Presidency University
STAT03SEE
DATA ANALYSIS PROJECT

Spandan Ghoshal
Reg.No :- 18214110001
Roll.No :- 18214001

January 16, 2020

Introduction

Young generation of today are exposed to the risk of being victims of HIV/AIDS - which was quite unknown to their predecessors a few decades ago. The epidemic of HIV/AIDS is now progressing at a rapid pace among young people. Studies have reported that young people form a significant segment of those attending sexually transmitted infection (STI) clinics and those infected by HIV.

Programme managers and policy makers have often recommended that schools and colleges can act at the center point for disseminating information and education on HIV/AIDS. Hence any materials prepared to enhance awareness constitute a potent weapon in the hands of public health personnel. In the following study a group of students are taken an examination signifying their awareness regarding HIV/AIDS, before and after the awareness programme. Here we are interested to examine any notable diffference in POST-SCORE in comparison with PRE-SCORE.Also we are interested to check the effect of different groups on the scores. Hence our hypothesis for the statistical study will be :-

$$\mathscr{H}_o : \mu_{\text{Post}} = \mu_{\text{Pre}}$$
$$\mathscr{H}_1 : \mu_{\text{Post}} > \mu_{\text{Pre}}$$

Since the given data has PRE.TEST and POST.TEST scores for 5 disciplines :

- Chemistry (Special)
- Botany (Special)
- Microbiology (SYBSc level)
- Microbiology (Special level)
- Zoology (Special)

Hence for different groups, we can write :-

$$\mathscr{H}_o : \text{means of different groups are equal}$$
$$\mathscr{H}_1 : \text{at least one of them is unequal}$$

Exploratory Data Analysis

The mean value of PRE.TEST scores is :-

```
[1] 11.11111
```

The mean value of POST.TEST scores is :-

```
[1] 18
```

Observation:- Hence there is a clear difference in the two mean values which furthur supports our initial hypothesis.

The mean values of PRE.TEST and POST.TEST scores for different subjects :-

```
   Subjects  PRE_mean POST_mean     Diff
1 Chemistry  9.533333  18.00000 8.466667
2    Botany  9.368421  16.68421 7.315789
3 Micro_Bsc 10.857143  18.28571 7.428571
4  Micro SP 14.600000  19.00000 4.400000
5   Zoology 15.000000  19.28571 4.285714
```

Observation:- Even for different Subjects, the POST.TEST socres are quite higher than PRE.TEST scores though the difference varies from subject to subject.

Variance of PRE.TEST and POST.TEST scores for different subjects :-

```
   Subjects   PRE_var  POST_var
1 Chemistry 6.6952381 2.4285714
2    Botany 6.8011696 3.0058480
3 Micro_Bsc 5.4285714 1.8142857
4  Micro SP 2.9333333 0.4444444
5   Zoology 0.6666667 0.2380952
```

Observation:- For different Subjects, variances of POST.TEST socres are quite small for PRE.TEST scores which might be the indication for increase in consistency of the students as a result of the awareness programme.

