RProject

Nagma Pokharel

4/28/2020

## Summary

The researcher tries to address if there is a real difference in the math score between male and female. This is important to know because there is various theories and belief that females can outperform in reading score better while males can outperform in math. This also might be due to various socio-cultural factor. The researcher would like to check whether there is difference between the math and reading score of both male and female.Male compared to female have different priorities when they choose their career which is also a reason why some researchers claim male being more effective in maths score in exmas.

The following would be the research question that the researcher would like to ask:

1. What are the average number of male and female math score?
2. Show if it is possible to test the differences using another dependable variable?
3. Do the math and reading scores differ according to gender? If yes/ no, why?

The researcher believes that females are more proficient in reading condition while males are more proficient in math. The researcher mentions such comment because there has been theoretical evidences and news on this problem. Hence, the researcher would like to purpose that the mean of math score is different between male and female.

The researcher found their data in Kaggle.com. The analysis of data was with first being approaching the data. Second analyzing categorical and numerical values. Third, finding data types, outliers and missing values. The researcher also analyzed the data and discovered various methods to find the accuracy in the results from the data.

The basic findings of the data is that the researcher’s assumption was right. Based on the gender, male score better than female in math exam while female score better in reading exam compared to that with male. Based on the researcher’s findings, researcher would like to recommend that there have been substantial change in the gap between male and female math score. There still lies huge gap however, it has been decreasing with time. Another interesting finding is also about the income differences that has been resulted with the gap in the math score of both male and female. There has been substantial increase in the pay of women as well. Nevertheless, we cannot eliminate the fact that the gap is still there.

# Introduction

The researcher explains the problem of this research to be finding the difference between math and reading scores in both males and females. I would like to find out if there are certain differences in the score as males are more into technical subjects and females are in theoretical. The researcher does not aim to find anything associated with the score but only aims to compare how does the difference between males and females differs and what could be the reason. The math and reading score could also differ with the preparation the students might have made. However, it is also about how many students have registered. My data set involves 482 males and 518 females. The researcher also aims to find if the math score of female/male is low, why is so?

The researcher views the importance of this problem to measure the independent variables like gender, race/ethnicity and parental level of education to determine the dependable variable i.e. math. The researcher would also like to go ahead and develop two hypotheses one involving math score and the other involving the reading score. With the availability, the researcher aims to check what is the reason for the score increment in both males and females.

# Literature review

Education has been the most important discipline in human life. Race and ethnicity have also been important to determine how gender differs the score of math and reading.

Some researches show that males are more adaptable on most measures of quantitative and visuospatial ability and the reason behind it reasons why males are often more variable remain obscure which is still a problem in the world. Good careers in math require many types of cognitive abilities while various socio-cultural forces also contribute to the results.

However, due to more awareness, we can see changes in the society but still we yet to see the data which could determine whether female outstands in reading test and math test or vice versa.

# Theory

The researcher would like to analyze two hypotheses where one involves writing as a dependent variable and the other involves reading as a dependent variable.

Hypothesis number 1:

H0: µ0: The mean math score of male and female are not different from each other.

H1: µ1: The mean math score of male and female are different from each other.

Hypothesis number 2:

H0: µ0: The mean reading score of male and female are not different from each other.

H1: µ1: The mean reading score of male and female are different from each other.

# Data

## R Markdown

library(stringr)  
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 3.6.3

## Warning: package 'tibble' was built under R version 3.6.3

## Warning: package 'tidyr' was built under R version 3.6.3

## Warning: package 'readr' was built under R version 3.6.3

## Warning: package 'dplyr' was built under R version 3.6.3

## Warning: package 'forcats' was built under R version 3.6.3

library(lubridate)

## Warning: package 'lubridate' was built under R version 3.6.3

library(dplyr)  
library(readr)  
library(mltools)

## Warning: package 'mltools' was built under R version 3.6.3

library(data.table)

## Warning: package 'data.table' was built under R version 3.6.3

###### Response variable is math.score

###### Read Data

sp = read.csv('C:/Users/nagma/Desktop/Rproject/StudentsPerformance.csv')  
head(sp)

