UTM

UNIVERSITI TEKNOLOGI MALAYSIA

SEM II 2023/2024

SECI1143-02

PROBABILITY AND STATISTICAL DATA ANALYSIS

PROJECT 2

Lecturer: Dr. Noorfa Haszlinna Mustaffa

MEMBERS NAME	MATRIC NO.
NAWWARAH AUNI BINTI NAZRUDIN	A23CS0143
NABIL AFLAH BOO BINTI MOHD YOSUF BOO YONG CHONG	A23CS0252
NUR FIRZANA BINTI BADRUS HISHAM	A23CS0156
LUBNA AL HAANI BINTI RADZUAN	A23CS0107

TABLE OF CONTENT

1.0 Introduction

The goal of this study is to understand the various factors that may affect students' performance at GS Ling Academy. We examine academic scores, weekly self-study hours, involvement in extracurricular activities, attendance records and part-time job status from the provided dataset. The aim is to identify the patterns using statistical analysis tests such as one- sample hypothesis tests, correlation and regression tests and chi square test of independence to gain insights into what helps students succeed in school.

We were interested in this data and the questionnaire because understanding what influences students to keep success is crucial. This analysis will be very helpful for educators or administrators to determine and support students facing challenges. With the growing emphasis on personalised learning, it is important to identify the key elements that contribute to a student's academic achievement and well-being.

Based on the data, we expect to see clear relationships between academic scores and any variables like attendance, self-study hours, participation in extracurricular activities and doing part-time jobs during their studies. Specifically, we predict that students with higher self-study hours and fewer absences would be more likely to achieve perfect academic scores. Additionally, we are curious to see if involvement in extracurricular activities has a positive or negative impact on academic performance and if students holding part-time jobs will affect the academic performance. This study will provide valuable information to improve the educational practices and students' support.

This project is conducted to achieve these objectives :

1. Apply and execute statistical test analysis on secondary data
2. Demonstrate whether the selected variables from the dataset are dependent on each other

2.0 Dataset

2.1 Description of Data

Population: Student academic scores in GS Ling Academy
Sample: 2000 students
Data description:

Variable (description)	Type of variable	Measurement level
id (student from 1 to 2000)	Quantitative	Nominal
gender (students sex either male or female)	Qualitative	Nominal
part_time_job (whether or not the students have part time job)	Qualitative	Nominal
absence_days (number of days the students did not attend to school)	Quantitative	Ratio
extracurricular_activities (whether or not the students have extracurricular activities in school)	Qualitative	Nominal
weekly_self_study_hours (number of hours the students self study per week)	Quantitative	Ratio
career_aspiration (students desire career to pursue)	Qualitative	Nominal
math_score (score for subject mathematic that each of the students obtain)	Quantitative	Interval
history_score (score for subject history that each of the students obtain)	Quantitative	Interval
physics_score (score for subject physics that each of the students obtain)	Quantitative	Interval
biology_score (score for subject biology that each of the students obtain)	Quantitative	Interval
english_score (score for subject english that each of the students obtain)	Quantitative	Interval
geography_score (score for subject geography that each of the students obtain)	Quantitative	Interval

2.2 Statistical Test Analysis

Selected Variables	Objectives	Test Analysis and Expected Outcome
part_time_job, math_score	Analyse the connection between students' part time job status and their mathematical score.	**Analysis:** 2 Sample Hypothesis Testing (Test on mean, variance unknown) **Expected outcome:** We may determine the relationship between part_time_job and math_score using a two-sample t-statistical test. If the t-statistical test yielded more than the t critical value, it indicates that there is a relationship between these two variables.
absence_days, physics_score	Identify the relation between students' absence days from school and their physics scores.	**Analysis:** Correlation Analysis **Expected outcome:** We can determine the strength of the link between absence_days and physics_score using correlation analysis tests. If the correlation coefficient, r is positive and close to one, it implies a strong positive linear relationship between absence_days and physics_score. It demonstrates that the number of absence days increases as the physics score rises.
weekly_self_study _hours, chemistry_score	Evaluate the association between students' self study hours per week and their chemistry score	**Analysis:** Simple Linear Regression **Expected outcome:** We get to measure the relationship between weekly_self_study_hours and chemistry_score. If the test statistical value for the regression is greater than the critical value, it will show that there are relationships between the students' self study hours per week and their chemistry score.
gender, part_time_job	Determine the relationship between students' gender and their part time job status	**Analysis:** Chi-Square Test of Independence **Expected outcome:** Using a chi-square test of independence, we can see if there is a relation between gender and part_time_job. If the chi-square test statistic value exceeds the chi-square critical value, it indicates that there is a link between the students' gender and their part-time job status.

