SESSION 2019/2020 SEMESTER 2

COURSE CODE

SECI 2143 –PROBABILITY & STATISTICAL DATA ANALYSIS

LECTURER'S NAME

DR CHAN WENG HOWE

PROJECT 2- INDIVIDUAL

HOME LOAN

GROUPS' MEMBER

INTAN MARINA BINTI SULAIMAN (A19EC0053)

SECTION

02

Table of Contents

INTRODUCTION

A house loan or home loan simply means a sum of money borrowed from a financial institution or bank to purchase a house. Home loans consist of an adjustable or fixed interest rate and payment terms. So , based on the proposal of project 2, this project is about home loan provided by Dream Housing Finance company. The company predict the loan eligibility by automate the loan eligibility process based on customer detail provided while filling online application. The specification of taeget population of this project is the applicant of the home loans from all properties area.

This study aims to figure out the relationship between the details that the applicant fill in and the loan amount and the term of the loan. So that can predicts validates the customer eligibility for loan.

The data in this project was obtained from the dataset “Bank_loan” file name (madfhntr.csv). In this dataset there were these detailsLoan ID, Gender, Marital Status, Education, Number of Dependents, Self Employed status, Applicant Income, Coapplicant Income, Loan Amount, Loan Amount Term, Credit History, Property Area and Loan Status.

In this dataset there were 614 applicant was registered. To do the test analysis in R, 100 sample of 7 variables was taken from the population. Which is the sample file name is (Pro2). The variables also selected for analysis, Gender, Dependents, Education, Applicant Income, Coapplicant Income, Loan Amount, Loan Amount Term.

RESULT

1 SAMPLE TEST

```
mean(madfhantr$Dependents, na.rm = TRUE)
[1] 0.771615
```

Based on the sample data obtain form the dataset, we claim that the average of the dependents not 0.771615 (0.77). Hence the null hypothesis, H0 and alternative hypothesis H1 is as follows:

H0 : $\mu = 0.77$

H1 : $\mu > 0.77$

critical value = Z(0.05) = 1.644854 (Right-tailed test)

Where μ is the population mean of dependents. n is the number of data. A simple random sample of 100 data are used to figure out the Z-score of the sample population. The sample mean and standard deviation can be calculated using the formula:

Sample mean, $\bar{x} = \sum X / n$

Standard deviation, $\sigma = \sqrt{((\sum(X-\bar{x})2))/((n-1))}$

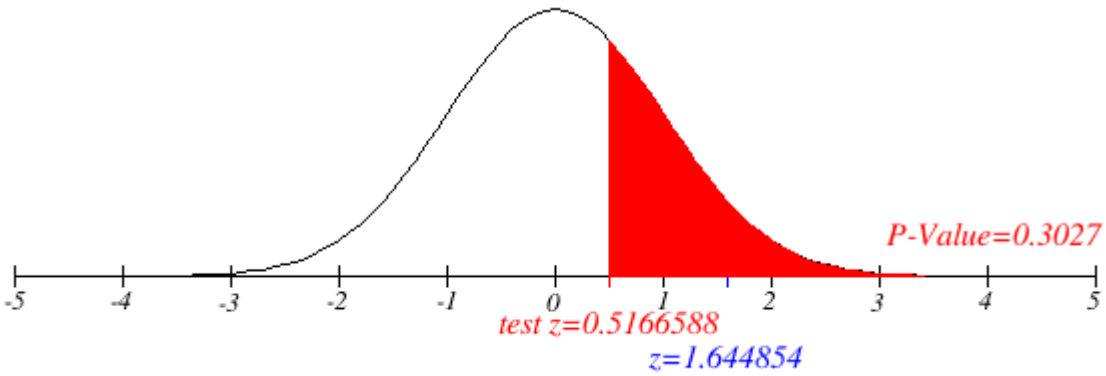
After calculated, the sample mean is 0.84 while standard deviation is 1.323601.

A significant level of 0.05 is used to test the claim that the average of the dependents are more than 0.77. The critical value of 0.05 significant level is 1.644854. The Z-value of the sample mean is 0.84 can be calculated by:

