PROBABILITY & STATISTICAL DATA ANALYSIS
SECI 2143

PROJECT 2

SECTION 04 GROUP 4

LECTURER : DR ARYATI BAKRI

GROUP MEMBERS :

NO.	NAME	MATRIC NO
1.	SAFURA BALQIS BINTI AZMAN	A21EC0224
2.	NUR ATHIRA NABILA BINTI LUKMAN	A21EC0109
3.	AYESHA IMELDA BINTI ROHAIZAN	A21EC0164
4.	MUHAMMAD ZULFADHLY BIN MUHAMMAD AZHAR	A21EC0209

TABLE OF CONTENT

1.0 INTRODUCTION

Student behavior is an important aspect to be observed as a person’s behavior affects many other angles in his or her life. For example, a good student can be judged in terms of its specifications, such as certification courses, college mark, salary expectation, possibility of choosing their career based on their degree, part time jobs, etc., so that we can apply the use of statistics to demonstrate whether there is a connection between the data. To accomplish this goal, a few possible variables are chosen, and several test studies are conducted.

2.0 BACKGROUND OF STUDY

The student at Birla Institute of Applied Sciences in Haldwani, Uttarakhand, India, Akshat Giri, obtained the secondary data source for the dataset on student behaviour, which was retrieved through the Kaggle website. This dataset was compiled to display information about 235 samples of students, including their gender, department, hobbies, amount of study time per day, certification courses, height, weight, grade point averages in high school and college, and preferences for various degrees.

3.0 OBJECTIVE OF STUDY

1. To apply and carry out statistical test analysis on secondary data sources.
2. To demonstrate whether the given dataset variables are interdependent.

4.0 DESCRIPTION OF DATASET

Selected Variables	Objectives	Test Analysis and Expected Outcome
Part Time, College Mark	To test whether the mean of college marks when students do part time jobs is larger than the mean of college marks when students do not do part time jobs.	**Analysis:** 2 Sample Hypothesis Testing (Test on Mean, Variance Unknown) **Expected Outcome:** The mean of college marks when students do part time jobs is larger than the mean of college marks when students do not do part time jobs, at 95% confidence level, assuming variances unequal.
Salary Expectation, Possibility of choosing their career based on their degree	To test whether a linear relationship exists between the daily salary expectation and possibility of choosing their career based on their degree using Pearson's Product-Moment Correlation Coefficient, at 95 % confidence level.	**Analysis:** Correlation Analysis **Expected Outcome:** There is a strong linear relationship between the salary expectation and the possibility of choosing their career based on their degree, at a confidence level 95 %.
College Mark, Expected Salary	To test whether the expected salary depends on the value of college mark, use college mark as the independent variable(x) and expected salary as the dependent variable(y).	**Analysis:** Regression Analysis **Expected Outcome:** The expected salary depends on the value of the college mark. The higher the college mark, the higher the expected salary.
Gender, Part time Job	To test whether the Part Time and the gender of students are related using Two Way Contingency Table, at 95% confidence level.	**Analysis:** Chi-square Test Independence **Expected Outcome:** The part time job and the gender of the students are interrelated and independent at 95% confidence level.

5.0 DATA ANALYSIS AND DISCUSSION

A) 2-SAMPLE HYPOTHESIS TESTING

In this analysis, we are using variables part_time and college_mark, where we will test whether the mean of college marks when students do part time jobs is larger than the mean of college marks when students do not do part time jobs at 95% confidence level, assuming unequal variances. From the data, frequency(n), mean($\bar{x}$), standard deviation(s) are calculated.

Now that we have calculated the data, we can group them:

$\bar{x}_1 = 72.12195$	$\bar{x}_2 = 70.3517$
$s_1 = 13.59631$	$s_2 = 16.15662$
$n_1 = 41$	$n_2 = 194$

where group$_1$ is for students who do part-time jobs and group$_2$ is for students who do not do part-time jobs.

1.Hypothesis statement:

H_0: $\mu_1 = \mu_2$

H_1: $\mu_1 > \mu_2$

where µ1 equals the mean of college marks of students who do part-time jobs, and μ_2 equals the mean of college marks of students who do not do part-time jobs.

2. Given 95% confidence level, $\alpha = 0.05$. The test statistics, t_0 can be calculated by:

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

By using RStudio, test statistics, $t_0 = -\ 0.73164$.

3. Calculate the degree of freedom by:

Calculating the test statistic, degree of freedom and p-value.

By using RStudio, degree of freedom, v = 66.215.

