PROBABILITY AND STATISTICAL DATA ANALYSIS (SECI2143-08)

PROJECT 2

GROUP AMONG US

NAME	MATRIC NO.
1. Cheng Chea Hee	A20EC0025
2. Amr Faisal Hamad	A20EC0259
3. Bilkis Musa	A20EC0233
4. Fatin Aimi Ayuni Binti Affindy	A20EC0190

Section: 08

Lecturer: Dr Sharin

TABLE OF CONTENT

INTRODUCTION

Wherever a person may be, maybe in grade school, highschool, university or even work environments at some point in time most of us have been exposed to the idea that females are better than males when it comes to academics, some may have even thought of it themselves and thus in this study we will dive into the numbers behind this comparison and attempt to determine the assumption is correct or not.

DATA SET

In this project, we have chosen to use the data provided by the lecturer which is the Student Performances dataset. In the dataset provided, there is a total of 518 females and 482 males datas been collected. Our objective for this project is to prove that the females have better performances compared to males. Thus, we randomly chose 200 females and 200 males dataset to conduct our testing to see if our assumption is valid to avoid any bias because the original number of datasets for both genders are not equal.

The unprocessed datasets contain four qualitative datas and three quantitative datas. Throughout all the testing we did, we only used the quantitative data which are; math score, writing score and reading score. As for the hypothesis testing 2 samples and chi square testing, we decided to manipulate the data by adding the three scores to create a new variable which named total scores to continue making the testing. For correlation testing, we use the variable math score and reading score to see if there is any correlation between the two variables. For regression testing, we use reading score as an independent variable and writing score as dependent variable to see if they both have a relationship to each other.

Variable	**Level of measurement**
Gender	Nominal
Math Score	Ratio
Reading Score	Ratio
Writing Score	Ratio
Total Score	Ratio

Table 1 shows the processed dataset that will be used in this project

DATA ANALYSIS

Hypothesis 2 Sample Test

The basis of our hypothesis analysis stems from a very common somewhat stereotypical comparison that most of us at some point in our lives have come across; provided our dataset, we decided to study the performance of both male and female students relative to their scores on a few tests and determine whether female students perform better than their male counterparts.

Since we are comparing the means of both genders' performances, a 2 sample test was in favor.
A t-statistical test was conducted with the following assumptions:

- A 5% significance level (95% confidence level).
- Variances are considered unequal (since it wasn't specified otherwise).
- $\mu 1$ is the mean score for female students.
- $\mu 2$ is the mean score for male students.
- Null hypothesis (H0: $\mu 1 = \mu 2$)
  Alternative hypothesis (H1: $\mu 2 < \mu 1$)

Based on the above assumptions, we reject the null hypothesis(H0) if: **$t0 < t(0.05, 397) = -1.64$.** However, since we have found our calculated value of t0 to be +3.04, which is well outside of the rejection zone, we can conclude that we fail to reject the null hypothesis and that there is in fact not enough evidence to support the claim that female students perform better than males.

Correlation Test

We did spearman correlation with the same dataset as the previous test. Using scatter plot to show the relation between math score and reading score. We divided the dataset into two which are female and male. To measure the strength of linear relationship between math score and reading score of two dataset. From the scatter plot, we can also know the relationship of female dataset and male dataset.

Hypothesis statement for correlation test:

Null hypothesis, H0 : $\rho = 0$ (no linear correlation)
Alternative hypothesis, H1: $\rho \neq 0$ (linear correlation exist)

