Math1324 - Applied data Project - Assignment 2

Introduction to Statistics - Assignment 2

Chander Mohan(s3905185) and Alex Thomas(s3925735)

Last updated: 28 May, 2023

RPubs link information

- Rpubs link:https://rpubs.com/S3925735/1046682

Introduction

- Our project is about investigating potential gender-related impact on academic performance among students. In this project our goal is to analyse whether there is a significant difference in average test score between male and female students in a public school. By studying the relationship between gender and exam score, we can gain understanding about the potential impact of gender on academic achievement and their implication on future opportunities.It is important to identify any gender gaps in academic performance to ensure equal opportunities.
- By using statistical analysis, we will inspect the test score among male and female students. By using null hypothesis and alternative hypothesis, we will conclude whether there are significant differences in the average score of male and female or there are not any differences in the average score of male and female.
- The information we will get by using statistical analysis can guide educational policymakers, school administrators, and teachers to design support systems to fill the identify gap.

Introduction Cont.

- It is a well known fact that the female population has lesser opportunities to work in various major sector compared to the male population.
- To Understand whether there is equal opportunity for everyone in terms of academic results.
- i.e. We want to check whether there is a difference in merit results between the male and female population
- This Data is based on the results of examination scores of males and females in public school

Problem Statement

- Understanding the impact of gender on students' academic performance is crucial for promoting equal opportunities and addressing potential disparities in educational outcomes.
- The aim of this project is to investigate whether there are significant differences in test scores between male and female students, which can provide insights into the potential influence of gender on academic performance.
- To achieve this, we will conduct hypothesis testing to compare the mean test scores of male and female students. By performing statistical analysis, we can determine whether the observed differences are statistically significant.
- By using hypothesis testing, we gain some insights in identifying gender-related factors that may impact academic performance. This will provide valuable information for educational interventions and support strategies aiming to promote equal opportunities.

Data

- The Dataset used is open sourced and it is retrieved from Kaggle
- Dataset link: https://www.kaggle.com/datasets/desalegngeb/students-exam-scores?select=Original_data_with_more_rows.csv/
- Data contains information about Examination results of students in public schools

Data Cont.

Data Characteristics - Data Contains examination results of more than 30000 records of students in a public schools

- Data consists of variables of different types characters as well as numeric
- Gender: Gender of the student whether they are male or female
- WritingScore: Writing test results of the student value from (0 to 100)
- MathScore: Math test results of the student value from (0 to 100)
- ReadingScore: Reading test results of the student value from (0 to 100)

Data Pre Processing:

- Data was checked for null values and outliers in marks which was beyond the range of 0-100
- We have included a new metric of AVG_Score which is the average value of Writing score, Math Score and Reading Score so that we can have a combined knowledge of the performance of the student against all three exam types.

Descriptive Statistics and Visualisation

- Reading and Transforming the data

```
exam_data <- read.csv("Original_data_with_more_rows.csv")
dim(exam_data)
```

```
## [1] 30641     9
```

```
sum(is.na(exam_data))
```

```
## [1] 0
```

```
#Finding the average scores
exam_data <- exam_data %>%
  mutate(Avg_Score = (MathScore + ReadingScore + WritingScore) / 3)
```

```
gender_counts <- table(exam_data$Gender)

print(gender_counts)
```

```
##
## female   male
##  15424  15217
```

