UTM

UNIVERSITI TEKNOLOGI MALAYSIA

# PROBABILITY AND STATISTICAL DATA ANALYSIS

## SECI2143

## SECTION 01

## PROJECT 2

## LECTURER: DR SHARIN HAZLIN BINTI HUSPI

| NO | NAME | MATRIC NO |
|---|---|---|
| 1 | HUDA NAJIHAH BINTI AHMAD ASRI | A20EC0045 |
| 2 | INDIRA A/P THANGARAJ | A20EC0049 |
| 3 | MUHAMMAD AIMAN BIN ABDUL RAZAK | A20EC0082 |
| 4 | NUR HAZNIRAH BINTI HAZMAN | A20EC0114 |

Contents

## Introduction

Student performance is affected by most of factors. These includes gender, race, parental level of education, test preparation and lunch. The gender of the student is a factor to determine student’s performance. Gender differences in attitudes, personality, teacher’s expectation and behaviours, differential course taking and biological giving rise to gender difference in achievement (Feingold, 1988). Next, research has provided evidence that race and ethnicity continue to be important factors in explaining achievement differences whereas the parental level of education is a factor that considered as a key of the students to perform. In 1994, 13- and 17-year-olds whose parents had at least one or 4 more years of college had higher math and science proficiency scores than those whose parents did not finish high school. In the same study, parents' educational attainment was positively related to reading and writing scores as well (Young & Smith, 1997). Test preparation is very important because it is to make sure that we can practice and perform well in any examination or test. Other than that, lunch can affect the development of school aged children. One study discovered that 5th grade students who ate more fast food fared worse on math and reading scores (Li & O’Connell, 2012).

## Dataset

Figure 1 shows the raw dataset of student’s performance that is represented in a table consisting of the names, variables and data types represented in StudentsPerformance.csv file. The dataset contains information regarding the gender of students, race of students, level of education for the parents of the students, lunch status of students, test preparation course status of students, math score of students, reading score of students, and writing score of students. For this research, some of the variables were taken and compared to get adequate results. For the hypothesis testing two-sample, gender variable and math score variable were chosen to test if the mean for male students and female students are equal or not. Math score and reading score variables were chosen to conduct a correlation test to determine if the relationship between the students’ math scores and reading scores are related or not. For the regression test, variables math score and reading score were also chosen to conduct the test. Meanwhile, for Chi Square test of Independence, the parental level of education and the test preparation course were chosen to see if the variables are independent of each other. Other than that, for ANOVA test, we used the math score and race/ethnicity to test whether the different group of race/ethnicity have the same mean or not.

| NAME | VARIABLE | DATA TYPE |
|---|---|---|
| gender | gender | Nominal |
| race.ethnicity | race/ethnicity | Nominal |
| parental.level.of.education | parental level of education | Nominal |
| lunch | lunch | Nominal |
| test.course | test preparation course | Nominal |
| math.score | math score | Ratio |
| reading.score | reading score | Ratio |
| writing.score | writing score | Ratio |

*Figure 1 The name, variable and data type of StudentsPerformance.csv*

## Data Analysis

### Hypothesis Testing Two-sample

We would like to determine whether there will be a difference between the mean of math scores for male students and the mean of math scores for female students under 95% significance level in t-test. The number of male students is 482 while the number of female students is 518. The mean value of math scores for male students (rounded to 3 decimal places) is 68.728 while the mean value of math scores for female students (rounded to 3 decimal places) is 63.633. The standard deviation value of math scores for male students (rounded to 3 decimal places) is 14.356 while the standard deviation value of math scores for female students (rounded to 3 decimal places) is 15.491. We assume unequal variance. The variables used in this test are gender (gender of the students) and math.score (math score of the students).

$\mu_1$ is the mean of math scores for male students

$\mu_2$ is the mean of math scores for female students

$H_0$: $\mu_1 = \mu_2$

$H_1$: $\mu_1 \neq \mu_2$

We will reject $H_0$ if:

$t_0 < t_{0.025,997} = -1.9623$

or if $t_0 > t_{0.025,997} = 1.9623$

Since $t_0 = 5.398 > t_{0.025,40} = 1.9623$, we reject the null hypothesis. There is sufficient evidence to conclude that the mean of math scores for male students is not equal to the mean of math scores for female students. This tells us that genders play a role in how much students will score in math.