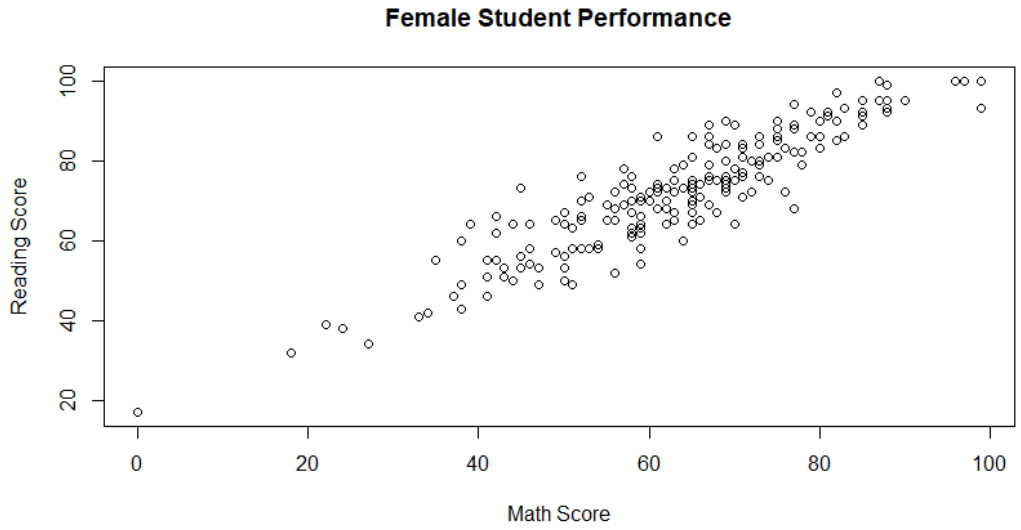


Figure 1 show scatter plot of female student performance in math score vs. reading score

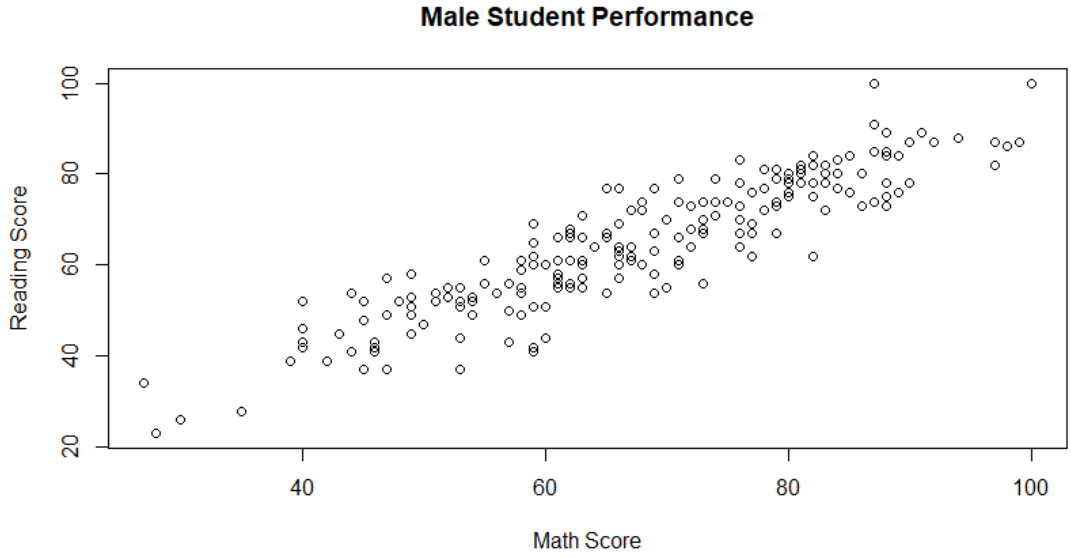


Figure 2 show scatter plot of male student performance in math score vs. reading score

From figure 1 and 2, we can see that both dataset show strong positive linear relationships.
For female student performance, it produced rho = 0.8867995. For male student performance, rho = 0.9099209. Means that both rejected the null hypothesis. From the scatter plot, we can see that the female dataset has a high concentration at range 60-80. However, male data set has a high concentration at range 40-80. It has shown that females perform better than males.