$Z = (\bar{x}-\mu)/ (\sigma/\sqrt{N})$ which is 0.5166588

Z-value and Critical value table

$\bar{x}$	μ	Z-value	Critical value
0.84	0.77	0.5288603	1.644854

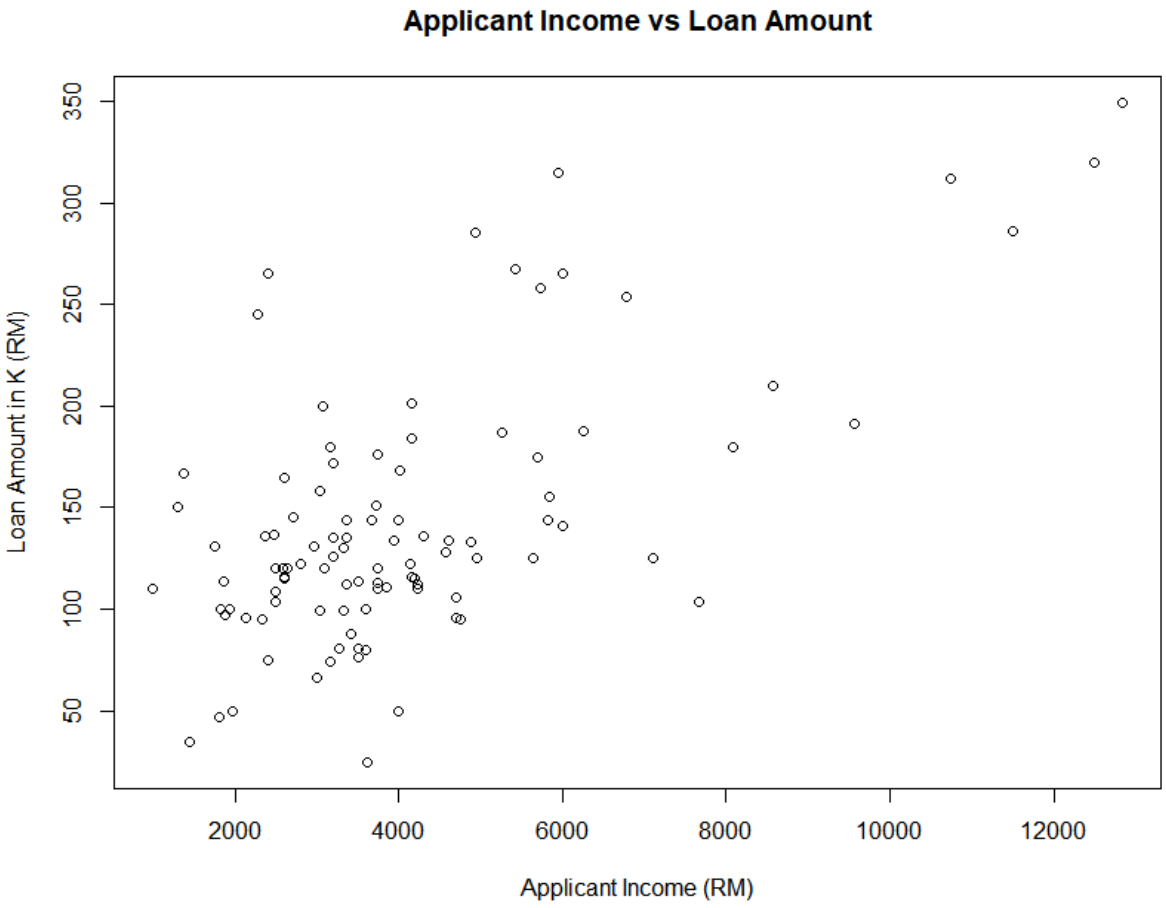


Since the Z-value < critical value (0.5288603 < 1.644854), the Z-value not fall in the critical region, and we can say that fail reject null hypothesis.

Hence, there is insufficient evidence to reject the null hypothesis where the average of the dependents is 0.77.

CORRELATION

The correlation was test using variables Applicant Income and Loan Amount.



In the correlation test, we analyses the strength and relationship between the Applicant Income and Loan Amount with a sample size of 100. The correlation coefficient is calculated to test the relationship between these two variables. The correlation coefficient is calculated using Pearson's technique, r is found to be 0.6394578. The results show that Applicant Income and Loan Amount have a moderate positive relationship. Which means if high Applicant Income, then high Loan Amount.

The value of linear regression model is $\hat{y}$= 67.61480 + 0.01824 x.

```
> cor (x , y)
[1] 0.6394578
> model <- lm(y ~ x)
> model

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
   67.61480      0.01824

> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-108.648  -32.808   -4.062   24.201  153.698

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 67.614803  10.402869   6.500 3.38e-09 ***
x            0.018241   0.002215   8.234 8.03e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.79 on 98 degrees of freedom
Multiple R-squared:  0.4089,    Adjusted R-squared:  0.4029
F-statistic: 67.79 on 1 and 98 DF,  p-value: 8.026e-13
```

REGRESSION

The regression of the two variables, are analysed to predict the of Loan Amount based on Applicant Income . The independent variable, x is Applicant Income (RM) while the dependent variable, y is the Loan Amount. Estimated regression model is used.