### Correlation

Based on this dataset of random sample of 1000 students, we are going to check the linear relationship between the math score and reading score at the significance level of 0.05. The variable used in this correlation test is the math score (math score obtained by the students) and reading score (reading score obtained by the students).

The null hypothesis, $H_0$: $\rho = 0$, there is no linear correlation between the math score and reading score.

The alternative hypothesis, $H_A$: $\rho \neq 0$, the linear correlation exists between the math score and reading score.

Based on the result from the R code, we can conclude that, the $t_0 = 44.855 > t_{0.025,\ 1000} = 2.021075$ and the null hypothesis, $H_0$ is rejected. Hence there is sufficient evidence that the linear correlation exists between the math score and reading score at the 0.05 or 5% significance level.

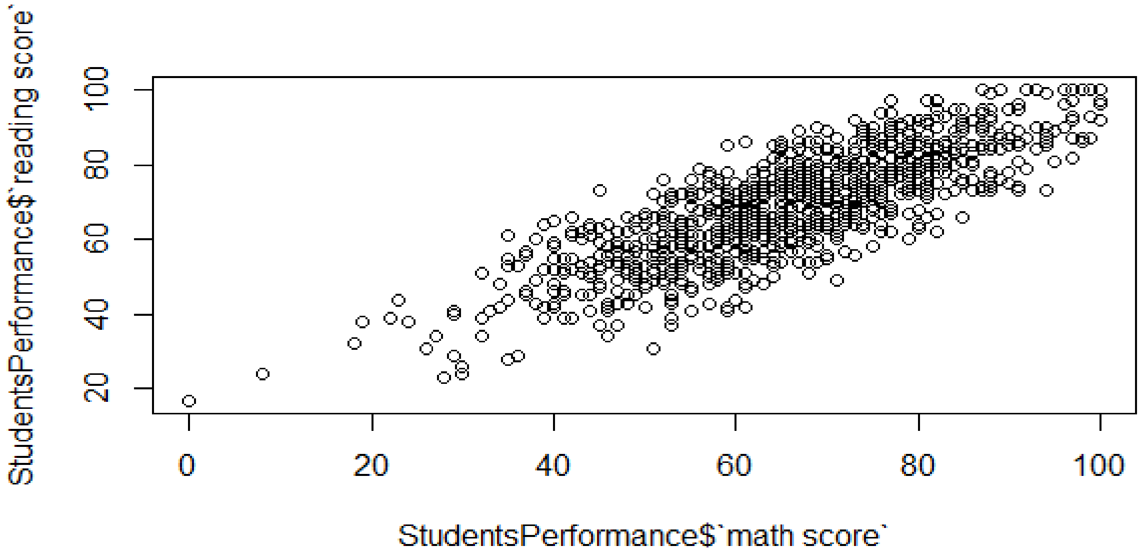


*Figure 2 Scatter plot of reading score and math score*

From the graph, we can see that the math score increases as the reading score increases. The scatter plot and correlation point out that there is a positive relationship between the math score and reading score. The R value is 0.8175797 which exhibits that there is a strong positive between the variables tested which are the math score and reading score.

## Regression

A random sample of 1000 students for this test have been selected and we want to check if math scores can predict reading scores at the 0.05 level of significance. The dependent variable (*y*) is reading score while the independent variable (*x*) is math score. The goal of this test is to see if reading score and math score have a linear relationship.