3.0 Data Analysis

3.1 Hypothesis 2 Sample Test

We use hypothesis testing on mean with unknown variances on 2 variables, **part time job status** and **maths score** to see if there is any difference between the mean of maths score for people that have a part time jobs and the mean of maths score for people that don't have a part time jobs at 95% confidence level, and we assume that the variances for these two groups are not equal. From the dataset, frequency(n), means($\bar{x}$), standard deviations(s) and variances(s^2) are calculated. (Refer to figure 3.1.1 in appendix for Rstudio code)

Variables	Frequency, n	Mean, $\bar{x}$	Std dev, s	Variance, s^2
Have a part time job	316	78	16.9974	288.9105
Don't have a part time job	1684	85	12.1041	146.5090

Table 3.1.1 Frequency, mean, standard deviation and variances calculated

1) **Hypothesis Statements**

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

Where μ_1 is the mean for people that have a part time job and μ_2 is the mean for people that don't have a part time job.

2) **Test Statistics**

t_0 can be calculated using the formula : $T_0^* = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n1} + \frac{S_2^2}{n2}}}$

Using, α = 0.05, by using Rstudio : t-statistic, t_0 = -6.9955 (refer to figure 3.1.2)

3) **Degrees of freedom(v) and t-alpha($t_{\alpha/2, v}$)**

Using Rstudio (refer to figure 3.1.3 in appendix),
Degree of freedom(v) is calculated using the formula : $v = \frac{\left(\frac{S_1^2}{n1} + \frac{S_2^2}{n2}\right)^2}{\frac{\left(S_1^2/n1\right)^2}{n1-1} + \frac{\left(S_2^2/n2\right)^2}{n2-1}}$.

v = 377.1628 ≈ 377

Since it is two-tailed : $t_{\alpha/2, v} = t_{0.025, 377}$,

therefore we will reject H_0 if $t_0 < t_{0.025, 377}$ = -1.9663 or $t_0 > t_{0.025, 377}$ = 1.9663.

4) **Conclusion**

By referring to figure 3.1.4 and 3.1.5 in appendix, t_0 = -6.9955 < $t_{0.025, 377}$ = -1.9663, thus we reject the null hypothesis, H_0.

Thus, there is enough reason to conclude that the mean of maths score for people that have a part time job is different from the mean of maths score for people that don't have a part time job.

3.2 Correlation Test

In this analysis, we use the variables **absences_days** and **physics_scores** to determine whether there is a relationship between these two variables at a 95% confidence level using Pearson's Product-Moment Correlation Coefficient. Based on the analysis, we aim to understand if there is a statistically significant linear relationship between the number of days a student is absent and their score in physics, which could provide insights into how attendance affects the academic performance in this subject,

