UTM UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

SECI2143
Probability & Statistical Data Analysis

Project 2

Theme : Inference Statistical Analysis - Diabetes

Section : 09

Lecturer : Dr Rozilawati Binti Dollah

NO.	NAME	MATRIX NO.	PROGRAM
1.	HENG XING YU	A21EC0183	SECVH
2.	GAN JIA HUI	A21EC0177	SECVH
3.	SONDOS TAMER MOUSTAFA	A21EC0277	SECVH
4.	TAN KAI YUAN	A21EC0230	SECVH

Table of Contents

No.	Contents	Page
1	1.0 Introduction and Background	3
2	2.0 Dataset	3
3	3.0 Data Analysis 3.1 Hypothesis testing 1 sample 3.2 Correlation 3.3 Regression 3.4 Chi-Square Test of Independence	4 4 5-7 8-9 10
4	4.0 Conclusion	11
5	5.0 Appendix	11
6	6.0 Video Link	11
7	7.0 Eportfolio Link	11

1.0 Introduction and Background

Diabetes has been an ongoing issue in Malaysia for years. In fact, Malaysia has been known as the "sweetest nation in Asia". With diabetes comes a range of threatening health concerns, therefore we have decided to study a range of factors affecting the risk of diabetes which we decided to focus on and investigate.

We collected our data from a secondary source online and used it from a sample of 100 patients for our calculations. The sample consists of 66 females and 34 males, along with their ages, weight, and height and their calculated BMI in addition to other basic measurements like cholesterol and glucose.

The purpose of this study was to analyze the data and try to link it together, for example, studying the relationship between the BMI values and the gender of the patients to predict if it has an effect on it (the chi-square test). We predict that women will have higher values size naturally they have more body fat. In addition to that is finding the link between BMI and glucose levels of the patients which are expected to be directly related to each other.

Studying these factors will hopefully educate us more about diabetes, what can accelerate/ cause it and how to reduce its risks. It is crucial that people have basic knowledge on what can influence their health negatively in order to abstain from it.

2.0 Dataset

We collected all of the data that will be used in this project from the website Kaggle. The data that we used came from the National Institute of Diabetes and Digestive and Kidney Diseases. In order to improve visual understanding of the scatter plot, we decided to minimize the sample size from 350 to 100 sample patients. So, we randomly pick 100 sets information of about patients from the dataset. There are several variables in the data, including patient number, cholesterol, glucose, HDL chol, chol HDL ratio, age, gender, height, weight, BMI, systolic bp, diastolic bp, waist, hip, and diabetes. We decided to use variables BMI in hypothesis testing 1 sample after consulting with all members. Then, we select the variables cholesterol and age to be used in correlation Next, the variable BMI and glucose is used for the regression. In the Chi-Square test of independence, we choose variables BMI and gender. In this project, correlation and regression are required tests, and we choose the chi-square test as optional. Many mediums such as R studio and Excel, helped us during this project. R studio is mainly used in the project for data analysis and testing. All the tests will be analyzed by using 0.05 significance level.

3.0 Data Analysis

3.1 Hypothesis testing 1 sample

From Wikipedia, we know the population average body mass index (BMI) in Malaysia is 25.3 kg/m^2 (Wikipedia, 2015). In order to check whether the sample means BMI of Malaysia in our data set is equal to or higher than 25.3 kg/m^2, we decided to use hypothesis testing 1 sample. We assume the confidence level is 95% ($\alpha = 0.05$) in this hypothesis testing 1 sample.

Hypothesis Statement:

- The null hypothesis is $H_0 : \mu = 25.3 kg/m^2$
- The alternative hypothesis is $H_1 : \mu > 25.3 kg/m^2$ (right-tailed test)

We used RStudio to calculate the sample mean of BMI with formula $\bar{X} = \frac{\sum x}{n}$ and the sample mean is 28.374kg/m^2 . The sample standard deviation of BMI is calculated by the formula:

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

The result of sample standard deviation (s) is 6.651. Since we have 100 sample which is more than 30, the test statistic for mean can be calculated by using formula: $z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$

alpha	0.05
n	100
pmean	25.3
sd	6.65109727409658
smean	28.374
z	4.62179377825675
z.alpha	1.64485362695147

Denotation:
n = total sample
pmean = population mean
smean = sample mean
sd = standard deviation of sample
z = z-score
z-alpha = critical value

By using this formula, we found the z-value is 4.622. The critical region, α is equal to 0.05 and the critical value can be obtained from the Z-score table which is 1.645. Since the z-value is greater than the critical value (4.622 >1.645), this means that the test statistic falls in the critical region and the null hypothesis, H_0 is rejected. There is sufficient evidence to conclude that $\mu > 25.3 kg/m^2$. These results suggest that the BMI from the sample test is greater than the population means BMI in Malaysia which is 25.3.