The Scatterplot for both the Scores for all the subjects and subjectwise :-

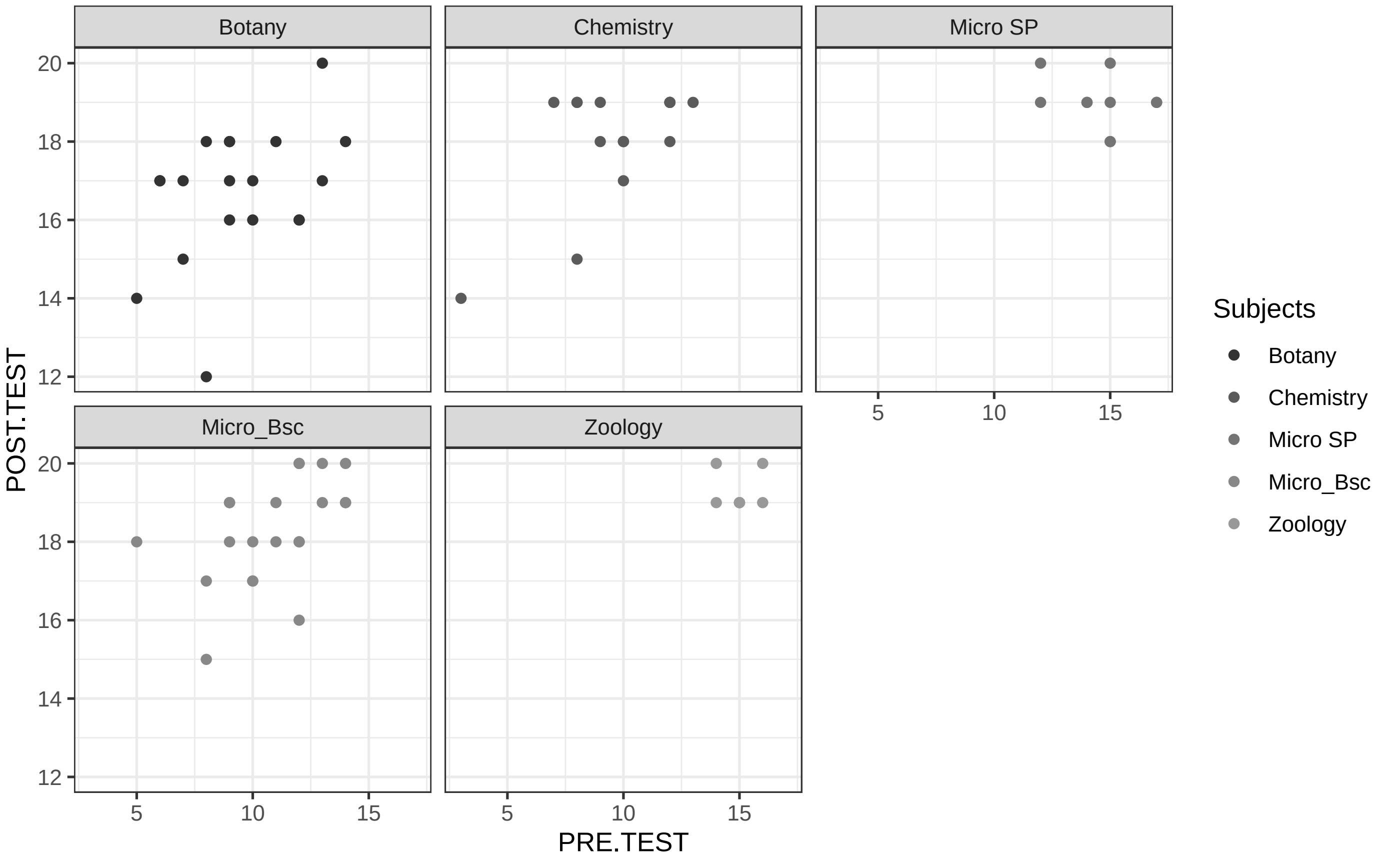

Figure 0.1: Scatterplot of Pre and Post test Scores

Observation:- Hence there is a slight dependence between PRE.TEST and POST.TEST scores as we can verify from the correlation between them :-

```
[1] 0.588057
```

Subject wise correlations :-programme

```
   Subjects Correlation
1 Chemistry   0.6199855
2    Botany   0.4203509
3 Micro_Bsc   0.4916206
4  Micro SP  -0.2919371
5   Zoology   0.0000000
```

Observation:- The change of correlation for different groups is a very interesting observation.

Now we make boxplot and jittered plots of both scores to see the difference between their mean and variablity.