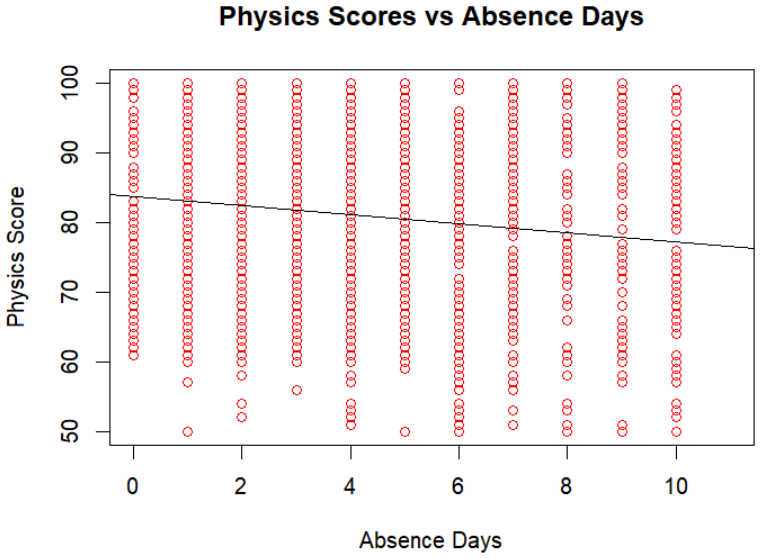


Figure 3.2.1 Visualise data in scatter plot using RStudio

According to the scatter plot above, there will be a weaker negative correlation relationship between absence days and physics score. The relationship states that the more days that students are absent, the lower their score in physics.

1) **Sample Correlation Coefficient using Pearson's method**

$$r = \frac{\sum xy - \left(\sum x \sum y\right)/n}{\sqrt{[(\sum x^2) - (\sum x)^2/n][(\sum y^2) - (\sum y)^2/n]}}$$

By using RStudio, the result of sample correlation coefficient, r is `-0.1364192`, which indicates that there is a relatively weak negative linear correlation between these variables.

2) **Significance Test**

- Hypothesis Statement

  H_0 : ρ = 0 (no linear correlation)

  H_1 : ρ ≠ 0 (linear correlation exists)

- Test Statistic

  t_0 can be calculated by using this formula:

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

  By using RStudio, the result state was t_0 = `-6.1553`.

- Critical Value

  Using α = 0.05 , df = (n-2) = 1998. From the t-table, there are two critical values. Critical value, $t_{\alpha/2=0.025,1998} = \pm 1.960$

- Decision

  Since t_0 = -6.1553 > $t_{\alpha/2=0.025,1998}$ = -1.960, we reject the null hypothesis, H_0 .

There is sufficient evidence of a linear relationship between the number of days a student is absent and students' physics score at the 5% level of significance.

3.3 Regression Test

In this analysis, we will examine whether **chemistry_score** depend on the number of **weekly_self_study_hours**. Using weekly_self_study_hours as the independent variable (x) and chemistry_score as the dependent variable (y), we will employ a simple linear regression model. We assume that changes of value in chemistry_score are influenced by changes in the amount of weekly_self_study_hours.

1. **Estimated Regression Model**

From RStudio, we get b_0 = 76.22904 and b_1 = 0.21210. Then, substitute the values of b_0 and b_1 into the regression model equation:

$$\widehat{y_i} = 76.229 + 0.212x$$

From the equation, we can interpret the values of the intercept coefficient, b_0 and the slope coefficient, b_1. In this data, $b_0 = 76.229$ represents the portion of the chemistry scores that is not explained by weekly self-study hours. Meanwhile, $b_1 = 0.212$ suggests that, on average, the chemistry scores increases by 0.212 points for each additional hour of weekly self-study.

Using RStudio, we also calculated the total sum of squares (SST) as 313155, the sum of squares error (SSE) as 13230.99, and the sum of squares regression (SSR) as 326386. The coefficient of determination (R^2) is 0.04054, indicating a weak linear relationship between weekly self-study hours and chemistry scores, with only 4.05% of the variation in chemistry scores explained by the variation in weekly self-study hours. Additionally, the standard error of the estimate (S_ε) is 12.52, suggests that the actual chemistry scores deviate from the predicted scores by an average of 12.52 points and the standard deviation of the slope (S_{b1}) is 0.02308 which is relatively small. This indicates that the estimate of the slope is quite precise and the positive relationship between study hours and chemistry scores observed in the sample is likely to be a true effect.