3.2 Correlation

For the correlation part, we will be using the relationship between age and cholesterol. Thus, the correlation test is used to determine whether higher age will cause higher cholesterol. Naturally, age is regarded as an independent variable, whereas the cholesterol variable is a dependent variable. age is initialised to a x-value and the cholesterol to a y-value. We are using Pearson's product-moment correlation coefficient technique to examine the correlation between the variables because the variables for age and cholesterol are ratio type variables.

Values	
total_x	2698L
total_x2	74814L
total_xy	516968L
total_y	19060L
total_y2	3769418L

Values calculated in R studio:
- $\Sigma x = 2698$
- $\Sigma y = 19060$
- $\Sigma xy = 516968$
- $\Sigma x^2 = 74814$
- $\Sigma y^2 = 3769418$

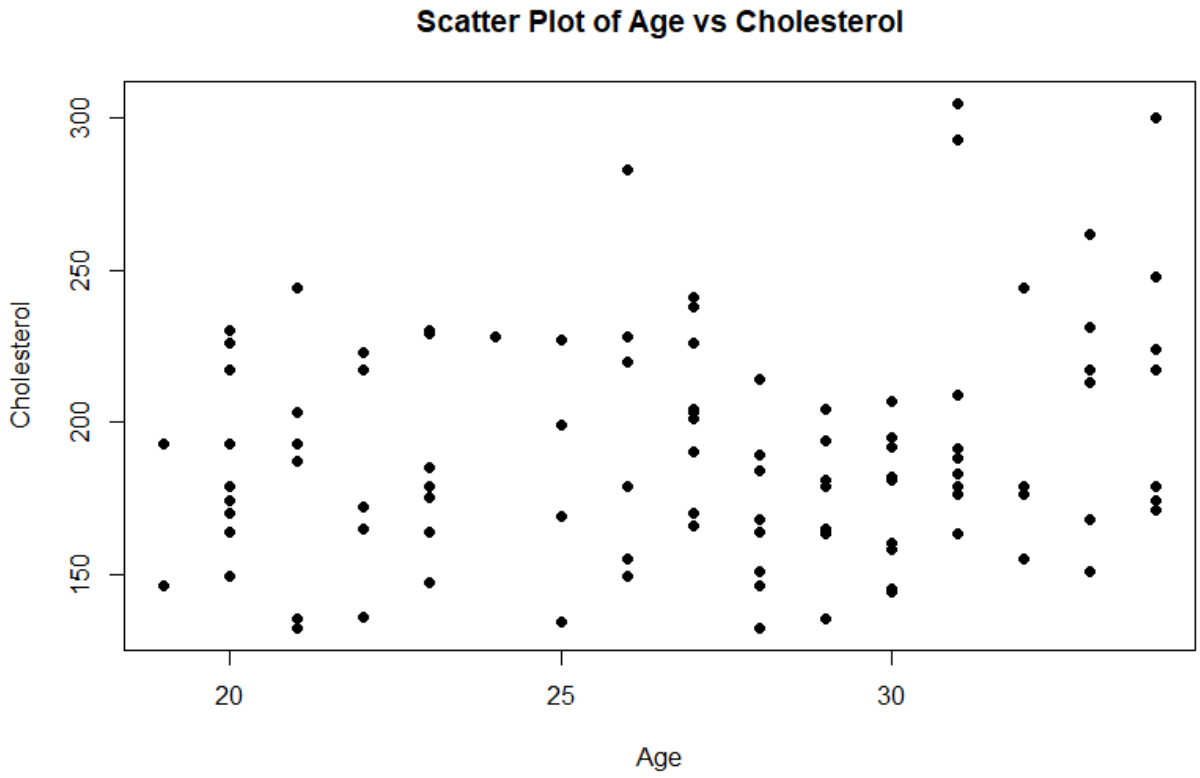
Based on the data, we have calculated the required value for the Pearson's product-moment correlation coefficient whereas n = 100. By using the function [cor (x, y, method = "pearson")] in R studio, we get the result of r = 0.1642299, which is the same result as the formula:

$$r = \frac{\Sigma xy - \frac{(\Sigma x \Sigma y)}{n}}{\sqrt{\left(\Sigma x^2 - \frac{(\Sigma x)^2}{n}\right)\left(\Sigma y^2 - \frac{(\Sigma y)^2}{n}\right)}}$$

$$r = \frac{516968 - \frac{(2698)(19060)}{100}}{\sqrt{\left(74814 - \frac{(2698)^2}{100}\right)\left(3769418 - \frac{(19060)^2}{100}\right)}}$$

$$r = 0.1642299$$

where,
r = Sample correlation coefficient
n = Sample size
x = Value of the independent variable
y = Value of the dependent variable