Bar chart between female and male population

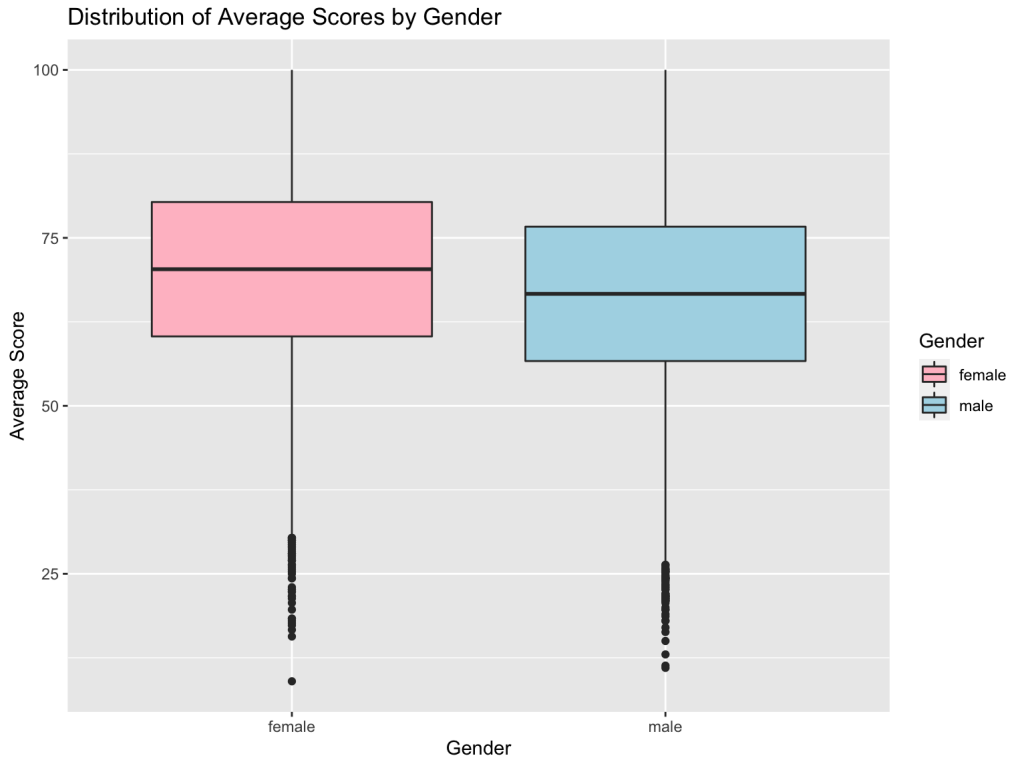
By using a bar plot we can analyse how much difference there is between male and female category. By seeing the plot we can say that there is not much difference between these two categories.

```
ggplot(exam_data, aes(x = Gender, y = Avg_Score, fill = Gender)) +
  geom_bar(stat = "identity") +
  scale_fill_manual(values = c("pink", "lightblue")) +
  ylim(0, 100) +
  labs(x = "Gender", y = "Average Score", title = "Average Score by Gender")
```

Boxplot between female population and male population

- Boxplot was ploted to check if there is any outliers due to data entry for results out of 100 i.e.no value outside the range of 0 to 100 for the Average score

```
ggplot(exam_data, aes(x = Gender, y = Avg_Score, fill = Gender)) +
  geom_boxplot() +
  scale_fill_manual(values = c("pink", "lightblue")) +
  labs(x = "Gender", y = "Average Score", title = "Distribution of Average Scores by Gender")
```



Decsriptive Statistics Cont.

- Statistical summary of exam score with respect to male and female.
- We can see that there is no missing values in our dataset

```
exam_data %>% group_by(Gender) %>% summarise(Min = min(Avg_Score,na.rm = TRUE),
                                             Q1 = quantile(Avg_Score,probs = .25,na.rm = TRUE),
                                             Median = median(Avg_Score, na.rm = TRUE),
                                             Q3 = quantile(Avg_Score,probs = .75,na.rm = TRUE),
                                             Max = max(Avg_Score,na.rm = TRUE),
                                             Mean = mean(Avg_Score, na.rm = TRUE),
                                             SD = sd(Avg_Score, na.rm = TRUE),
                                             n = n(),
                                             Missing = sum(is.na(Avg_Score))) -> table1
knitr::kable(table1)
```

Gender	Min	Q1	Median	Q3	Max	Mean	SD	n	Mis
female	9	60.33333	70.33333	80.33333	100	70.08480	14.17638	15424	
male	11	56.66667	66.66667	76.66667	100	66.45243	14.25669	15217	