2. **Inference about the Slope Test: t-Test**

- **Hypothesis Statement:**
  H_0 : $\beta_1 = 0$ (no linear relationship)
  H_1 : $\beta_1 \neq 0$ (linear relationship does exist)

- **Critical Value:**
  Using significance level, $\alpha = 0.05$ and degree of freedom, df = 1998

Since this is two tailed test, two critical values we got from RStudio are

$-t_{0,025,\,1998} = -1.961152$

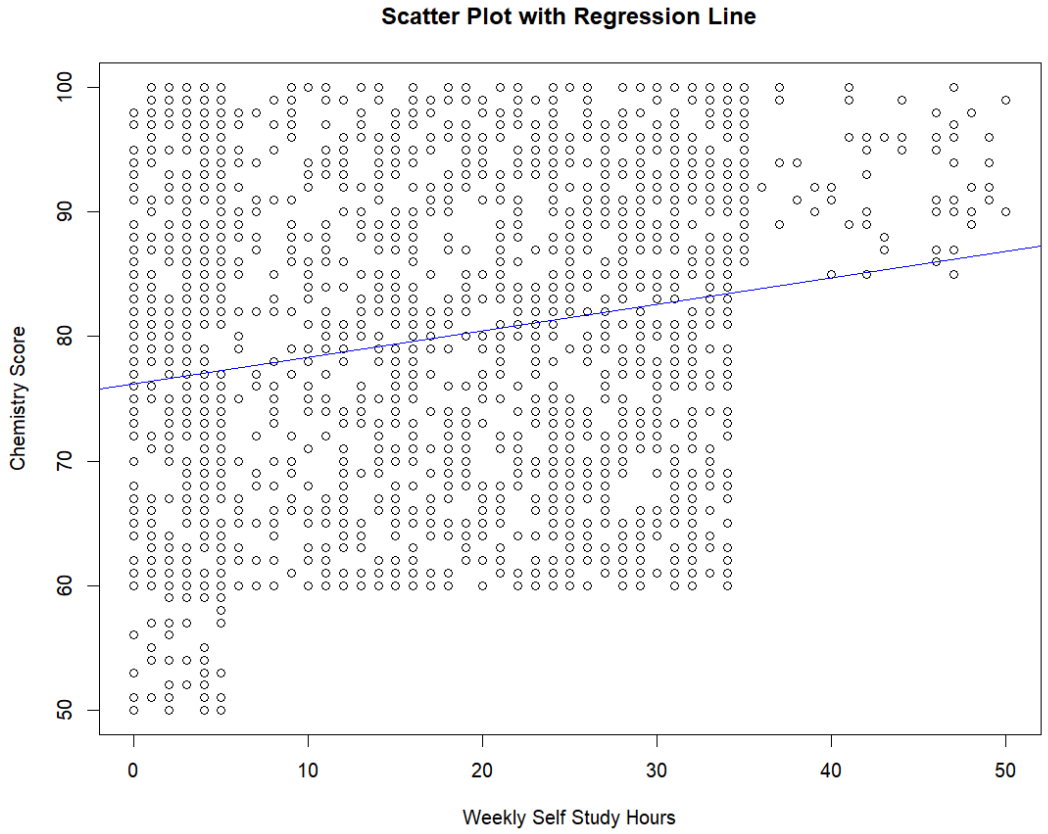
$t_{0,025,\,1998} = 1.961152$

Hence, we reject null hypothesis, H_0 if test statistics < -1.96 or test statistics > 1.96.

- Test Statistic
  Using RStudio, we get t = 9.188

- State the decision:
  Since $t = 9.188 > t_{0,025,\,1998} = 1.96$, we reject the null hypothesis, H_0. There is sufficient evidence that weekly self-study hours affect chemistry scores.

3. Scatter Plot with Regression Line



The line has a positive slope, which suggests that there is a positive correlation between the number of weekly self-study hours and chemistry scores. As study hours increase, chemistry scores tend to increase.

3.4 Chi-Square Test of Independence

To determine the independence of students' gender and their part time job status, we will utilise a two way contingency table with a 95% confidence level. As a consequence, we use the Chi-Square Test of Independence together with a two-way contingency table.

1. **State the hypothesis statement:**
   H_0: No relationship between variables.
   H_1: There is a relationship between variables.

2. **Find the critical value:**
   $\alpha = 0.05$
   Degree of freedom, df = (2-1)(2-1) = 1
   From the Figure 3.4.1, it shows that the critical value of $\chi^2_{k=1,\alpha=0.05}$ is 3.841459.