Significance Test for Correlation

Hypotheses:-

H_0 : $\rho = 0$ (no linear correlation)

H_1 : $\rho \neq 0$ (linear correlation exist)

Degree of freedom:-

df = 100 – 2 = 98

Assume:-

$\alpha = 0.05$

$t_{0.05,98} = 1.6606$

Test statistics:

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

$$t = \frac{0.1642299}{\sqrt{\frac{1 - 0.1642299^2}{100 - 2}}}$$

$$t = 1.6482$$

where,
t = T score
r = Sample correlation coefficient
n = Sample size

In conclusion, we saw that r = 0.1642299, which is closer to 0. Thus, it is shown that there is no linear relationship between the variables of age and cholesterol. The test statistics calculated using formula are 1.6482, which is lower than the critical value of 1.6606 using r = 0.1642299. As a result, the null hypothesis, H_0 is not rejected. There is not sufficient proof that age and cholesterol are linearly related. So, age does not affect the value of cholesterol no matter how high or low it is.

3.3 Regression

We use two variables selected from the dataset in this regression analysis.

Variable:

- Variable x: Body mass index (BMI) of 100 sampled patients (Independent variable)
- Variable y: Glucose of 100 sample patients (Dependent variable)

Hence, this regression test is used to find the relationship between BMI and glucose among 100 sample patients. To test whether there has a linear relationship between x and y, we determine the hypothesis as follows:

- H_0: $\beta 1 = 0$ (no linear relationship)
- H_1: $\beta 1 \neq 0$ (linear relationship does exist)

First, we use the **lm()** function to perform the linear regression in R.

```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    69.6907       0.6351
```

and we get an estimated regression model:

$$\hat{p} = 69.6907 + 0.6351x$$

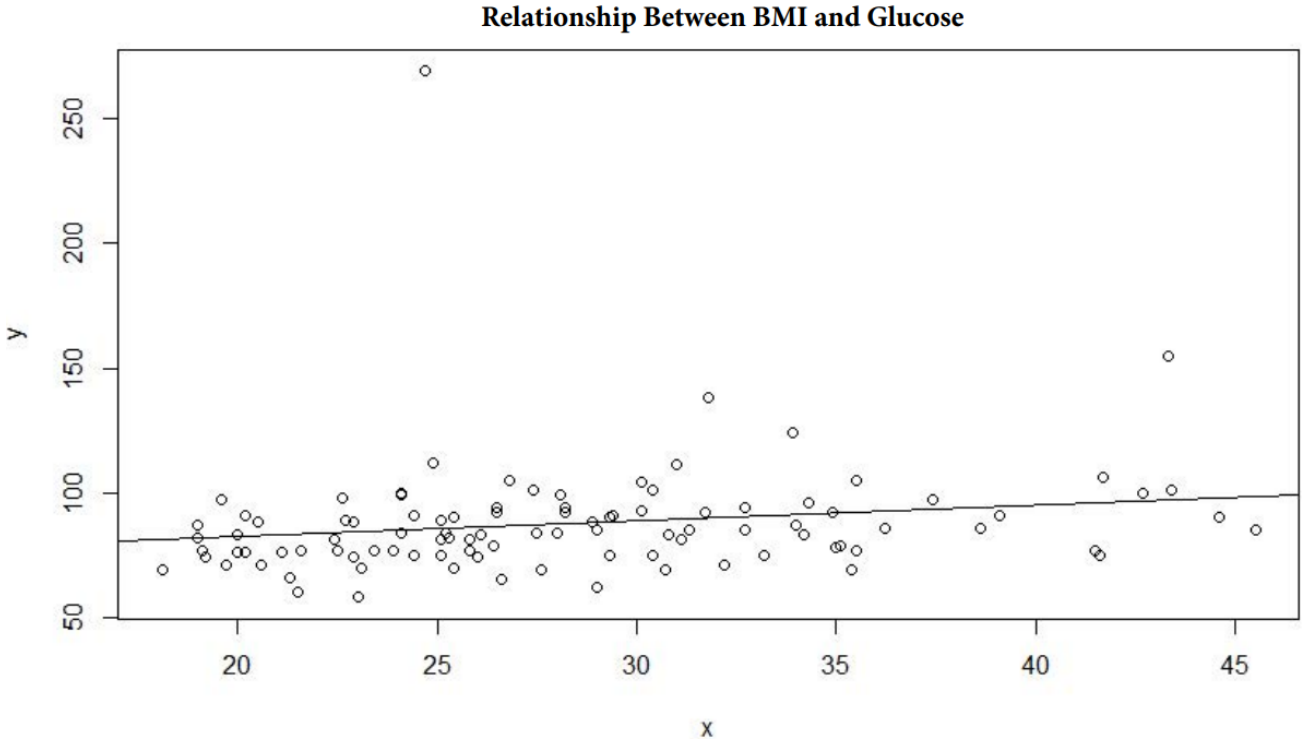
After we get the estimated regression model, we can do a summary of this model and we can find the relationship between BMI and glucose.

```
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-26.297 -10.333  -3.640   4.618 183.623

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  69.6907    10.2596   6.793 8.57e-10 ***
x             0.6351     0.3521   1.803   0.0744 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.3 on 98 degrees of freedom
Multiple R-squared:  0.03212,   Adjusted R-squared:  0.02225
F-statistic: 3.253 on 1 and 98 DF,  p-value: 0.07439
```

