
# UNIVERSITI TEKNOLOGI MALAYSIA

## SCHOOL OF COMPUTING

## FACULTY OF ENGINEERING

## PROBABILITY AND STATISTICAL DATA ANALYSIS

## (SECI2143-04)


| No. | Name | No. Matric |
|---|---|---|
| 1. | MUHAMMAD ABDUL AZIM BIN MISDAN | A20EC0081 |
| 2. | NOR FADLI BIN AHMAD | A20EC0109 |
| 3. | NUR FATEHAH BINTI HAIROL NIZAM | A20EC0113 |
| 4. | MUQRISYA BINTI MAT NOOR | A20EC0218 |

## TABLE OF CONTENTS

## 1.0 INTRODUCTION

Cars are the most common transport for everyone. It also has become essential needs to anyone these day since having a car will help them in doing some errand or work. For a long-term use, consumer should know more about the car quality and ability. Some of people would desire a very sporty car like Ferrari and some would desire a normal but versatile model such as Honda CR-V. According to Diana (7 January,2017), a car means more than a machine, but it also symbolized of freedom, status and it also of someone wealth.

So, in this project 2, our group purpose is to identify different types of cars with 5 different characteristics such as MPG, displacement, horsepower, weight, and acceleration. We use these four tests on our project to analyse the result. The tests are hypothesis testing on 1 sample test, correlation, regression, chi square and Anova. For the hypothesis testing on 1 sample test, it is used to determine whether unknown population mean is different from specific value.

Correlation is defined as a measure of the statistical relationship between two comparable variables or quantities (bivariate data). The values are positive when it values increase together meanwhile it will positive when one value decrease as the other. Regression is used to analyse the value of a dependent variable based on the value of at least one independent variable. Chi-square ($X^2$) statistic is a test that measures how a model compares to actual observed data. It usually often used in hypothesis testing.

## 2.0 DATASET

This dataset was taken from website https://perso.telecom-paristech.fr/eagan/class/igr204/datasets which from trusted website that have provided a lot of dataset for students doing their project and research regarding to statistic and data analysis. For this dataset, it has 9 column and 408 rows. The 408 rows are the type of car that have been analysed before this. For the column 2 in dataset is MPG (miles per gallon), third column is cylinders, fourth column is displacement (engine capacity), fifth column is horsepower, sixth column is weight, seventh column is acceleration, column eight is model and the last column is the origin of the car.

We only take 25 types of cars and 5 variables inside the dataset. The 5 variables that we take are MPG (meter per gallon), displacement (engine capacity), horsepower, weight, and acceleration. Inside the car variable has 25 different types of car like Chevrolet Chevelle Malibu, Buick Skylark 320, Plymouth Satellite,AMC Rebel SST, Ford Torino, Ford Galaxie 500,Chevrolet Impala, Plymouth Fury iii, Pontiac Catalina, AMC Ambassador DPL,Dodge Challenger SE,Plymouth 'Cuda 340,Chevrolet Monte Carlo,Buick Estate Wagon (sw),Toyota Corolla Mark ii, Plymouth Duster, AMC Hornet, Ford Maverick, Datsun PL510,Peugeot 504,AMC Gremlin, Ford F250,Chevy C20,Dodge D200 and Hi 1200D.

## 3.0 DATA ANALYSIS

### 3.1 Hypothesis Testing on 1 Sample Test

We made a hypothesis statement where we claim that the average of the cars' acceleration is 10 m/s$^2$. To determine whether this statement is to be rejected or accepted, we need to use hypothesis testing on 1 sample test using R Studio.