Regression Test

Using the same dataset as the previous tests, in the regression test, we would like to see if the writing score of a student is significantly related to the reading score at a level significant of 0.05. In these tests, we did two different regression tests on male dataset and female dataset to compare which gender performs better in the tests. Both dataset used have writing score as the dependent variable(y) meanwhile reading score as the independent variable(x). The tests are done to learn the existence of a linear relationship between the two variables. Both tests for female and male students have the same hypothesis statement which are:

The null hypothesis, $H_0 : \beta 1 = 0$
The alternative hypothesis, $H_1 : \beta 1 \neq 0$

```
Call:
lm(formula = writingscore ~ readingscore)

Residuals:
     Min       1Q   Median       3Q      Max
-11.0572  -2.8321   0.0116   2.6642  14.1758

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.45984    1.62522   0.283    0.778
readingscore  0.98941    0.02206  44.847   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.579 on 198 degrees of freedom
Multiple R-squared:  0.9104,    Adjusted R-squared:  0.9099
F-statistic:  2011 on 1 and 198 DF,  p-value: < 2.2e-16
```

Figure 3 shows the summary of regression test on female dataset

From the summary in figure 3, we can see that the linear regression equation for the female dataset for reading score and writing score is y = 0.45984 + 0.98941x which means the writing score as a dependent variable have a positive linear relationship with reading score as an independent variable. This means every 1 mark increase in reading score will cause the average writing score to increase by 0.98941. The R-squared indicates that 91.04% of the variation in writing score can be explained by reading score. This model is said to be statistically significant because the p-value is very close to zero which is smaller than the significant level which is 0.05. Thus, we reject the null hypothesis as there is sufficient evidence to support the claim that there exists a linear relationship between the two variables at a non-zero value.

```
Call:
lm(formula = writing ~ reading)

Residuals:
     Min       1Q   Median       3Q      Max
-11.3771  -3.1034   0.0997   2.9246  11.5764

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.40585    1.41565   0.993    0.322
reading      0.94391    0.02137  44.174   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.381 on 198 degrees of freedom
Multiple R-squared:  0.9079,    Adjusted R-squared:  0.9074
F-statistic:  1951 on 1 and 198 DF,  p-value: < 2.2e-16
```

Figure 4 shows the summary of regression test on male dataset

Meanwhile for the summary in figure 4, we can see that the linear regression for the male dataset for reading score and writing score is y = 1.40585 + 0.94391x which means the writing score as a dependent variable have a positive linear relationship with reading score as an independent variable. Every 1 mark increase in reading score will cause an average of 0.94391 mark increase in writing score. The R-squared shows that 90.79% of the variation in writing score can be explained by reading score. As the p-value is very close to zero which is smaller than the significant level which is 0.05, this model is also said to be statistically significant. Thus, we reject the null hypothesis as there is sufficient evidence to support the claim that there exists a linear relationship between the two variables, which is a non-zero value.