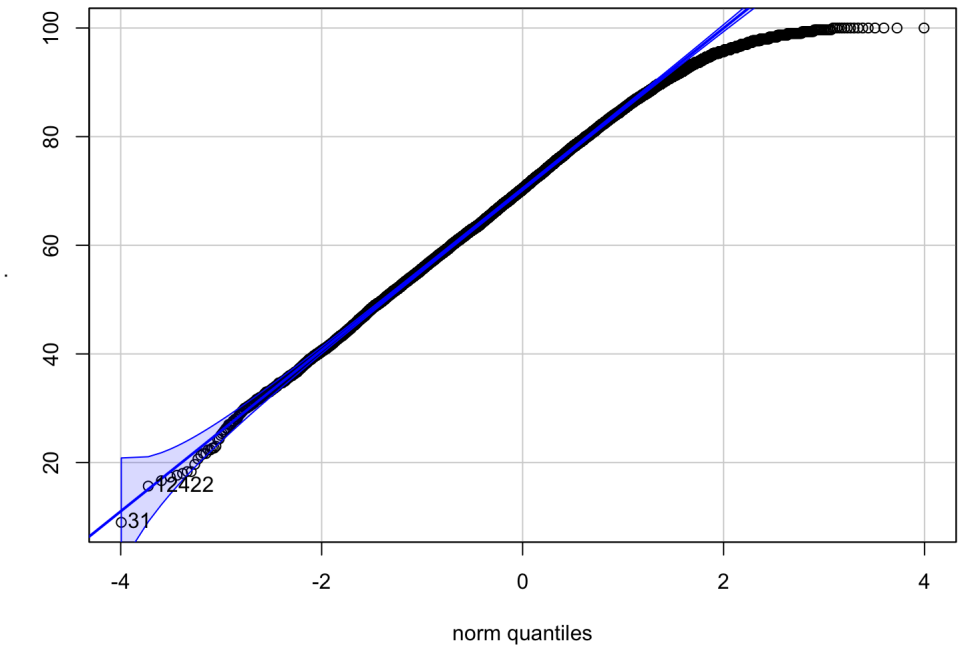
```
female_exam <- exam_data%>%filter(Gender=="female")
male_exam<-exam_data%>%filter(Gender=="male")
```

Hypothesis Testing

- Two sample t- test are used to compare the significant difference between two populations by using their mean.
- We are using this two sample independent t-test because we have Examination scores of two independent populations i.e. female and male.
- Before conducting the hypothesis test, we are checking normality and variance homogeneity by using qqPlot and leveneTest for equal variance.
- for normal distribution, our data sample size is greater than 30 so we can say that data is normally distributed by using Central limit theorem

Hypothesis Testing - qqPlot for Avg_score of females

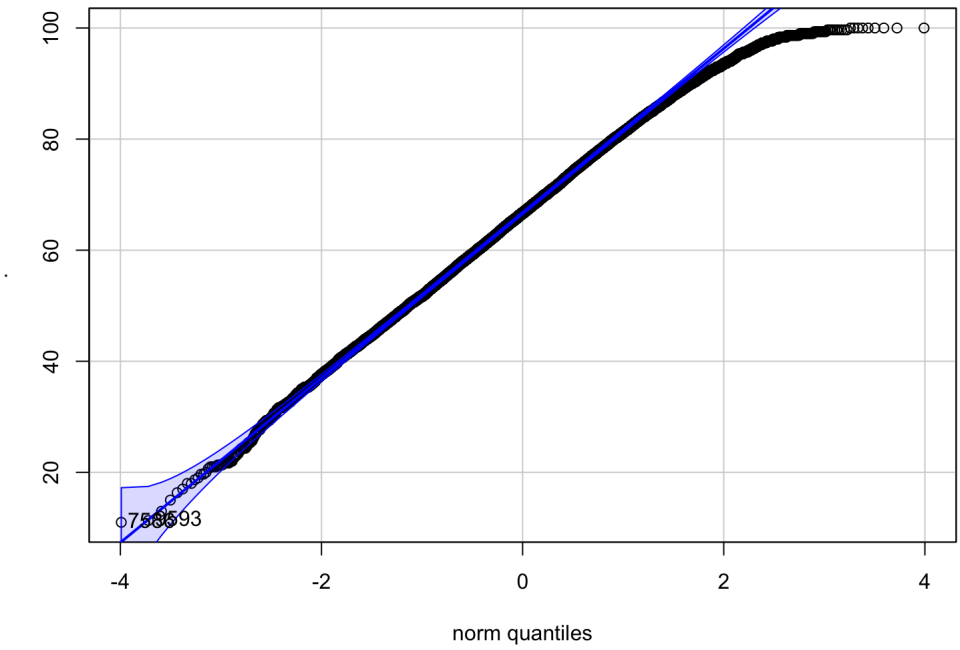
```
female_exam$Avg_Score%>%qqPlot(dist="norm")
```




```
## [1]    31 12422
```