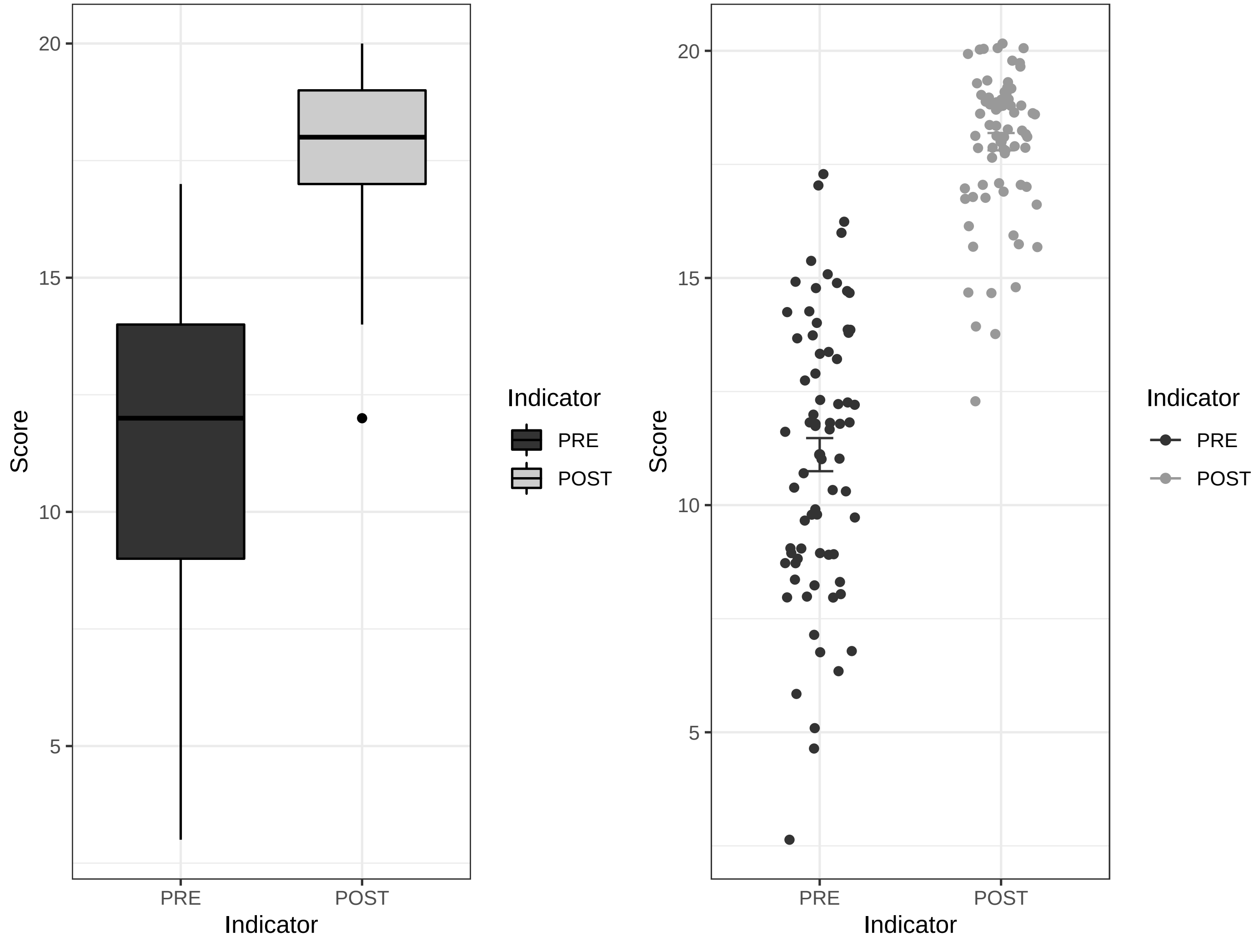

Figure 0.2: Boxplot of Scores and Jittered Plots of Scores(with standard errors)

Observation:- Though PRE.TEST scores has less mean than POST.TEST scores, but also has higher variance in comparison.