The null hypothesis, $H_0$: $\beta_1 = 0$

The alternative hypothesis, $H_1$: $\beta_1 \neq 0$

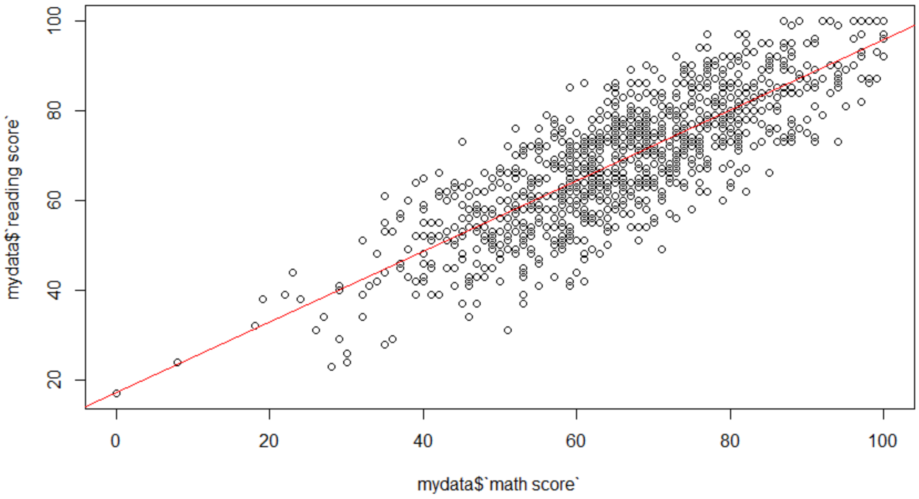


*Figure 3 Scatter plot and regression line of reading score against math score*

Through regression graph analysis from Figure 3, we can anticipate that there is a positive linear relation between math score and reading score with the least squares equation of ($y$ = 17.1418 + 0.7872$x$). So here, $\beta_0$ = 17.1418 indicate the estimated average value of reading score when the value of the math score is 0. It implies that for the reading score to be in the observed range 17.1418 is the proportion of the reading score not explained by the math score. The value of $\beta_1$ = 0.7872 implies that the average reading score increases by 0.7872 on average for each increasing on math score. As a result of our findings, we can conclude that reading score and math score have a positive linear relationship.

Coefficient of determination, $R^2 = 0.6684$

66.84% of the variation in reading score is explained by variation in the math score.

t-test value = 44.85

Based on the result, the degree of freedom is 998. As the p-value is 2e-16 = 0.0000000000000002 < 0.05, we reject the null hypothesis, $H_0$. Hence, there is a sufficient evidence of a positive linear relationship between reading score and math score at 0.05 significance level.

## Chi Square Test of Independence

A random sample of 1000 students were chosen to see if there exists any relationship between the two variables – Parental level of education, and Test preparation course at 0.05 significance level. In this test, we chose the variables 'parental level of education' which include Associate's Degree. Bachelor's Degree, High School, Master's Degree, Some College, and Some High School, together with 'test preparation course' – Completed, and None.

| Parental's level of education | Test preparation course | |
|---|---|---|
| | **Completed** | **None** |
| Associate's Degree | 82 | 140 |
| Bachelor's Degree | 46 | 72 |
| High School | 56 | 140 |
| Master's Degree | 20 | 39 |
| Some College | 77 | 149 |
| Some High School | 77 | 102 |

*Figure 4 Number of students based on test preparation course according to parental's level of education*

H$_0$: Parental level of education is independent of the test preparation course

H$_1$: Parental level of education and the test preparation course is dependent from each other

From the test, the result showed that the chi-square value ($x^2$) = 9.5441, and the critical value ($x^2_{k = 5, \alpha = 0.05}$) = 11.0705. Since $x^2 = 9.5441 < x^2_{k = 5, \alpha = 0.05} = 11.0705$, thus we fail to reject H$_0$ at $\alpha = 0.05$. We can conclude that the parental level of education and the test preparation is independent as there is sufficient evidence to support the claim.

## **ANOVA**

Based on the dataset of 1000 random students, we test whether the math score to different group of race/ethnicity have the same mean or not using a 0.05 significance level. In this test, we are using the variable Math score which has various numerical data from 0-100, and variable Race/ethnicity which includes Group A, Group B, Group C, Group D and Group E.

| Math score according to race/ethnicity | N | Mean |
|---|---|---|
| Group A | 89 | 61.63 |
| Group B | 190 | 63.45 |
| Group C | 319 | 64.46 |
| Group D | 262 | 67.36 |
| Group E | 140 | 73.82 |
| **Total** | **1000** | **66.144** |