3. **Calculate the expected counts:**

gender	part_time_job				Total
	FALSE		TRUE		
	Obs.	Exp.	Obs.	Exp.	
female	835	$\frac{1002\times1684}{2000}$ = 843.684	167	$\frac{1002\times316}{2000}$ = 158.316	1002
male	849	$\frac{998\times168}{2000}$ = 840.316	149	$\frac{998\times316}{2000}$ = 157.684	998
Total	1684	1684	316	316	2000

4. **Calculate the test statistic value:**

   - Calculate manually

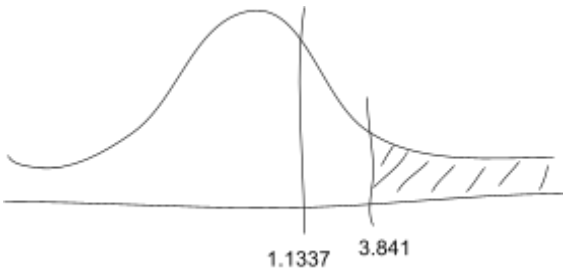
Cell, ij	Observed Count, o_{ij}	Expected Count, e_{ij}	$(o_{ij} - e_{ij})^2 / e_{ij}$
1,1	835	843.684	$(835-843.684)^2/843.684 = 0.089$
1,2	167	158.316	$(167-158.316)^2/158.316 = 0.476$
2,1	849	840.316	$(849-840.316)^2/840.316 = 0.090$
2,2	149	157.684	$(149-157.684)^2/157.684 = 0.478$
		$\chi^2 =$	1.133

Test statistic for chi-square value is 1.133 when we calculate it manually.

- By using RStudio

  By using RStudio, we obtain the test statistic value, χ^2 = 1.1337, and the p-value = 0.287.

5. **State the decision**:



Since, the test statistic value < critical value which is 1.1337 < 3.841, does not fall within the range of the critical region. Thus, we fail to reject the null hypothesis, H0. It shows that there is sufficient evidence to support the claim that there is no relationship between the variables gender and part_time_job, at α = 0.05.

4.0 Conclusion

This project provided us with valuable experience in the data analysis process, from selecting a dataset to performing complex statistical tests. We selected a comprehensive dataset from Kaggle, which included academic scores, self-study hours, extracurricular activities, attendance records, part-time job status, and other data.

Our analysis yielded several interesting findings. The hypothesis testing on the mean of math scores for students with and without part-time jobs revealed a statistically significant difference. We reject the null hypothesis and conclude that students with part-time jobs have significantly different math scores compared to those without part-time jobs. This finding suggests that part-time employment may have an impact on students' math performance,

For the correlation test analysis between absence days and physics scores indicated a weak negative correlation, $r = -0.136$, leading to the rejection of the null hypothesis at a 95% confidence level. This suggests that there is a statistically significant linear relationship between the number of days a student is absent and their physics scores. Specifically, more absences are associated with lower physics scores, highlighting the importance of regular attendance for academic success in physics.

Next, the simple linear regression analysis between weekly self-study hours and chemistry scores showed a weak positive relationship. The regression equation indicates that for each additional hour of weekly self-study, chemistry scores increase by an average of 0.212 points. We reject the null hypothesis. The line has a positive slope, which suggests that there is a positive correlation between the number of weekly self-study hours and chemistry scores. As study hours increase, chemistry scores tend to increase.

Lastly, the chi-square test of independence was conducted to determine the relationship between students' gender and their part-time job status. We fail to reject the null hypothesis. This indicates that there is no statistically significant relationship between gender and part-time job status at a 0.05 significance level. Therefore, gender does not appear to be a determining factor in whether a student holds a part-time job.

In summary, this project highlighted the complexity of factors influencing students' academic performance and emphasized the importance of helping students in various ways. By learning about the detailed connections between different factors and academic results, schools can better tailor their strategies to support student success and well-being.

We performed various test analysis, including two-sample hypothesis testing, correlation analysis, regression analysis, and the chi-square test of independence, using RStudio. This project greatly benefited our future by enhancing our data analysis skills. We would like to express our gratitude to our lecturer, Dr. Noorfa, for her assistance and direction throughout this project.