## gender race.ethnicity parental.level.of.education lunch  
## 1 female group B bachelor's degree standard  
## 2 female group C some college standard  
## 3 female group B master's degree standard  
## 4 male group A associate's degree free/reduced  
## 5 male group C some college standard  
## 6 female group B associate's degree standard  
## test.preparation.course math.score reading.score writing.score  
## 1 none 72 72 74  
## 2 completed 69 90 88  
## 3 none 90 95 93  
## 4 none 47 57 44  
## 5 none 76 78 75  
## 6 none 71 83 78

###### number of rows in sp

###### number of columns in sp

len\_cols = length(names(sp))  
print(c("number of rows:", nrow(sp)))

## [1] "number of rows:" "1000"

print(c("number of columns:", len\_cols))

## [1] "number of columns:" "8"

head(sp)

## gender race.ethnicity parental.level.of.education lunch  
## 1 female group B bachelor's degree standard  
## 2 female group C some college standard  
## 3 female group B master's degree standard  
## 4 male group A associate's degree free/reduced  
## 5 male group C some college standard  
## 6 female group B associate's degree standard  
## test.preparation.course math.score reading.score writing.score  
## 1 none 72 72 74  
## 2 completed 69 90 88  
## 3 none 90 95 93  
## 4 none 47 57 44  
## 5 none 76 78 75  
## 6 none 71 83 78

Response variable has NA values, remove the data when response variables have na values

sp = sp[!is.na(sp$math.score),]  
print(c("number of rows:", nrow(sp)))

## [1] "number of rows:" "1000"

###### Fill na values for Reading score by using median values for Male and Female

###### median of female reading score

med\_female\_rs = median(sp$reading.score[sp$gender=='female'],na.rm=TRUE)

###### median for male reading score

med\_male\_rs = median(sp$reading.score[sp$gender=='male'], na.rm=TRUE)  
  
sp$reading.score[is.na(sp$reading.score) & (sp$gender == "female")] <-med\_female\_rs  
sp$reading.score[is.na(sp$reading.score) & (sp$gender == "male")] <-med\_male\_rs

# Do the same for the writing Score

med\_female\_ws = median(sp$writing.score[sp$gender=='female'],na.rm=TRUE)

###### median for male reading score

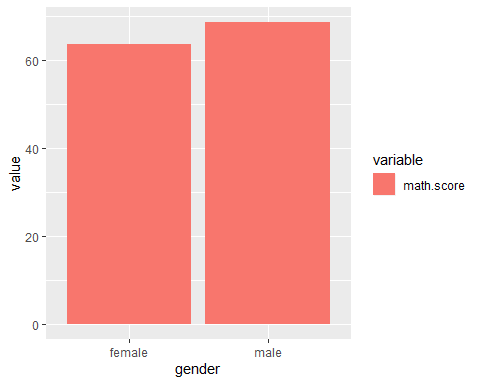
med\_male\_ws = median(sp$writing.score[sp$gender=='male'], na.rm=TRUE)  
  
sp$writing.score[is.na(sp$writing.score) & (sp$gender == "female")] <-med\_female\_ws  
sp$writing.score[is.na(sp$writing.score) & (sp$gender == "male")] <-med\_male\_ws

###### check if there is a difference of mean between girls and boys math score

gend <- sp%>% select(gender, math.score)  
gend <- gend %>% group\_by(gender) %>% summarise\_all(mean)  
gend <- data.frame(gend)  
d <- melt(gend)

## Warning in melt(gend): The melt generic in data.table has been passed a  
## data.frame and will attempt to redirect to the relevant reshape2 method;  
## please note that reshape2 is deprecated, and this redirection is now  
## deprecated as well. To continue using melt methods from reshape2 while both  
## libraries are attached, e.g. melt.list, you can prepend the namespace like  
## reshape2::melt(gend). In the next version, this warning will become an error.

ggplot(data = d,mapping = aes(x = gender,   
 y = value, fill = variable)) +   
 geom\_col(position = position\_dodge())



###### check if the difference betwween the mean of math.score is significant

t.test(math.score ~ gender, data = sp, paired = FALSE)

##   
## Welch Two Sample t-test  
##   
## data: math.score by gender  
## t = -5.398, df = 997.98, p-value = 8.421e-08  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -6.947209 -3.242813  
## sample estimates:  
## mean in group female mean in group male   
## 63.63320 68.72822