Subjectwise boxplots for the two scores :-

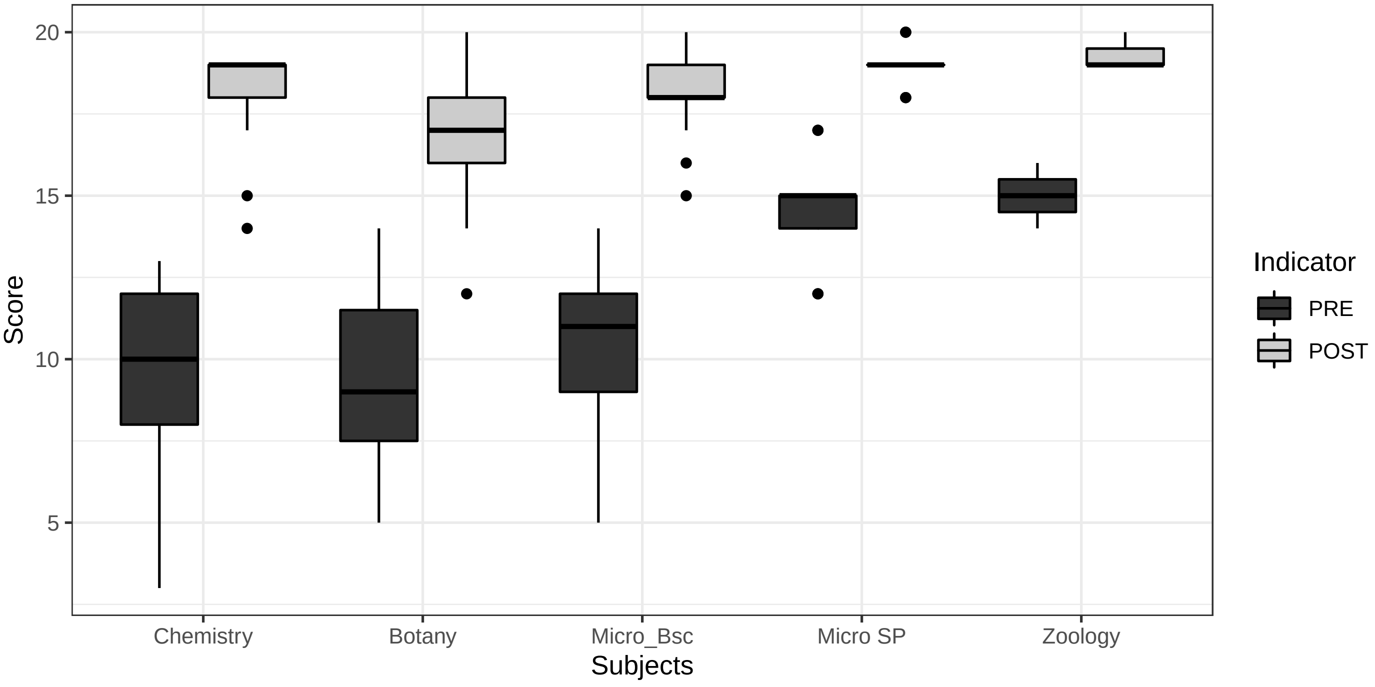

Figure 0.3: Subjectwise Boxplots

Subjectwise jittered plots for the two scores :-

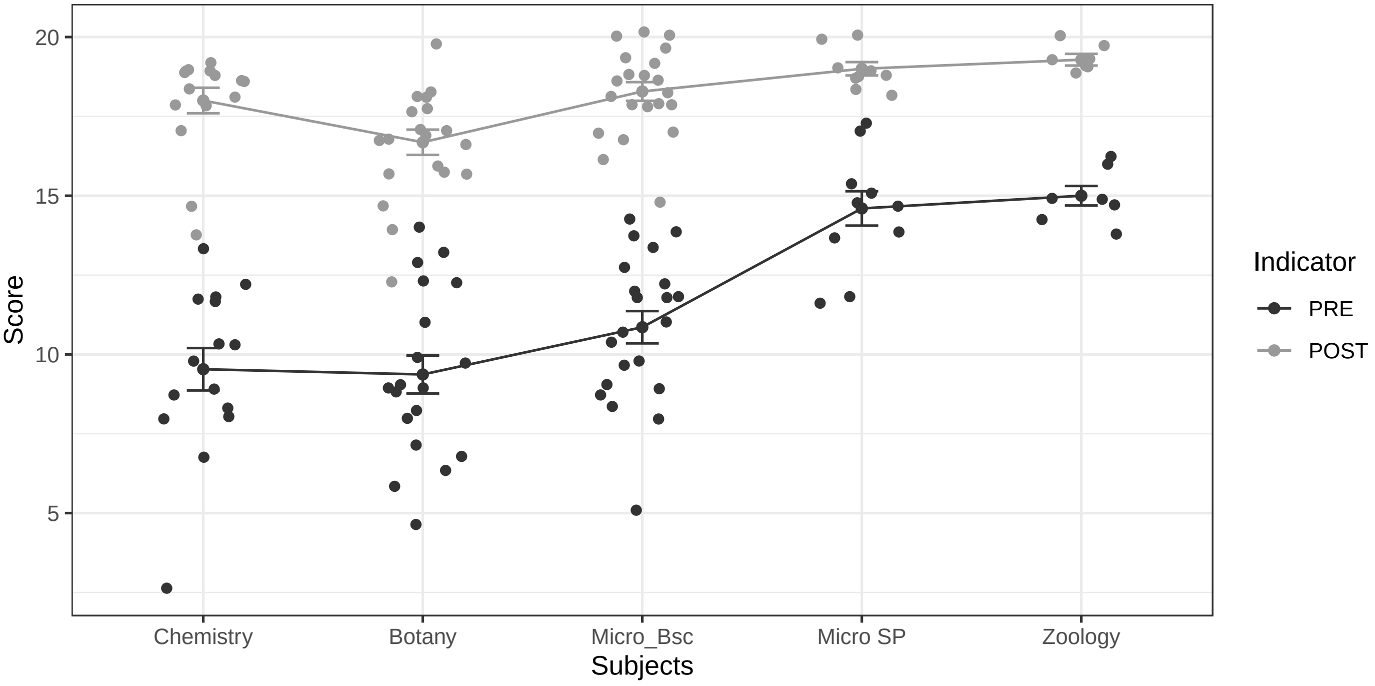

Figure 0.4: Subjectwise Jittered Plots