Hypothesis Testing - qqPlot for Avg_score of males

```
male_exam$Avg_Score%>%qqPlot(dist="norm")
```




```
## [1] 7586 9593
```

- For Both groups the Avg_Score follows a normal distribution. Also as n>30 in both groups the sampling distribution will approximate a normal distribution,

Hypothesis Testing - Levene Test for Homogenity of Variance

- Assuming our Null-hypothesis of levenetest: sigma_1 = sigma_2 (population variances are homogeneous)
- Alternate hypothesis of levenetest: sigma_1 != sigma_2 (population variances are not homogeneous)

```
knitr::kable(round(leveneTest(exam_data$Avg_Score ~ exam_data$Gender, data = exam_data),3))
```

	Df	F value	Pr(>F)
group	1	0.013	0.911
	30639	NA	NA

- We can see that by leveneTest, the p-value is grater than 0.05 at 0.911 so we can say that population variances are homogeneous.
- As we can see that our data is normally distributed and the population variances are homogenous, Now we can apply the two-sample t-test

Hypothesis Testing - Two Sample t test

After checking all the condition, we can apply the t-test

Taking the assumptions - Null Hypothesis: In this null hypothesis we are assuming there is no difference between the mean average score of both populations.

- Alternative Hypothesis: In this alternative Hypothesis, we will assume that there is a difference between the mean average score of both populations.
- H0 : μ1 = μ2 HA : μ1 ≠ μ2
- μ1 is the mean exam score of male students
- μ2 is the mean exam score of female students

```
ttest<-t.test(`Avg_Score` ~ Gender,
data = exam_data,
var.equal = TRUE,
alternative = "two.sided"
)
ttest
```

```
##
##  Two Sample t-test
##
## data:  Avg_Score by Gender
## t = 22.362, df = 30639, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group female and group male is
not equal to 0
## 95 percent confidence interval:
##  3.313994 3.950747
## sample estimates:
## mean in group female   mean in group male
##             70.08480             66.45243
```

Discussion

- Since the test static t from the t- test is 22.362 which indicates an extreme difference in means between the female and male groups.
- The p value is also less than 0.05 so we are rejecting the null hypothesis. This suggests that there is a significant difference between the mean exam score of female and male.
- As 95% confidence interval [3.313994,3.950747]. Which does not capture H0 : $\mu 1 = \mu 2$, Therefore results of the test was statistically significant
- The estimates from the sample data indicate that females have a higher average exam score (70.08480) compared to males (66.45243).
- Based on the evidence, it supports the hypothesis that the average exam scores of females and males are different. The mean estimates from the sample data further support that the female population performed better in the exam compared to the male population.
- The results suggest that females have better merit results and outperform males in the exam. This may indicate potential disparities in work opportunities, with females having lesser opportunities despite their better performance.

References

- Data and information related to the data gathered from https://www.kaggle.com/datasets/desalegngeb/students-exam-scores?select=Original_data_with_more_rows.csv/