5.0 Appendix

1. **Presentation Video Link**
   https://youtu.be/rJJxLwXTGCE

2. **Reflection Link:**
   https://www.linkedin.com/posts/nawwarah-auni-262895299_reflection-for-probability-and-statistical-activity-7214171653918851072-0kdL?utm_source=share&utm_medium=member_desktop

3. **Processed data using Rstudio**
   - **3.1 Hypothesis Testing :**

```
> # separating part time data to true and false
> part_time_true <- project_2_dataset[project_2_dataset$part_time_job == TRUE, ]
> part_time_false <- project_2_dataset[project_2_dataset$part_time_job == FALSE, ]
> # Calculate mean xbar1 and xbar2
> xbar1_math_score <- round(mean(part_time_true$math_score, na.rm = TRUE))
> xbar1_math_score
[1] 78
> xbar2_math_score <- round(mean(part_time_false$math_score, na.rm = TRUE))
> xbar2_math_score
[1] 85
> # Calculate variance var1 and var2
> var1_math_score <- round(var(part_time_true$math_score, na.rm = TRUE), 4)
> var1_math_score
[1] 288.9105
> var2_math_score <- round(var(part_time_false$math_score, na.rm = TRUE), 4)
> var2_math_score
[1] 146.509
> # Calculate std dev s1 and s2
> s1_math_score <- round(sqrt(var1_math_score), 4)
> s1_math_score
[1] 16.9974
> s2_math_score <- round(sqrt(var2_math_score), 4)
> s2_math_score
[1] 12.1041
> # Calculate number of sample n1 and n2
> n1_part_time_true <- nrow(part_time_true)
> n1_part_time_true
[1] 316
> n2_part_time_false <- nrow(part_time_false)
> n2_part_time_false
[1] 1684
```

Figure 3.1.1 Calculating frequencies, means, standard deviations and variances

```
> alpha <- 0.05
> alpha
[1] 0.05
> t0 =  round((xbar1_math_score-xbar2_math_score-0)/
+       (sqrt((var1_math_score/n1_part_time_true)+(var2_math_score/n2_part_time_false))),
4)
> t0
[1] -6.9955
```

Figure 3.1.2 Test Statistic (t_0) Calculation

```
> v = round(((var1_math_score/n1_part_time_true)+(var2_math_score/n2_part_time_false))^2/
+     ((((var1_math_score/n1_part_time_true)^2)/(n1_part_time_true-1))+
+     (((var2_math_score/n2_part_time_false)^2)/(n2_part_time_false-1))), 4)
> v
[1] 377.1628
> t.alpha = round(qt(alpha/2, floor(v)), 4)
> t.alpha
[1] -1.9663
```

Figure 3.1.3 Degree of Freedom and t alpha ($t_{\alpha/2,\,v}$) Calculation

```
> cat("Critical Value : ", t.alpha, "\n")
Critical Value :  -1.9663
> cat("T-statistical : ", t0, "\n" )
T-statistical :  -6.9955
> cat("t0 = ", t0, " < talpha = ", t.alpha)
t0 =  -6.9955  < talpha =  -1.9663
```

Figure 3.1.4 Comparing test statistic with critical value

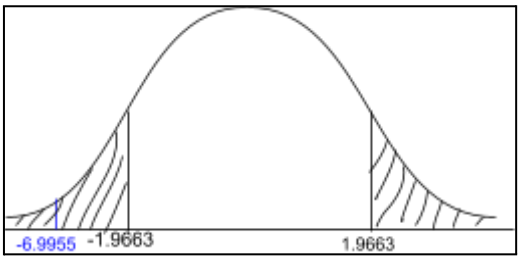


Figure 3.1.5 Decide the conclusion to reject H_0 or fail to reject H_0

- **3.2 Correlation Test :**

```
> t_stat <- (r/(sqrt((1-(r*r)/(n-2)))))
> cat("Test Statistic", t_stat)
Test Statistic -0.1364198
```

Figure 3.2.2 Test Statistic Calculation

```
> cor.test(x,y)

        Pearson's product-moment correlation

data:  x and y
t = -6.1553, df = 1998, p-value = 9.029e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.17917876 -0.09314522
sample estimates:
       cor
-0.1364192
```