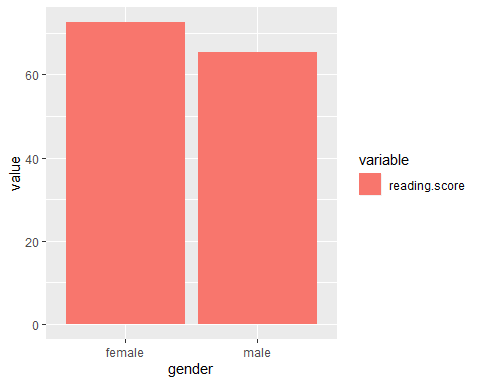
###### p-value is much less than 0.05, hence the different statistically significant

###### check if there is a difference of mean between girls and boys reading score

gend <- sp%>% select(gender, reading.score)  
gend <- gend %>% group\_by(gender) %>% summarise\_all(mean)  
gend <- data.frame(gend)  
d <- melt(gend)

## Warning in melt(gend): The melt generic in data.table has been passed a  
## data.frame and will attempt to redirect to the relevant reshape2 method;  
## please note that reshape2 is deprecated, and this redirection is now  
## deprecated as well. To continue using melt methods from reshape2 while both  
## libraries are attached, e.g. melt.list, you can prepend the namespace like  
## reshape2::melt(gend). In the next version, this warning will become an error.

ggplot(data = d,mapping = aes(x = gender,   
 y = value, fill = variable)) +   
 geom\_col(position = position\_dodge())

 ###### check if the difference betwween the mean of reading.scoreis significant

t.test(reading.score ~ gender, data = sp, paired = FALSE)

##   
## Welch Two Sample t-test  
##   
## data: reading.score by gender  
## t = 7.9684, df = 996.36, p-value = 4.376e-15  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 5.377941 8.892218  
## sample estimates:  
## mean in group female mean in group male   
## 72.60811 65.47303

###### p-value is much less than 0.05, hence the different statistically significant

###### One hot encode the categorical variables for mOdeling in R

table\_sp <- one\_hot(data.table(sp), dropCols = TRUE)

###### Convert back to data frame

sp = data.frame(table\_sp)

###### Print the correlation between all the variables with the response variable

for (i in names(sp)){  
 correlation\_res <- cor(sp[[i]], sp$math.score)  
 print(c("corr with respose of", i, correlation\_res))  
}

## [1] "corr with respose of" "gender\_female" "-0.167982238100356"   
## [1] "corr with respose of" "gender\_male" "0.167982238100356"   
## [1] "corr with respose of" "race.ethnicity\_group.A" "-0.0919770999021443"   
## [1] "corr with respose of" "race.ethnicity\_group.B" "-0.0842500599939814"   
## [1] "corr with respose of" "race.ethnicity\_group.C" "-0.0733868729013714"   
## [1] "corr with respose of" "race.ethnicity\_group.D" "0.0500707135088943"   
## [1] "corr with respose of" "race.ethnicity\_group.E" "0.205854580946756"   
## [1] "corr with respose of"   
## [2] "parental.level.of.education\_associate.s.degree"  
## [3] "0.0632281684190594"   
## [1] "corr with respose of"   
## [2] "parental.level.of.education\_bachelor.s.degree"  
## [3] "0.0796635696072283"   
## [1] "corr with respose of"   
## [2] "parental.level.of.education\_high.school"  
## [3] "-0.128725262522626"   
## [1] "corr with respose of"   
## [2] "parental.level.of.education\_master.s.degree"  
## [3] "0.0604168113742921"   
## [1] "corr with respose of"   
## [2] "parental.level.of.education\_some.college"  
## [3] "0.0370563175226976"   
## [1] "corr with respose of"   
## [2] "parental.level.of.education\_some.high.school"  
## [3] "-0.0798518764152474"   
## [1] "corr with respose of" "lunch\_free.reduced" "-0.35087664559186"   
## [1] "corr with respose of" "lunch\_standard" "0.35087664559186"   
## [1] "corr with respose of" "test.preparation.course\_completed"  
## [3] "0.177702469304395"   
## [1] "corr with respose of" "test.preparation.course\_none"  
## [3] "-0.177702469304395"   
## [1] "corr with respose of" "math.score" "1"   
## [1] "corr with respose of" "reading.score" "0.817579663672054"   
## [1] "corr with respose of" "writing.score" "0.802642045949808"