Hypothesis statement,
H0: μ = 10 (The mean of the acceleration of the cars is 10 m/s$^2$)
H1: μ =/= 10 (The mean of the acceleration of the cars is different from 10 m/s$^2$)

We use one sample t-test because the sample size is less than 50 and the population variance is unknown.
By using the formula of t-test,

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Where,

t = t-test score
$\bar{x}$ = sample mean
μ = population mean
s = standard deviation
n = number of observations
We use **x** as the variable of the acceleration of the cars.

```
> mean(x)
[1] 12.42
> sd(x)
[1] 3.050546
> t.test(x, mu=10)
```

Figure 1.1

From the result in Figure 1.1, the mean for the acceleration is 12.42 m/s$^2$. The standard deviation is 3.051.

```
        One Sample t-test

data:  x
t = 3.9665, df = 24, p-value = 0.0005734
alternative hypothesis: true mean is not equal to 10
95 percent confidence interval:
 11.1608 13.6792
sample estimates:
mean of x
    12.42
```

Figure 1.2

Based on Figure 1.2,
t-test value = 3.9665
Degree of freedom = 24
p-value = 0.0005
Confidence interval = [11.1608, 13.6792]

To find the critical value,

```
> qt(p=0.05/2, df=24, lower.tail=FALSE)
[1] 2.063899
```

Thus,

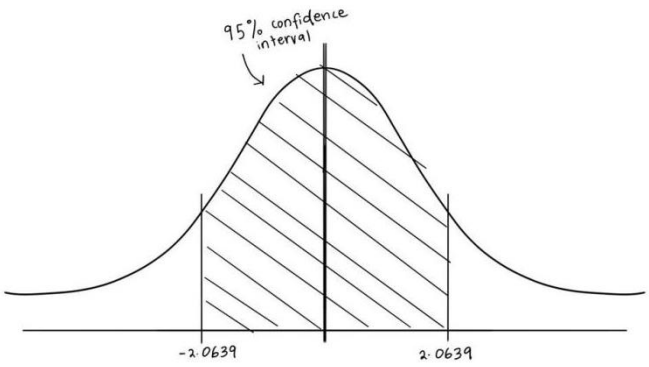


Figure 1.3

The t-test value is 3.9665, which is greater than the critical value of the right side, 2.0639. Plus, the t-test value is in the rejection region. Thus, it rejects the null hypothesis because the p-value is lower than the significance level, 0.0005<0.05 and the t-test value is greater than the critical value, 3.9665>2.0639. In short, we can conclude that we have sufficient evidence to the true mean is not 10 $m/s^2$.

### 3.2 Correlation

In correlation data analysis, we want to find the linear correlation between the horsepower and the acceleration of the cars.

To determine the correlation, use formula:

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

where,

$r_{xy}$ = correlation coefficient of the linear relationship between the x and y
$x_i$ = values of the x variable
$\bar{x}$ = mean of x variable
$y_i$ = values of the x variable
$\bar{y}$ = mean of x variable

To find the coleration coefficient, we use R Studio,

Hypothesis statement:

H0 : p = 0 ( There is no linear correlation between two variables)
H1 : p =/= 0 ( There is a linear correlation between two variables)

Variable declaration:

x = horsepower of cars
y = acceleration of cars

```
>   View(correlation)
> x=correlation$V1
> y=correlation$V2
> cor.test(x, y)

        Pearson's product-moment correlation

data:  x and y
t = -3.0207, df = 23, p-value = 0.006087
alternative hypothesis: true correlation is not
equal to 0
95 percent confidence interval:
 -0.7666396 -0.1745867
sample estimates:
       cor
-0.5329521
```

Figure 2.1

Based on Figure 2.1,
t-test value = -3.0207
Degree of freedom = 23
p-value = 0.0061
Confidence interval=[-0.7666, -0.1746]
Correlation coefficient = -0.533

Based on the result given, the p-value of the test is 0.0061 which is less than the significance level apha=0.05, 0.0061<0.05.

By using Pearson's Correlation Table with significance level of 0.05 and the degree of freedom 23, the critical value is [-0.3961, 0.3961].

Since the correlation coefficient is -0.533, it is lesser than the critical value, -0.533<-0.396 thus we fail to reject the null hypothesis.

There is no sufficient evidence of a linear correlation between horsepower and acceleration of the cars at the 0.05 significance level.