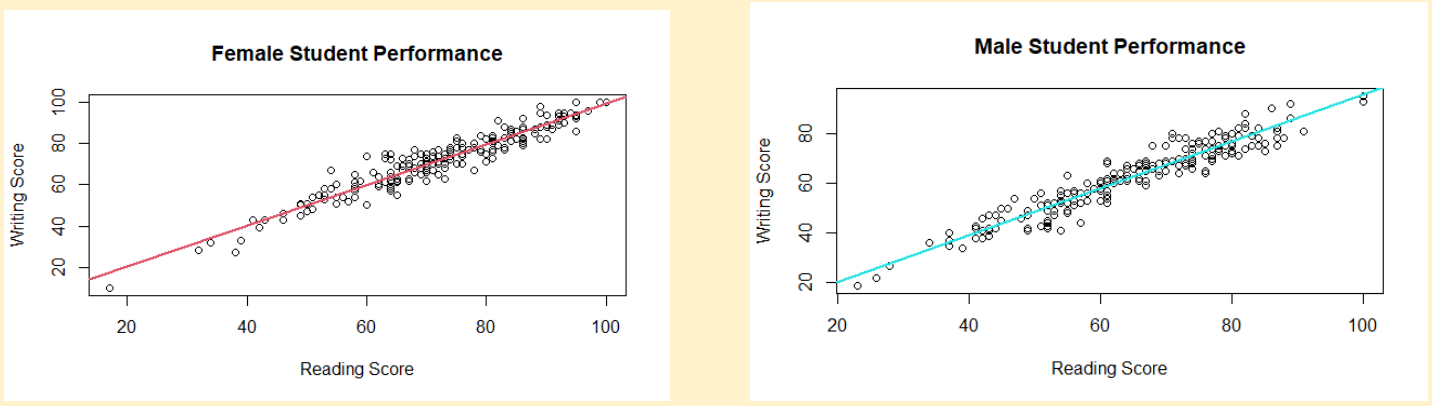


Figure 5 shows two scatter graphs for female and male dataset

Referring to figure 5, we can prove that females have a better performance than male. It is proven in the graph that female's scores scatter mostly at the range of 60-100 marks, meanwhile in the male graph, the plots scatter mostly at the range of 40-80 marks. We can see from the graphs that both have the similarity of having a positive strong linear relationship between the two variables as the points are closely scattered along the straight line and we can see there are no outliers in either graph.

Chi-Square Test of Independence

By using the same data set, a Chi-Square Test of Independence was conducted to determine whether there is a significant association between the gender of a student and their academic performance. The academic performance was determined by their total score which was the sum of their math score, reading score and writing score. The observed frequencies are presented in the following contingency table.

	total<=75	75< total <= 150	150 < total <= 225	total >225
Female	1	19	111	69
Male	1	35	108	56

Table 2 shows the observed frequencies of the marks of students by gender

Test hypothesis:

H_0 : The gender of a student and their academic performance is independent

H_1 : The gender of a student and their academic performance is not independent

Significance level:

$\alpha = 0.1$

The expected frequencies were calculated using the following formula and a contingency table was used to present the expected frequencies.

$$\text{Expected frequencies, } e_{ij} = \frac{(i^{th} \text{ Row Total})(j^{th} \text{ Column Total})}{\text{Total Sample Size}}$$

	total<=75	75< total <= 150	150 < total <= 225	total >225
Female	1	27	109.5	62.5
Male	1	27	109.5	62.5

Table 3 shows the expected frequencies of the marks of students by gender

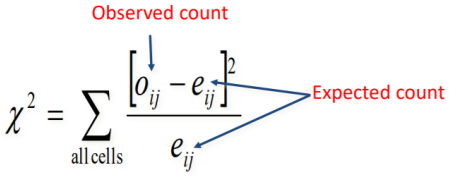
Since all the expected frequencies for the Chi-Square test should be larger or equal to 5, the column for "total <= 75" and "75 < total <= 150" is merged together into one column.

	Total <= 150	150 < total <= 225	total >225
Female	28	109.5	62.5
Male	28	109.5	62.5

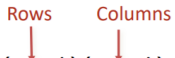
Table 4 shows the merged expected frequencies of the marks of students by gender

The test statistics of the Chi-Square test and degree of freedom were calculated using the following formula.

- Test Statistic:



- Degrees of Freedom: $(r - 1)(c - 1)$




```
> #perform chi-square test on the data
> chisq.test(ctbl, correct = FALSE)

        Pearson's Chi-squared test

data:  ctbl
X-squared = 5.9645, df = 2, p-value = 0.05068

>
> #critical value
> alpha <-0.1
> x2.alpha <- qchisq(alpha, df = 2, lower.tail =  FALSE)
> print(x2.alpha)
[1] 4.60517
```

Figure 6 shows the calculated test statistic and critical value from R-programming

The degree of freedom = (2 - 1) (3 - 1) = 2. By using R-programming, the value of test statistic is calculated, χ^2=5.964524 and the critical value for 2 degrees of freedom at 10% level is given by 4.60517.