###### Build a simple regression model

linearMod <- lm(math.score ~ ., sp)  
print(summary(linearMod))

##   
## Call:  
## lm(formula = math.score ~ ., data = sp)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -17.4995 -3.6824 0.1218 3.3932 14.1178   
##   
## Coefficients: (5 not defined because of singularities)  
## Estimate Std. Error t value  
## (Intercept) 13.98079 1.08470 12.889  
## gender\_female -13.24045 0.37193 -35.599  
## gender\_male NA NA NA  
## race.ethnicity\_group.A -5.07770 0.73714 -6.888  
## race.ethnicity\_group.B -4.24233 0.60419 -7.022  
## race.ethnicity\_group.C -4.89947 0.54748 -8.949  
## race.ethnicity\_group.D -4.97930 0.57300 -8.690  
## race.ethnicity\_group.E NA NA NA  
## parental.level.of.education\_associate.s.degree -0.55216 0.54989 -1.004  
## parental.level.of.education\_bachelor.s.degree -1.59906 0.65701 -2.434  
## parental.level.of.education\_high.school 0.01557 0.55859 0.028  
## parental.level.of.education\_master.s.degree -2.40823 0.82760 -2.910  
## parental.level.of.education\_some.college -0.15190 0.54678 -0.278  
## parental.level.of.education\_some.high.school NA NA NA  
## lunch\_free.reduced -3.21271 0.37420 -8.585  
## lunch\_standard NA NA NA  
## test.preparation.course\_completed -3.50227 0.39658 -8.831  
## test.preparation.course\_none NA NA NA  
## reading.score 0.26351 0.04205 6.266  
## writing.score 0.70156 0.04352 16.120  
## Pr(>|t|)   
## (Intercept) < 2e-16 \*\*\*  
## gender\_female < 2e-16 \*\*\*  
## gender\_male NA   
## race.ethnicity\_group.A 1.00e-11 \*\*\*  
## race.ethnicity\_group.B 4.08e-12 \*\*\*  
## race.ethnicity\_group.C < 2e-16 \*\*\*  
## race.ethnicity\_group.D < 2e-16 \*\*\*  
## race.ethnicity\_group.E NA   
## parental.level.of.education\_associate.s.degree 0.3156   
## parental.level.of.education\_bachelor.s.degree 0.0151 \*   
## parental.level.of.education\_high.school 0.9778   
## parental.level.of.education\_master.s.degree 0.0037 \*\*   
## parental.level.of.education\_some.college 0.7812   
## parental.level.of.education\_some.high.school NA   
## lunch\_free.reduced < 2e-16 \*\*\*  
## lunch\_standard NA   
## test.preparation.course\_completed < 2e-16 \*\*\*  
## test.preparation.course\_none NA   
## reading.score 5.52e-10 \*\*\*  
## writing.score < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.362 on 985 degrees of freedom  
## Multiple R-squared: 0.8767, Adjusted R-squared: 0.875   
## F-statistic: 500.3 on 14 and 985 DF, p-value: < 2.2e-16

###### get all dependent variables

sp\_dependent <- sp  
sp\_dependent$math.score <- NULL  
predicted <- predict(linearMod, sp\_dependent)

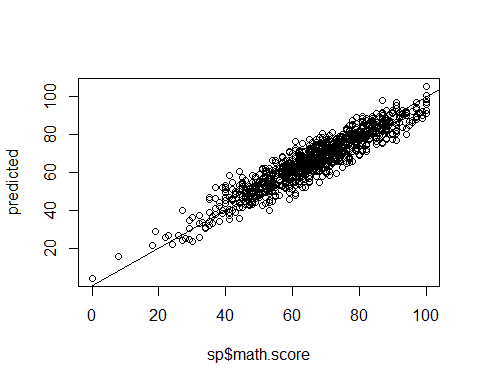
## Warning in predict.lm(linearMod, sp\_dependent): prediction from a rank-deficient  
## fit may be misleading

print(c("correlation between predicted and actual:", cor(predicted, sp$math.score)))