In Figure 2.2, the scatter plot shows the relation between two variables as stated. It was created by the R studio.
x = Horsepower of the cars
y = Acceleration of the cars

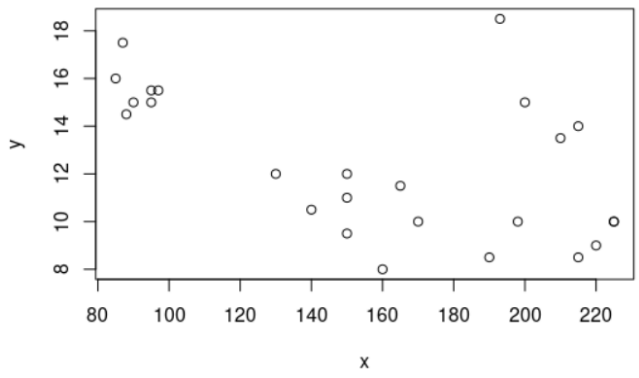


Figure 2.2
This graph plot indicates that there is no linear correlation between the two variables.

### 3.3 Regression

Linear regression is a measure for the relationship between an independent variable and dependent variable. For this case, we examined the relationship between Horsepower and Acceleration of a car. ($\alpha = 0.05$)

Hypothesis statement:
H0: $\boldsymbol{\beta}_1 = 0$ (no linear relationship)
H1: $\boldsymbol{\beta}_1 \neq 0$ (linear relationship does exist)

| Horsepower (x) | Acceleration (y) | xy | $x^2$ |
|---|---|---|---|
| 130 | 12 | 1560 | 16900 |
| 165 | 11.5 | 1897.5 | 27225 |
| 150 | 11 | 1650 | 22500 |
| 150 | 12 | 1800 | 22500 |
| 140 | 10.5 | 1470 | 19600 |
| 198 | 10 | 1980 | 39204 |
| 220 | 9 | 1980 | 48400 |
| 215 | 8.5 | 1827.5 | 46225 |
| 225 | 10 | 2250 | 50625 |
| 190 | 8.5 | 1615 | 36100 |
| 170 | 10 | 1700 | 28900 |
| 160 | 8 | 1280 | 25600 |
| 150 | 9.5 | 1425 | 22500 |
| 225 | 10 | 2250 | 50625 |
| 95 | 15 | 1425 | 9025 |
| 95 | 15.5 | 1472.5 | 9025 |
| 97 | 15.5 | 1503.5 | 9409 |
| 85 | 16 | 1360 | 7225 |
| 88 | 14.5 | 1276 | 7744 |
| 87 | 17.5 | 1522.5 | 7569 |

| | | | |
|---|---|---|---|
| 90 | 15 | 1350 | 8100 |
| 215 | 14 | 3010 | 46225 |
| 200 | 15 | 3000 | 40000 |
| 210 | 13.5 | 2835 | 44100 |
| 193 | 18.5 | 3570.5 | 37249 |
| $\sum x = 3943$ | $\sum y = 310.5$ | $\sum xy = 47010$ | $\sum x^2 = 682575$ |

```
> model = lm(cars$`Acceleration` ~ cars$`Horsepower`)
> summary(model)

Call:
lm(formula = cars$Acceleration ~ cars$Horsepower)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3463 -1.6696 -0.2447  1.1168  7.2207

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     17.51938    1.76860   9.906 9.14e-10 ***
cars$Horsepower -0.03233    0.01070  -3.021  0.00609 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.637 on 23 degrees of freedom
Multiple R-squared:  0.284,     Adjusted R-squared:  0.2529
F-statistic: 9.125 on 1 and 23 DF,  p-value: 0.006087
```

Figure 3.1

Based on figure 3.1, output from software R:

$b_1 = -0.0323$ $S_{b1} = 0.0107$ $t = -3.021$

d.f = 25 – 2 = 23, $\alpha = 0.05$
$t_{\alpha/2} = 2.069$ (from table)

Plotting for the regression data:

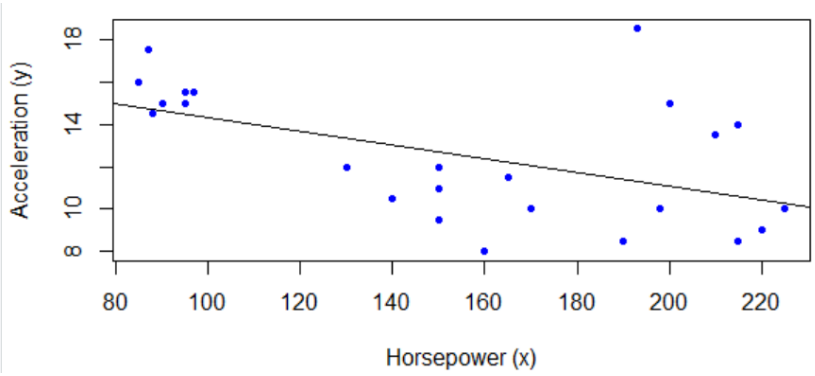


Figure 3.2

From the graph, we can conclude that the data has negative linear relationship.
**Decision:** Reject $H_0$ ($t_{\alpha/2} = -2.069 > -3.021$)
**Conclusion:** There is sufficient evidence that Horsepower of a car affects acceleration.

### 3.4 ANOVA

The ANOVA test is used to MPG of a car($x_1$) and the acceleration of a car($x_2$). The mean of each sample is calculated and have been labelled as $\bar{x}_1$ and $\bar{x}_2$.

| Cars | MPG ($x_1$) | Acceleration ($x_2$) |
|---|---|---|
| Chevrolet Chevelle Malibu | 18 | 12 |
| Buick Skylark 320 | 15 | 11.5 |
| Plymouth Satellite | 18 | 11 |
| AMC Rebel SST | 16 | 12 |
| Ford Torino | 17 | 10.5 |
| Ford Galaxie 500 | 15 | 10 |
| Chevrolet Impala | 14 | 9 |
| Plymouth Fury iii | 14 | 8.5 |
| Pontiac Catalina | 14 | 10 |
| AMC Ambassador DPL | 15 | 8.5 |
| Dodge Challenger SE | 15 | 10 |
| Plymouth 'Cuda 340 | 14 | 8 |
| Chevrolet Monte Carlo | 15 | 9.5 |
| Buick Estate Wagon (sw) | 14 | 10 |
| Toyota Corolla Mark ii | 24 | 15 |
| Plymouth Duster | 22 | 15.5 |
| AMC Hornet | 18 | 15.5 |
| Ford Maverick | 21 | 16 |
| Datsun PL510 | 27 | 14.5 |
| Peugeot 504 | 25 | 17.5 |
| AMC Gremlin | 21 | 15 |
| Ford F250 | 10 | 14 |
| Chevy C20 | 10 | 15 |
| Dodge D200 | 11 | 13.5 |
| Hi 1200D | 9 | 18.5 |
| N = 25 | $\bar{x}_1 = 412 / 25 = 16.480$ | $\bar{x}_2 = (310.5) / 25 = 12.420$ |

$H_0$: $\mu_1 = \mu_2$

$H_1$: at least one mean is different.

Standard deviation:

$$s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$$

$S_1 = 4.700$

$S_2 = 3.051$

Mean between samples:
$\bar{X} = (16.480 + 12.420) / (k = 2) = 14.45$

Standard deviation between samples:
$S_{\bar{x}} = 15.000$

Variance between sample:
$= 25(15.000)^2 = 5625.000$

Variance within sample:
$= [(4.700)^2 + (3.051)^2] / (k = 2) = 15.699$

Test statistic, F:
$= 15.000 / 15.699 = 0.955$

Numerator and denominator degree of freedom:

Numerator = 2 – 1 = 1

Denominator = 2(25-1) = 48

Critical value with $\alpha = 0.05$ from F-distribution table:

F-critical value = 4.08

**Conclusion:** since $F_{\text{test statistic}} < F_{\text{critical value}}$ (0.955 < 4.08), we fail to reject null hypothesis. There is sufficient evidence to claim that the MPG of a car has the same mean as acceleration of a car.

## 4.0 CONCLUSION

As a conclusion, we analyzed that every car maker will make the cars engine based on the suitable horsepower and MPG to get better acceleration. The data for both factors are collected from a reliable source for us to carry out multiple analyses. Results from the data analysis that we had done above, all the car maker more focus on the acceleration of their car that have connection to horsepower of its engine.