$$Y = b_0 + b_1 X$$

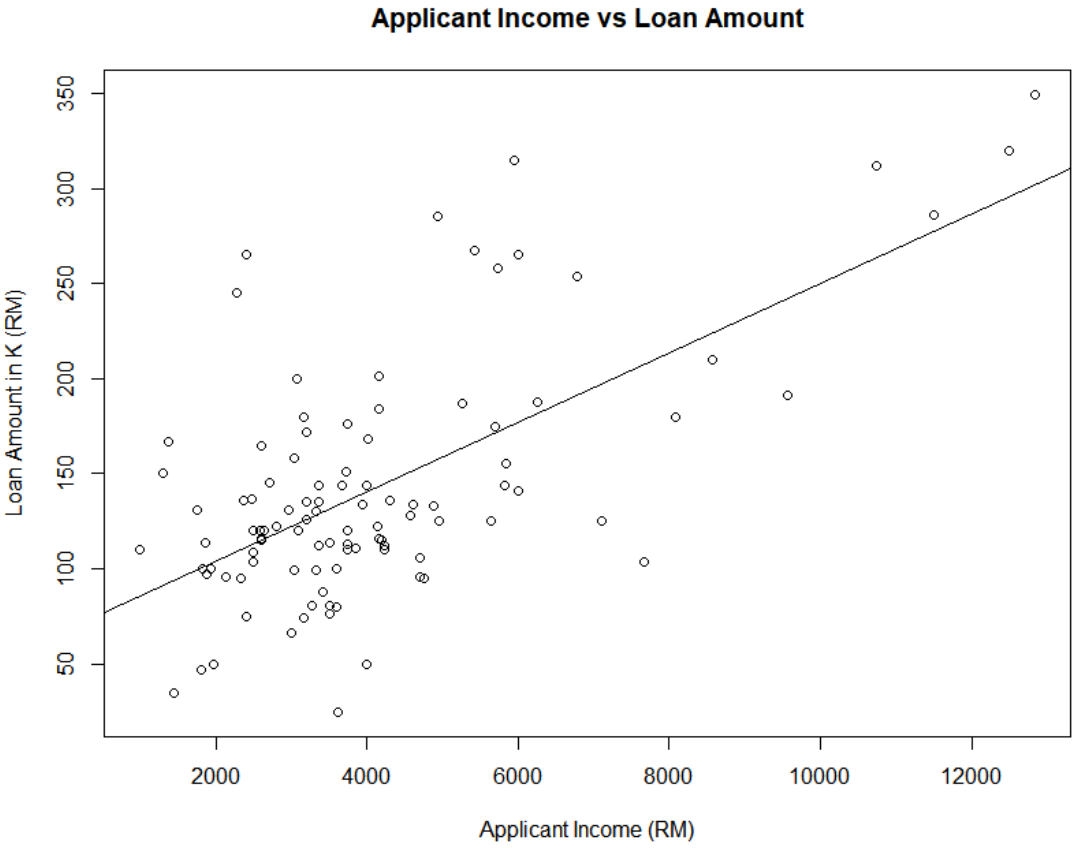
Y= estimated y value

b_0 = estimate of the regression intercept

b_1 = estimate of the regression slope

x = independent variable

The estimated regression model is calculated using RStudio, where obtain Y= 67.61480 + 0.01824 x.



CHI-SQUARE TEST OF INDEPENDENCE

```
> library(MASS)
>
> tb1 <- table(Pro2$Education,Pro2$Gender) #get contigency table
> chisq.test(tb1,correct = FALSE)

        Pearson's Chi-squared test

data:  tb1
X-squared = 1.6675, df = 1, p-value = 0.1966

> alpha <- 0.05
> x2.alpha <- qchisq(alpha,df=1,lower.tail = FALSE)
> tb1

              Female Male
  Graduate         24   53
  Not Graduate      4   19
```

Chi Square test of independent is used to test whether relationship is exists between the applicant Gender and Education. 100 data sample is used and the Gender is divided into group of Male and Female, while Education is distributed into Graduate and Not Graduate. We claim that the at 0.05 significant level, Gender is independent to the Education. The null hypothesis and alternative hypothesis are as follow:

H_0 : Gender is independent to the Education

H_1 : Gender is dependent to the Education

The expected count E_{ij} is calculated using this formula:

$$E_{ij} = \frac{(i^{th} Row\ total)(j^{th} Column\ total)}{Total\ sample\ size}$$

The $x^2 = 1.6675$, the degree of freedom: (2-1) (2-1) = 1. The critical region could be finds in the Chi-Square Distribution table with degree of freedom equal to 1 and significant level of 0.05. The critical values are found to be approximately 3.841459.

Since the test statistic, x^2 < critical value (1.6675 < 3.841459), the test statistic, x^2 did not fall within the critical region, therefore we fail to reject the null hypothesis. There is sufficient evidence to show that the Gender is independent to the Education.

CONCLUSSION

Based on the hypothesis we fail to reject null hypothesis because there is insufficient evidence to reject the null hypothesis where the average of the dependents is 0.77. Besides that, from the analysis, we found that there is a moderate positive relationship between the Applicant Income and Loan Amount with a correlation coefficient of 0.6394578. The estimated regression model obtained is Y= 67.61480 + 0.01824 x. This regression is important for the prediction on the Loan Amount based on the Applicant Income. The Chi-Square test of independence show that the applicant Gender is independent to the applicant Education.