## [1] "correlation between predicted and actual:"  
## [2] "0.936326111317846"

###### Plot actual vs predicted

plot(sp$math.score, predicted)  
abline(lm(sp$math.score~predicted))

 # Methodology and Results

The researcher used various techniques to analyze the data. First, the researcher checked whether there as a difference of mean between male and female math score. The researcher used ggplot to plot the data and see the differences. The mean math score for female is 63.63 whereas, the same for the male is 68.72. This shows a huge difference and disagrees with the hypothesis that the mean math score for both male and female is the same. Hence, we reject the null hypothesis (hypothesis number 1) and go along the alternative hypothesis.

In addition to that, the researcher also worked with the reading data and found that the mean reading score is 72.60 for female while its 65.47 for male. This also proves that the mean reading score of male and female is different from each other. Here, we reject the null hypothesis (hypothesis number 2) and go with the alternative hypothesis.

## t-test

The researcher performs t-test to check if the difference between the mean of reading score is significant. With the analysis, the researcher finds out that the t score is -5.398, degree of freedom (df) is 997.98 and the p-value is 0.001 with 95% confidence interval.Hence, the p value is less than 0.05 here we reject null hypothesis and proves that researcher’s finding is significance.

The researcher performs t-test to check if the difference between the mean of reading score is significant. It also means the true difference is greater than or less than 0. With the analysis, the researcher finds out that the t score is 7.9684, degree of freedom (df) is 996.36 and the p-value is 0.001 with 95% confidence interval. Hence, the p value is less than 0.05 here we reject null hypothesis and proves that researcher’s finding is significance.

## Correlation and Simple Linear Regression

Correlation and simple linear regression The researcher took a further step to find the correlation between the dependent variable (math score) with other independent variables where the independent variables are both numerical and categorical.

In addition to above, the researcher conducted a simple linear regression model. The analysis shows that the residual standard error is 5.362 on 985 degree of freedom (df) with R- squared of 0.8767 and adjusted R-squared of 0.875. The simple linear regression model shows the f-statistics to be 500.3 and the p- value less than 0.0001.

# Implications

The researcher has found that the mean difference between the male and female in their math score and reading score is difference because of various factor including different value priorities, working situations where male tend to enjoy more working with things while female like to work with people around, problem solving strategies where male tend to follow more noble kind of procedures while female like to follow school taught tactics and lastly teacher’s stereotypes, where the teacher perceives that male are better in maths compared to male.

# Conclusion

The researcher would like to conclude that the math mean score in male is higher than in female which is due to various factors mentioned in recommendation. The quantitative result shows that there is difference in the math score and reading score in both male and female where various factors, mainly value priorities and nature or choice of work in each gender is regarded very important.

# References

Diane F. Halpern and Gernsbacher ([n.d.](#ref-Halpern2006)) Green ([n.d.](#ref-Barrow2018)) (*Student Performance* [2018](#ref-Kaggle2018)) Todd ([n.d.](#ref-Todd2019))

Diane F. Halpern, David C. Geary, Camilla P. Benbow, and Morton Ann Gernsbacher. n.d. *The Science of Sex Differences in Science and Mathematics*. [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4270278/](%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4270278/%20).

Green, June Barrow. n.d. *Women in Mathematics: The History Behind the Gender Gap*. [https://www.open.edu/openlearn/science-maths-technology/mathematics-statistics/women-mathematics-the-history-behind-the-gender-gap](%20https://www.open.edu/openlearn/science-maths-technology/mathematics-statistics/women-mathematics-the-history-behind-the-gender-gap%20).

*Student Performance*. 2018. [https://www.kaggle.com/spscientist/students-performance-in-exams](%20https://www.kaggle.com/spscientist/students-performance-in-exams%20).

Todd, Sara. n.d. *The Test Results That Best Explains the Stem Gender Gap Are Not the Math Scores*. [https://qz.com/work/1708961/the-stem-gender-gap-comes-from-girls-advantage-in-reading-over-math/](%20https://qz.com/work/1708961/the-stem-gender-gap-comes-from-girls-advantage-in-reading-over-math/%20).