into our RStudio and conduct Chi-square test to check whether there is any relation between our two categorical variables Number of cylinders (3, 4, 5, 6, 8, and rotary) and type of car sold (Compact, Large, Midsize, Small, Sporty, and Van). Based on the findings, we will have some insights for study.
- Null Hypothesis: Number of cylinders and type of car sold are independent
- Alternative Hypothesis: Number of cylinders and type of car sold are related.
- Our significance level is 0.05 for this test.

Summary of Insights

Chi-square value is 78.935. Higher is chi-square value means lower is the p-value.

The degree of freedom is 25 which is calculated by multiplying (6-1) and (6-1). 25 numbers in the grid are independent.

Since p-value is 1.674e-07 which is smaller than our significance value 0.05, we reject the null hypothesis that Number of cylinders and type of car sold are independent. As we see that Number of cylinders and type of car sold are highly correlated(dependent).



```
        Pearson's Chi-squared test

data:  tbl
X-squared = 78.935, df = 25, p-value = 1.674e-07

Warning message:
In chisq.test(tbl) : Chi-squared approximation may be incorrect
```

As you see, we got a warning message. This is due to the numbers in our grid, which are less than 5. I tried combining columns to get rid of this warning, but I couldn’t be successful. If we have more data, we would be able to get a result without warning hopefully. See the next page please.

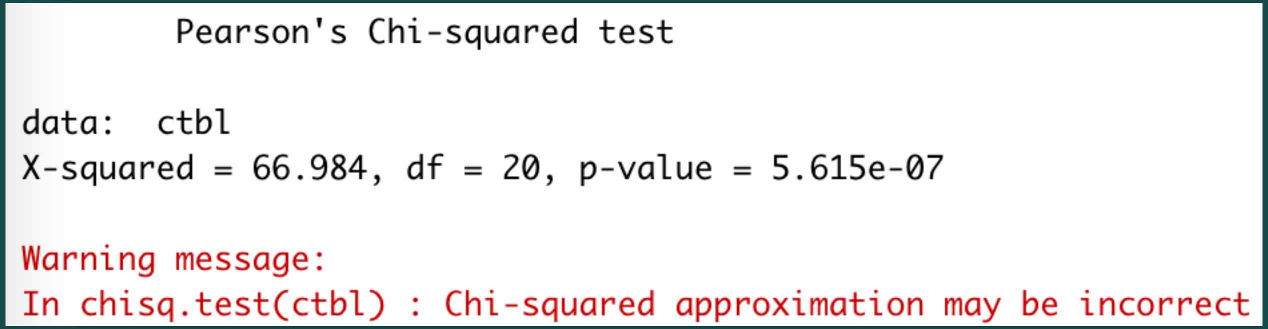
Two Combinations

1. Firstly, I would like to choose another two categorical variables such as Drive Train and Airbags to check whether there is any relations between them. Drive Train has categorical variables such as Front, Rear, 4WD, on the other hand Airbags has None, Driver only, Driver and Passenger.

2. Secondly, I would like to choose another two categorical variables such as Manual Transmission Available and Passengers to check whether there is any relations between them. Manual Transmission Available has categorical variables such as Yes/No, on the other hand Passengers has 2, 4, 5, 6, 7. It might be numbers but since Passengers is not infinite, we consider it as categorical variables.

R Script (Combining Columns)

- #New Table
- ctbl = cbind(tbl[,"Compact"],tbl[,"Midsize"],tbl[,"Small"],tbl[,"Sporty"], tbl[,"Large"] + tbl[,"Van"])
- ctbl
- #Chi-squared New Table
- chisq.test(ctbl)

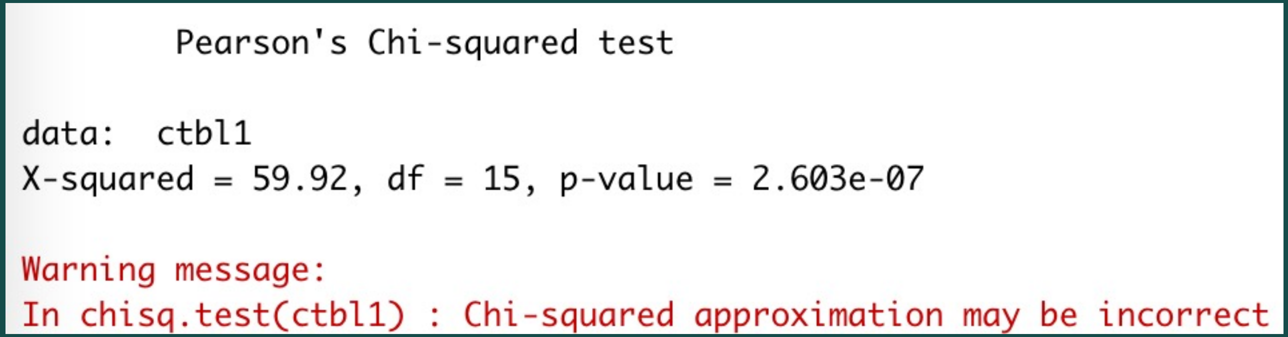


```
	Pearson's Chi-squared test

data:  ctbl
X-squared = 66.984, df = 20, p-value = 5.615e-07

Warning message:
In chisq.test(ctbl) : Chi-squared approximation may be incorrect
```

- #New Table1
- ctbl1 = cbind(tbl[,"Midsize"],tbl[,"Small"],tbl[,"Compact"]+tbl[,"Sporty"], tbl[,"Large"] + tbl[,"Van"])
- ctbl1
- #Chi-squared New Table
- chisq.test(ctbl1)



```
	Pearson's Chi-squared test

data:  ctbl1
X-squared = 59.92, df = 15, p-value = 2.603e-07

Warning message:
In chisq.test(ctbl1) : Chi-squared approximation may be incorrect
```

Thank You