Therefore, using α = 0.05, we reject if H_0 if $t_0 = - 0.73164 > t_{0.025, 66.215} = - 1.6683$

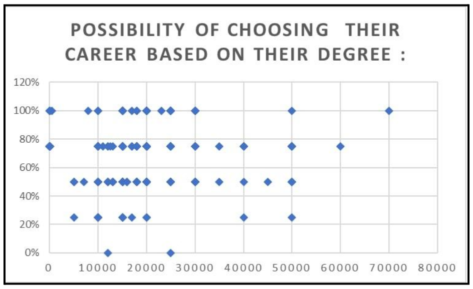
∴ Critical value, $t_{0.025, 66.215} = -1.668$, p-value = 0.467

4. Conclusion:

Since test statistics $t_0 = - 0.73164 >$ critical value $t_{0.025, 66.215} = - 1.6683$, we reject the null hypothesis. There is sufficient evidence to conclude that the mean of college marks of students who do part-time jobs is larger than college marks of students who do not do part-time jobs, at α = 0.05.

B) CORRELATION TEST

We conduct a correlation test to find out whether there is a linear relationship between the 'Salary Expectation' and 'possibility of choosing their career based on their degree' with a random sample of 235 students. Hence, in this analysis, we are using 'salary expectation' and 'possibility of choosing their career based on their degree' for this test using Pearson's Product-Moment Correlation Coefficient, at 95% confidence level and significance level of 0.05. Correlation analysis is used to measure the strength of association (linear relationship) between two variables. For the correlation coefficient, we use Pearson's Product-Moment Correlation Coefficient since variables 'salary expectation' and 'possibility of choosing their career based on their degree' are ratio-type data.



Scatter plot of 'salary expectation' and 'possibility of choosing their career based on their degree'

Based on the scatter plot above, the x-axis is the expected salary and the y-axis is the 'possibility of choosing their career based on their degree'. From the scatter plot above, it indicates that there is a weak correlation relationship (negative correlation) between 'salary expectation' and the 'possibility of choosing their career based on their degree'. It can be seen that, the possibility of choosing their career based on their degree decreases as the salary expectation decreases. However, since this Pearson's Product-Moment Correlation Coefficient is sensitive to outliers, there are a few outliers also on the top right side of the plot.We can conclude that the scatter plot above shows a negative linear correlation.

Analysis :

1. Calculate the sample correlation coefficient using Pearson's method by:

By using RStudio, we get the sample correlation coefficient, r = -0.02972485, which indicates that there is a relatively weak correlation between x ('salary expectation') and y (possibility of choosing their career based on their degree).

2. Significance Test for Correlation

Let x = salary expectation
and y = possibility of choosing their career based on their degree

a) Hypothesis Statement:

H0: ρ = 0 (There is no linear correlation between x and y)
H1: ρ ≠ 0 (There is a linear correlation exists between x and y)

b) Calculate test statistic:

Explanation:
By using RStudio we get the test statistic t = –0.4539307

c) Critical value

Significance level : α = 0.05
Degree of freedom : df = 233

Value from table :
Critical value tα/2=0.025, df=233 = 1.960

Value from R programming :
Critical value using R programming : 1.65142

d) State the decision and conclusion:

Based on calculating t-test and critical value from R programming, $t < t(\alpha/2=0.025, df=233)$. Thus we can reject the null hypothesis. Therefore by using a significance level of 0.05, there is sufficient evidence to conclude that a linear correlation exists between 'salary expectation' and the 'possibility of choosing their career based on their degree'.

C) REGRESSION TEST

In this analysis, we are using variables Salary Expectation and College Mark, where we will test whether the Salary Expectation depends on the College Mark, using College mark as the independent variable(x) and Salary expectation as the dependent variable(y). Our regression model is a linear model, hence simple linear regression is used. The changes in the values of expected salary are assumed to be caused by the changes in the college marks.
The mathematical equation for Population Linear Regression:

The mathematical equation for Population Linear Regression:

A) Estimated Regression Model

Dependent variable (Y) = Salary expectation
Independent variable (X) = College mark

Hence, we get the value of b1 and b0 from the RStudio which are b1 = -729.2492 and b0 = 84010.84 .

$$\hat{yi} = 84010.84+(-729.2492)x$$

From the regression model equation, we can interpret the intersection coefficient b0, and slope coefficient b1.

B) Inference about the Slope: t-Test

1) Hypothesis Statement:
   H0: $\beta1 = 0$ (no linear relationship)
   H1: $\beta1 \neq 0$ (linear relationship does exist)

2) Find critical value, using $\alpha = 0.05$, $df = n-2 = 233$

We get the critical value from the RStudio which is 1.65142

3) Calculate test statistic by using RStudio:

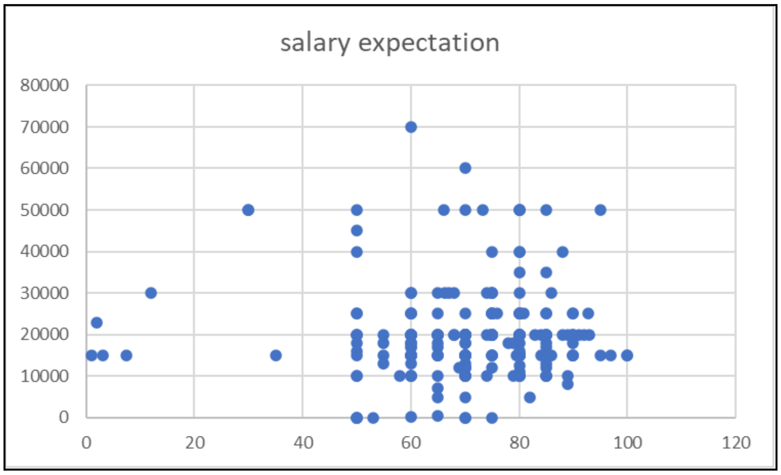
We get the value of test statistic t = -1.509535

4) Discussion :

Since test statistics t = -1.509535 < upper tail critical = 1.65142, we fail to reject the null hypothesis.