Figure 3.2.3 Correlation Test using Pearson's method

```
> summary(model)

Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max
-33.07 -10.17   1.58  10.58  22.13

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  83.7213     0.4768 175.604  < 2e-16 ***
x            -0.6506     0.1057  -6.155 9.03e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.43 on 1998 degrees of freedom
Multiple R-squared:  0.01861,   Adjusted R-squared:  0.01812
F-statistic: 37.89 on 1 and 1998 DF,  p-value: 9.029e-10
```

Figure 3.2.4 Summary model for correlation test

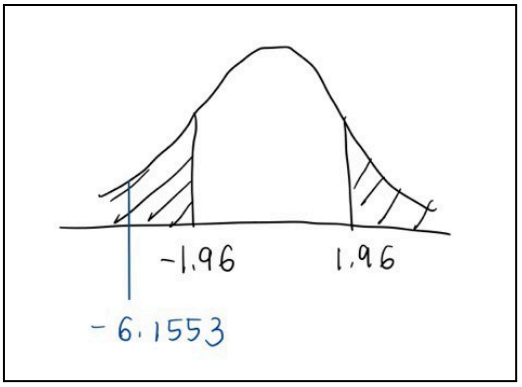


Figure 3.2.5 Decide the decision for correlation test

- **3.3 Regression Test :**

```
> summary(model)

Call:
lm(formula = chemistry_score ~ weekly_self_study_hours, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-27.2895 -10.9558   0.9378  10.3474  23.5589

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            76.22904    0.49636 153.576   <2e-16 ***
weekly_self_study_hours  0.21210    0.02308   9.188   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.52 on 1998 degrees of freedom
Multiple R-squared:  0.04054,   Adjusted R-squared:  0.04006
F-statistic: 84.42 on 1 and 1998 DF,  p-value: < 2.2e-16
```

Figure 3.3.1 Summary Model for Regression from RStudio

```
> # Print Least Squares Criterion
> cat("SSE:", sse, "\n")
SSE: 313155
> cat("SSR:", ssr, "\n")
SSR: 13230.99
> cat("SST:", sst, "\n")
SST: 326386
```

Figure 3.3.2 Least Squares Criterion for Regression from RStudio

```
> cat("-t", alpha / 2, ",", df, "=", critical_value_lower, "\n")
-t 0.025 , 1998 = -1.961152
> cat("t", alpha / 2, ",", df, "=", critical_value_upper, "\n")
t 0.025 , 1998 = 1.961152
```

Figure 3.3.3 Critical Value for Slope Test Inference from RStudio

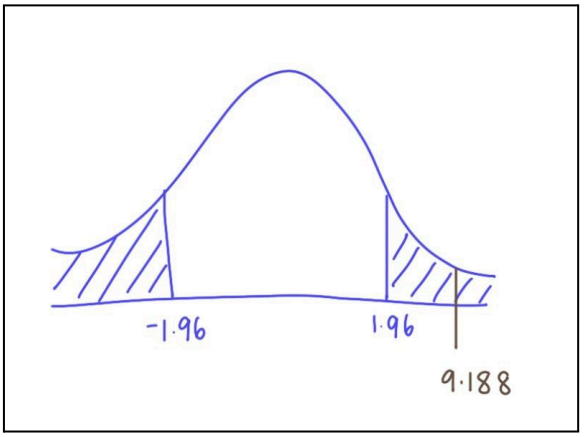


Figure 3.3.4 Decide the decision for inference for slope test

- **3.4 Chi-Square Test :**

```
> tbl = table(student_scores$gender, student_scores$part_time_job)
> tbl

        FALSE TRUE
  female   835  167
  male     849  149
```

Figure 3.4.1 Frequency of variables gender and part_time_job

```
> #find critical value
> alpha = 0.05
> df_tbl = 1
> tbl_alpha = qchisq(alpha, df=df_tbl, lower.tail = FALSE)
> print(tbl_alpha)
[1] 3.841459
```

Figure 3.4.2 Find critical value of Chi-Square using RStudio

```
> #perform chi-square test
> chisq.test(tbl, correct = FALSE)

        Pearson's Chi-squared test

data:  tbl
X-squared = 1.1337, df = 1, p-value = 0.287
```

Figure 3.4.3 Test Statistic for Chi-Square