*Figure 5 Number and mean of math score according to race/ethnicity*

H$_0$: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

H$_1$: At least one of the mean is different

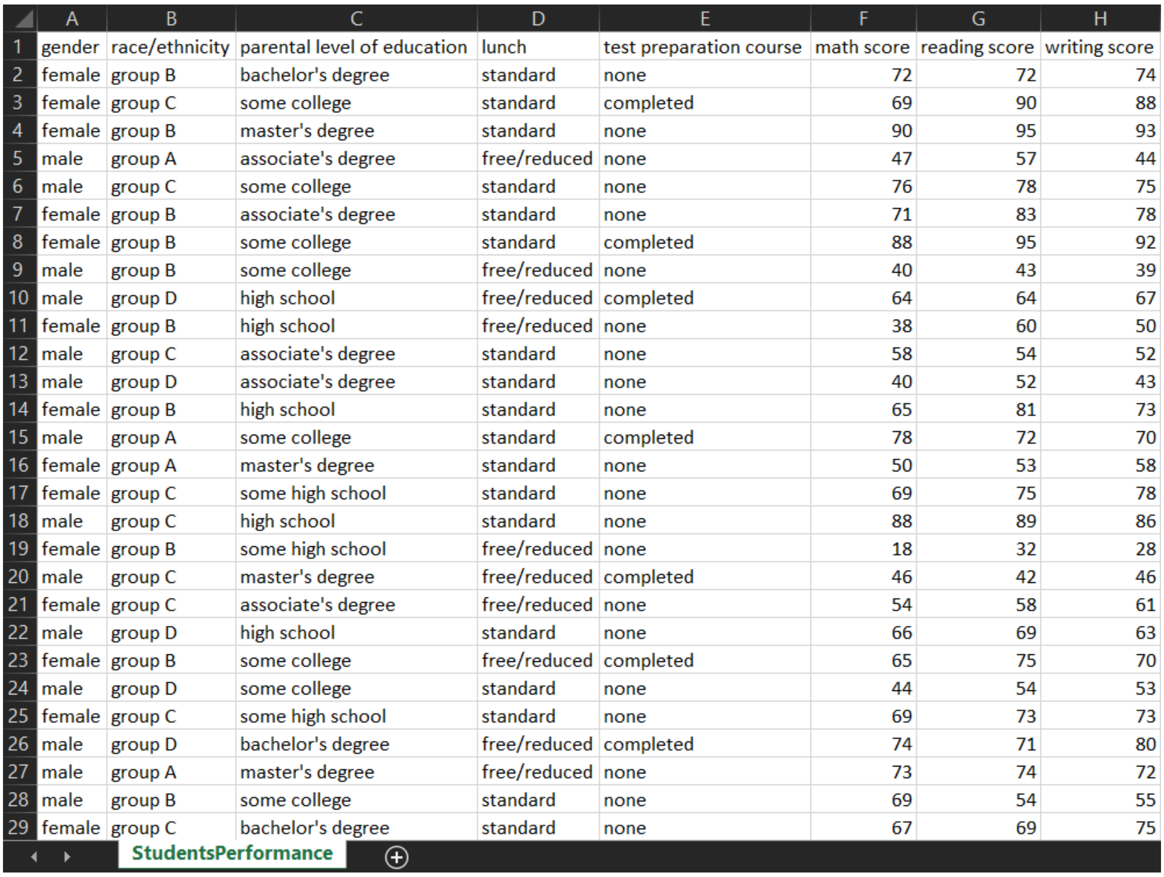
From the test, the result that we got showed that the degree of freedom for numerator = 4 while degree of freedom denominator = 995 and the *F* value is 14.59, thus *F (*4, 995) = 14.59. The *P*-value is 1.37*e*-11 = 0.0000000000137 < 0.05. Thus, we reject the $H_0$. In conclusion, there is sufficient evidence that at least one of the mean that is different.

## Conclusion

In a nutshell, we can conclude that scores of students can affected by many possible reasons. From the study conducted above, we get to know that gender is one of the factors that affects the score of the students. Not only that, we also get to know that there is a linear correlation exists between the math score and the reading score. The parental level of education does not affect the test preparation. Finally, from the study conducted above we can conclude that there is at least one mean that is different and the ethnicity affects the score obtained by the students. In conclusion, we have learned from this study, we have learned that using inference statistical analysis and proper tools such as R-language. We learned many new techniques like using the R-language, choosing an appropriate dataset, pre-processing the data and using proper analysis test or methods such as hypothesis testing, correlation analysis, regression analysis, chi square test of independence and ANOVA test to make inference in statistical analysis.