5) Conclusion :

There is no sufficient evidence to support that college marks affect salary expectation, at $\alpha = 0.05$.



D) CHI-SQUARE TEST - TWO WAY CONTINGENCY TABLE

In this analysis, we are using variables **part_time** and **Gender** where we will test whether the height and the gender of the students are related using a Two way Contingency Table, at 95% confidence level. Hence, we use the Chi-Square Test of Independence, with a two-way contingency table.

1. State the test hypothesis:
   H0:There is no relationship between Gender and part_time.
   H1:Gender and part_time are related and dependent.

2. Find the critical value :
   Critical value x^2= 3.841 (with df(2-1)(2-1)=1,alpha=0.05)

3.Calculate the expected frequency

Gender	part_time				
	Yes		No		Total
	obs	exp	obs	exp	
Female	71	194x79/235=65.2	8	41x79/235= 13.8	79
Male	123	194x156/235=128.8	33	41x156/235= 27.2	156
Total	194	194	41	41	235

*Remarks=eij >=5

4. Calculate the test statistic:

- Calculate manually:

Cell,ij	**Observed Count=oij**	**Expected Count=eij**	(*oij*−*eij*)**2/*eij***
1,1	71	194x79/235=65.2	(71-65.2)2/65.2=0.5156
1,2	123	194x156/235=128.8	(123-128.8)2/128.8=0.2612
2,1	8	41x79/235= 13.8	(8-13.8)2/13.8=2.4377
2,2	33	41x156/235= 27.2	(33-27.2)2/27.2=1.2368
		x2	4.4513

We calculate test statistics manually, we get test statistic x2=4.4513.When we calculate test statistics using R studio, we get test statistic x2= 4.4276, with p-value=0.03536.

5.State the decision:
Since the test statistic value (x2=4.4276)> critical value(x2 k=1, alpha=0.05=3,841), it does fall within the critical region. Thus, we reject H0. There is enough evidence to conclude that there is no relationship between the variables part_time and gender, at alpha=0.05.

6.0 CONCLUSION

For the 2 sample hypothesis testing, where we test on the mean assuming unequal variances, we found out that the mean of college marks of students who do part-time jobs is larger than the mean of college marks of students who do not do part-time jobs, hence we reject the null hypothesis. In the real world, this conclusive statement is not accurate considering that students who do part-time jobs have less time to study compared to students who do not do part-time jobs.

Next, for the correlation analysis test, we are using 'salary expectation' and 'possibility of choosing their career based on their degree' for this test using Pearson's Product-Moment Correlation Coefficient, at 95% confidence level and significance level of 0.05. Thus, we can conclude that there is sufficient evidence to support that the linear correlation between 'salary expectation' and the 'possibility of choosing their career based on their degree exists.

After that for the regression analysis, we want to test if the Expectation Salary variable is independent of College Mark. From this test, we found that there is no relation between these two variables because we have rejected the null hypothesis.

Lastly, for the chi-square test of independence, we found out that there is no relationship between gender and part time, hence we fail to reject the null hypothesis. In the real world, gender also does not affect whether someone does a part time job or not.

In conclusion, we can perform test analysis such as 2 Sample Hypothesis Testing, Correlation Analysis, Regression Analysis and Chi-square Test Independence using R Studio. We believe that this project is very useful for our future as this project has developed our data analysis skills. Also, special thanks to our lecturer, Dr Aryati Bakri for her help and guidance throughout this project.

7.0 APPENDIX

Original Dataset Link :

(Student_Behaviour)
https://bit.ly/3OOt8qV

Video Presentation Link :
https://youtu.be/D3bc8uCjAD4

Eportfolio Link :

1) Muhammad Zulfadhly :
https://eportfolio.utm.my/user/muhammad-zulfadhly-bin-muhamma/reflection-seci2143
2) Nur Athira Nabila :
https://eportfolio.utm.my/user/nur-athira-nabila-binti-lukman/seci2143-psda-project-2
3) Safura Balqis :
https://eportfolio.utm.my/user/safura-balqis-azman/probability-statistic-data-reflection
4) Ayesha Imelda :
https://eportfolio.utm.my/user/ayesha-imelda-binti-rohaizan/reflection-project-2-seci2143