As we can see from the result of hypothesis on 1 sample test, we can conclude that the mean of acceleration of each car are not equal to $10m/s^2$. Thus, the mean for the acceleration can be more than $10m/s^2$. Next, we can see correlation results from two indicators which are horsepower and acceleration. The results we got was that both indicators are not in a linear relationship which means we fail to reject our null hypothesis. Additionally, regression test is used to identify between the independent and dependent indicators. From the result above, we can conclude that the horsepower affects the acceleration of a car. Finally, ANOVA test, we fail to reject null hypothesis because there is sufficient evidence that shows MPG has the same mean with acceleration.

In a nutshell, this project has sharpened our own skill in statistical analysis and help us gaining a new knowledge. The test that we used to analyze our data are a good medium for other person to use it to analyze data because it can make our work much easier. All tests that we used in this dataset are useful in identifying all factors more accurate and scientific. We believe that this project can give useful example for the community to do research about cars.

## 5.0 APPENDIX

Raw Data

| | Column1 | Column2 | Column3 | Column4 | Column5 | Column6 | Column7 | Column8 | Column9 |
|---|---|---|---|---|---|---|---|---|---|
| 2 | Car | MPG | Cylinders | Displacemen | Horsepowe | Weight | Acceleratior | Model | Origin |
| 3 | STRING | DOUBLE | INT | DOUBLE | DOUBLE | DOUBLE | DOUBLE | INT | CAT |
| 4 | Chevrolet Chevelle Malibu | 18.0 | 8 | 307.0 | 130.0 | 3504. | 12.0 | 70 | US |
| 5 | Buick Skylark 320 | 15.0 | 8 | 350.0 | 165.0 | 3693. | 11.5 | 70 | US |
| 6 | Plymouth Satellite | 18.0 | 8 | 318.0 | 150.0 | 3436. | 11.0 | 70 | US |
| 7 | AMC Rebel SST | 16.0 | 8 | 304.0 | 150.0 | 3433. | 12.0 | 70 | US |
| 8 | Ford Torino | 17.0 | 8 | 302.0 | 140.0 | 3449. | 10.5 | 70 | US |
| 9 | Ford Galaxie 500 | 15.0 | 8 | 429.0 | 198.0 | 4341. | 10.0 | 70 | US |
| 10 | Chevrolet Impala | 14.0 | 8 | 454.0 | 220.0 | 4354. | 9.0 | 70 | US |
| 11 | Plymouth Fury iii | 14.0 | 8 | 440.0 | 215.0 | 4312. | 8.5 | 70 | US |
| 12 | Pontiac Catalina | 14.0 | 8 | 455.0 | 225.0 | 4425. | 10.0 | 70 | US |
| 13 | AMC Ambassador DPL | 15.0 | 8 | 390.0 | 190.0 | 3850. | 8.5 | 70 | US |
| 14 | Citroen DS-21 Pallas | 0 | 4 | 133.0 | 115.0 | 3090. | 17.5 | 70 | Europe |
| 15 | Chevrolet Chevelle Concours (sw) | 0 | 8 | 350.0 | 165.0 | 4142. | 11.5 | 70 | US |
| 16 | Ford Torino (sw) | 0 | 8 | 351.0 | 153.0 | 4034. | 11.0 | 70 | US |
| 17 | Plymouth Satellite (sw) | 0 | 8 | 383.0 | 175.0 | 4166. | 10.5 | 70 | US |
| 18 | AMC Rebel SST (sw) | 0 | 8 | 360.0 | 175.0 | 3850. | 11.0 | 70 | US |
| 19 | Dodge Challenger SE | 15.0 | 8 | 383.0 | 170.0 | 3563. | 10.0 | 70 | US |
| 20 | Plymouth 'Cuda 340 | 14.0 | 8 | 340.0 | 160.0 | 3609. | 8.0 | 70 | US |
| 21 | Ford Mustang Boss 302 | 0 | 8 | 302.0 | 140.0 | 3353. | 8.0 | 70 | US |
| 22 | Chevrolet Monte Carlo | 15.0 | 8 | 400.0 | 150.0 | 3761. | 9.5 | 70 | US |
| 23 | Buick Estate Wagon (sw) | 14.0 | 8 | 455.0 | 225.0 | 3086. | 10.0 | 70 | US |
| 24 | Toyota Corolla Mark ii | 24.0 | 4 | 113.0 | 95.00 | 2372. | 15.0 | 70 | Japan |
| 25 | Plymouth Duster | 22.0 | 6 | 198.0 | 95.00 | 2833. | 15.5 | 70 | US |
| 26 | AMC Hornet | 18.0 | 6 | 199.0 | 97.00 | 2774. | 15.5 | 70 | US |
| 27 | Ford Maverick | 21.0 | 6 | 200.0 | 85.00 | 2587. | 16.0 | 70 | US |
| 28 | Datsun PL510 | 27.0 | 4 | 97.00 | 88.00 | 2130. | 14.5 | 70 | Japan |
| 29 | Volkswagen 1131 Deluxe Sedan | 26.0 | 4 | 97.00 | 46.00 | 1835. | 20.5 | 70 | Europe |
| 30 | Peugeot 504 | 25.0 | 4 | 110.0 | 87.00 | 2672. | 17.5 | 70 | Europe |
| 31 | Audi 100 LS | 24.0 | 4 | 107.0 | 90.00 | 2430. | 14.5 | 70 | Europe |
| 32 | Saab 99e | 25.0 | 4 | 104.0 | 95.00 | 2375. | 17.5 | 70 | Europe |
| 33 | BMW 2002 | 26.0 | 4 | 121.0 | 113.0 | 2234 | 12.5 | 70 | Europe |