From the summary in R, we can see that R-squared, R^2 is equal to 0.03212. R^2 is calculated using the formula sum of squares explained by regression (SSR) divided by the total sum of squares (SST). Since 0.03212 is between 0 and 1, this shows a weaker relationship between BMI and glucose. Moreover, we also have the t-test for the slope. The test statistic is 6.793 and it is calculated by using formula: $t = \frac{b_1 - \beta_1}{s_{b_1}}$ when b_1 is sample regression slope coefficient, β_1 is the

hypothesis slope and s_{b1} represents the estimator of the standard error slope. The critical value for the slope is 1.9845 when the alpha, α is 0.05 and the degrees of freedom is 98. As we can see, 6.793 > 1.9845, it shows the H_0 is rejected. There is sufficient evidence to conclude that the linear relationship between BMI and glucose does exist. This also means that there is the relationship between BMI and glucose.

Next, we use **plot()** and **abline()** functions to generate the scatter plot to have a better visual understanding of this regression test.



As we can see, this scatters plot has a weak positive linear relationship and this means that there is a weak relationship between BMI and glucose. This observation is the same as the conclusion above.

3.4 Chi-Square Test of Independence

In this Chi-Square test, the Chi-Square test analyzed the relationship between two variables which are BMI and gender. BMI is represented as x and gender is represented as y. To measure this dataset, a modified dataset has been prepared. The BMI score is divided into 4 categories which are $x < 18.5$, $18.5 \leq x \leq 24.9$, $25 \leq x \leq 29.9$ and $x \geq 30$. Besides, the genders are divided into two which are male and female. Thus, we want to test whether there is a relationship between the BMI score and the genders.

Hypothesis:

H_0: There is no relationship between BMI and genders.

H_1: There is a relationship between BMI and genders.

BMI	Male	Female	Total
$X \leq 18.5$	0	1	1
$18.5 \leq X \leq 24.9$	18	15	33
$25 \leq X \leq 29.9$	8	21	29
$X \geq 30$	8	29	37
Total	34	66	100

```
        Pearson's Chi-squared test

data:  d
X-squared = 9.7808, df = 3, p-value = 0.02052
```

From the calculation, the chi-square value is 9.78 which is greater than the critical value calculated which is 7.82.

Conclusion:

Since the chi-square value is greater than the critical value ($9.78 > 7.82$), the null hypothesis, H_0 is rejected. Hence, there is strong evidence that there is a relationship between BMI and the genders. This hypothesis is supported by Hocem Benmansour's research, which predicts patients' BMI.

4.0 Conclusion

In short, we decided to pick a topic that we thought was interesting and still relevant. And in the end, we agreed on studying factors influencing diabetes and how they're related to each other and came up with ideas for each statistical test. Firstly, we used hypothesis testing to test if the sample means from our data set is equal to, higher or lower than 25.3 kg per meter squared. Moreover, we used correlation to investigate the relationship between age and cholesterol levels and looked further into how glucose levels can affect body mass index value using regression. Lastly, we were able to use the chi-square test of independence to find a link between the patient's gender and their BMI values. It was interesting to be able to dissect the data set we collected and study it in detail using statistical methods. And most importantly educate ourselves and understand the results of our tests.

5.0 Appendix

1. Population BMI
   List of countries by body mass index - Wikipedia
2. Dataset source
   Predict diabetes based on diagnostic measures | Kaggle

6.0 Video link: https://youtu.be/al3rkm4QgbE

7.0 Eportfolio link:

1. **https://eportfolio.utm.my/user/heng-xing-yu/seci2143-09-kebarangkalian-statistik-analisis-data-probability-statistical-data-analysis (Heng Xing Yu)**
2. **https://eportfolio.utm.my/user/tan-kai-yuan/seci2143-09-probability-statistical-data-analysis (Tan Kai Yuan)**
3. **https://eportfolio.utm.my/user/sondos-tamer-ezzat-youssef-mou/seci2143-probability-statistical-data-analysis (Sondos Tamer Mustafa)**
4. **https://eportfolio.utm.my/user/gan-jia-hui/seci1013-11-struktur-diskrit-discrete-structure (GAN JIA HUI)**