# Appendix

Sample of dataset



| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | gender | race/ethnicity | parental level of education | lunch | test preparation course | math score | reading score | writing score |
| 2 | female | group B | bachelor's degree | standard | none | 72 | 72 | 74 |
| 3 | female | group C | some college | standard | completed | 69 | 90 | 88 |
| 4 | female | group B | master's degree | standard | none | 90 | 95 | 93 |
| 5 | male | group A | associate's degree | free/reduced | none | 47 | 57 | 44 |
| 6 | male | group C | some college | standard | none | 76 | 78 | 75 |
| 7 | female | group B | associate's degree | standard | none | 71 | 83 | 78 |
| 8 | female | group B | some college | standard | completed | 88 | 95 | 92 |
| 9 | male | group B | some college | 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 degree | standard | none | 58 | 54 | 52 |
| 13 | male | group D | associate's degree | standard | none | 40 | 52 | 43 |
| 14 | female | group B | high school | standard | none | 65 | 81 | 73 |
| 15 | male | group A | some college | standard | completed | 78 | 72 | 70 |
| 16 | female | group A | master's degree | standard | none | 50 | 53 | 58 |
| 17 | female | group C | some high school | standard | none | 69 | 75 | 78 |
| 18 | male | group C | high school | standard | none | 88 | 89 | 86 |
| 19 | female | group B | some high school | free/reduced | none | 18 | 32 | 28 |
| 20 | male | group C | master's degree | free/reduced | completed | 46 | 42 | 46 |
| 21 | female | group C | associate's degree | free/reduced | none | 54 | 58 | 61 |
| 22 | male | group D | high school | standard | none | 66 | 69 | 63 |
| 23 | female | group B | some college | free/reduced | completed | 65 | 75 | 70 |
| 24 | male | group D | some college | standard | none | 44 | 54 | 53 |
| 25 | female | group C | some high school | standard | none | 69 | 73 | 73 |
| 26 | male | group D | bachelor's degree | free/reduced | completed | 74 | 71 | 80 |
| 27 | male | group A | master's degree | free/reduced | none | 73 | 74 | 72 |
| 28 | male | group B | some college | standard | none | 69 | 54 | 55 |
| 29 | female | group C | bachelor's degree | standard | none | 67 | 69 | 75 |

*Figure 6 shows 29 out of 1000 from the Students Performance dataset*

## Reference

Gooding, Y. (n.d.). *The relationship between parental educational level and academic success of college freshmen*. Retrieved from website: https://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=1428&context=rtd

Five key trends in U.S. student performance: Progress by blacks and Hispanics, the takeoff of Asians, the stall of non-English speakers, the persistence of socioeconomic gaps, and the damaging effect of highly segregated schools. (2017). Retrieved June 28, 2021, from Economic Policy Institute website: https://www.epi.org/publication/five-key-trends-in-u-s-student-performance-progress-by-blacks-and-hispanics-the-takeoff-of-asians-the-stall-of-non-english-speakers-the-persistence-of-socioeconomic-gaps-and-the-damaging-effect/#epi-toc-1

Desimone, L. (1999). Linking Parent Involvement with Student Achievement: Do Race and Income Matter? *The Journal of Educational Research*, *93*(1), 11–30. Retrieved from https://www.jstor.org/stable/27542243

Research, W. (2014). *Nutrition and Students' Academic Performance*. Retrieved from website: https://www.wilder.org/sites/default/files/imports/Cargill_lit_review_1-14.pdf

Factors Which Influence The Students Academic Performance. (2015). Retrieved June 28, 2021, from UKEssays.com website: https://www.ukessays.com/essays/education/factors-which-influence-the-students-academic-performance-education-essay.php

My Instant Essay | Factors that affects student's academic performance. How do we improve student learning on campus? (2011). Retrieved June 28, 2021, from Myinstantessay.com website: https://myinstantessay.com/sample/psychology/factors-that-affects-students-academic-performance