Processed Data

| | Car | MPG | Displacemen | Horsepowe | Weight | Acceleratior |
|---|---|---|---|---|---|---|
| 2 | Chevrolet Chevelle Malib | 18 | 307 | 130 | 3504 | 12 |
| 3 | Buick Skylark 320 | 15 | 350 | 165 | 3693 | 11.5 |
| 4 | Plymouth Satellite | 18 | 318 | 150 | 3436 | 11 |
| 5 | AMC Rebel SST | 16 | 304 | 150 | 3433 | 12 |
| 6 | Ford Torino | 17 | 302 | 140 | 3449 | 10.5 |
| 7 | Ford Galaxie 500 | 15 | 429 | 198 | 4341 | 10 |
| 8 | Chevrolet Impala | 14 | 454 | 220 | 4354 | 9 |
| 9 | Plymouth Fury iii | 14 | 440 | 215 | 4312 | 8.5 |
| 10 | Pontiac Catalina | 14 | 455 | 225 | 4425 | 10 |
| 11 | AMC Ambassador DPL | 15 | 390 | 190 | 3850 | 8.5 |
| 12 | Dodge Challenger SE | 15 | 383 | 170 | 3563 | 10 |
| 13 | Plymouth 'Cuda 340 | 14 | 340 | 160 | 3609 | 8 |
| 14 | Chevrolet Monte Carlo | 15 | 400 | 150 | 3761 | 9.5 |
| 15 | Buick Estate Wagon (sw) | 14 | 455 | 225 | 3086 | 10 |
| 16 | Toyota Corolla Mark ii | 24 | 113 | 95 | 2372 | 15 |
| 17 | Plymouth Duster | 22 | 198 | 95 | 2833 | 15.5 |
| 18 | AMC Hornet | 18 | 199 | 97 | 2774 | 15.5 |
| 19 | Ford Maverick | 21 | 200 | 85 | 2587 | 16 |
| 20 | Datsun PL510 | 27 | 97 | 88 | 2130 | 14.5 |
| 21 | Peugeot 504 | 25 | 110 | 87 | 2672 | 17.5 |
| 22 | AMC Gremlin | 21 | 199 | 90 | 2648 | 15 |
| 23 | Ford F250 | 10 | 360 | 215 | 4615 | 14 |
| 24 | Chevy C20 | 10 | 307 | 200 | 4376 | 15 |
| 25 | Dodge D200 | 11 | 318 | 210 | 4382 | 13.5 |