As the value of the calculated test statistic, 5.9645 is greater than the critical value 4.6052, there is evidence to reject the null hypothesis, H_0 . Therefore, we can conclude that the gender of a student and their academic performance is not independent, there is an association between the gender of a student and their academic performance.

CONCLUSION

In conclusion, we learned that for every assumption or conclusion we want to make, it must be based on research or further calculation to avoid any uncertainty in data or bias. We decided to choose the student's performance dataset just simply because we want to learn whether the idea of females having better academic performance than male is true. Throughout the four testing that have been done, we can say that this assumption still needs further research because in the hypothesis testing, we can conclude that average female does not have better performance than male, but in correlation testing and regression testing, based on the scatter plot, we can say that female have better performance than male. We also learn in chi square testing that the academic performance of a student depends on their gender. In correlation, we agreed that math score has a linear correlation with reading score, meanwhile in regression, we proved that the writing score of a student depends on their reading score. This means that if a student can excel in their reading, they will absolutely excel in their writing too because reading can improve their writing skills.

APPENDICES

Link for video presentation: https://youtu.be/JBiobzLsl-g

Sample of raw data

	A	B	C	D	E	F	G	H
1	gender	race/ethnicit	parental lev	lunch	test prepara	math score	reading scor	writing score
2	female	group B	bachelor's d	standard	none	72	72	74
3	female	group C	some colleg	standard	completed	69	90	88
4	female	group B	master's deg	standard	none	90	95	93
5	male	group A	associate's d	free/reduced	none	47	57	44
6	male	group C	some colleg	standard	none	76	78	75
7	female	group B	associate's d	standard	none	71	83	78
8	female	group B	some colleg	standard	completed	88	95	92
9	male	group B	some colleg	free/reduced	none	40	43	39
10	male	group D	high school	free/reduced	completed	64	64	67
11	female	group B	high school	free/reduced	none	38	60	50
12	male	group C	associate's d	standard	none	58	54	52
13	male	group D	associate's d	standard	none	40	52	43
14	female	group B	high school	standard	none	65	81	73
15	male	group A	some colleg	standard	completed	78	72	70
16	female	group A	master's deg	standard	none	50	53	58
17	female	group C	some high s	standard	none	69	75	78
18	male	group C	high school	standard	none	88	89	86
19	female	group B	some high s	free/reduced	none	18	32	28
20	male	group C	master's deg	free/reduced	completed	46	42	46
21	female	group C	associate's d	free/reduced	none	54	58	61
22	male	group D	high school	standard	none	66	69	63
23	female	group B	some colleg	free/reduced	completed	65	75	70
24	male	group D	some colleg	standard	none	44	54	53
25	female	group C	some high s	standard	none	69	73	73
26	male	group D	bachelor's d	free/reduced	completed	74	71	80

Sample of processed data

	A	B	G	H	I	J
1	**Number**	**gender**	**math score**	**reading score**	**writing score**	**total score**
2	**1**	female	72	72	74	218
3	**2**	female	69	90	88	247
4	**3**	female	90	95	93	278
5	**4**	female	71	83	78	232
6	**5**	female	88	95	92	275
7	**6**	female	38	60	50	148
8	**7**	female	65	81	73	219
9	**8**	female	50	53	58	161
10	**9**	female	69	75	78	222
11	**10**	female	18	32	28	78
12	**11**	female	54	58	61	173
13	**12**	female	65	75	70	210
14	**13**	female	69	73	73	215
15	**14**	female	67	69	75	211
16	**15**	female	62	70	75	207
17	**16**	female	69	74	74	217
18	**17**	female	63	65	61	189
19	**18**	female	56	72	65	193
20	**19**	female	74	81	83	238
21	**20**	female	50	64	59	173
22	**21**	female	75	90	88	253
23	**22**	female	58	73	68	199