For visualizing densityplots of marks we make Ridgeline Plots :-

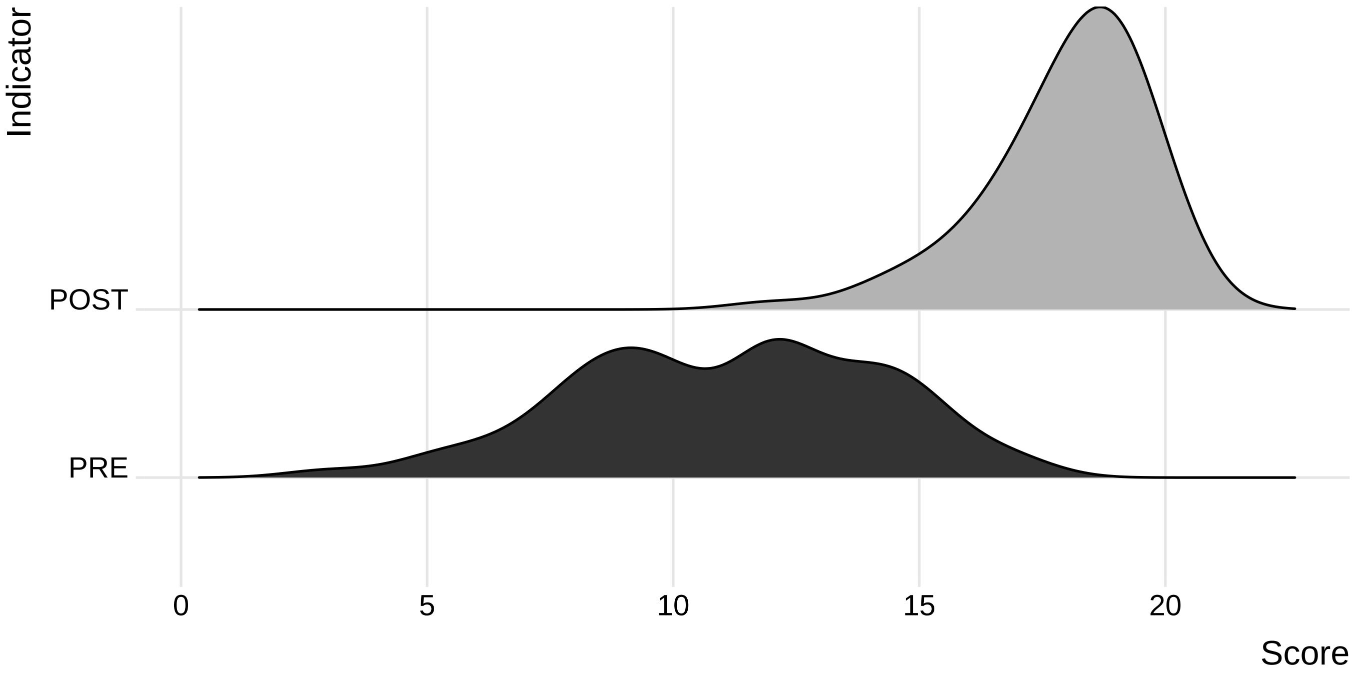

Figure 0.5: Ridgeline Plots for both tests

For visualizing densityplots of subjectwise marks we make Ridgeline Plots :-

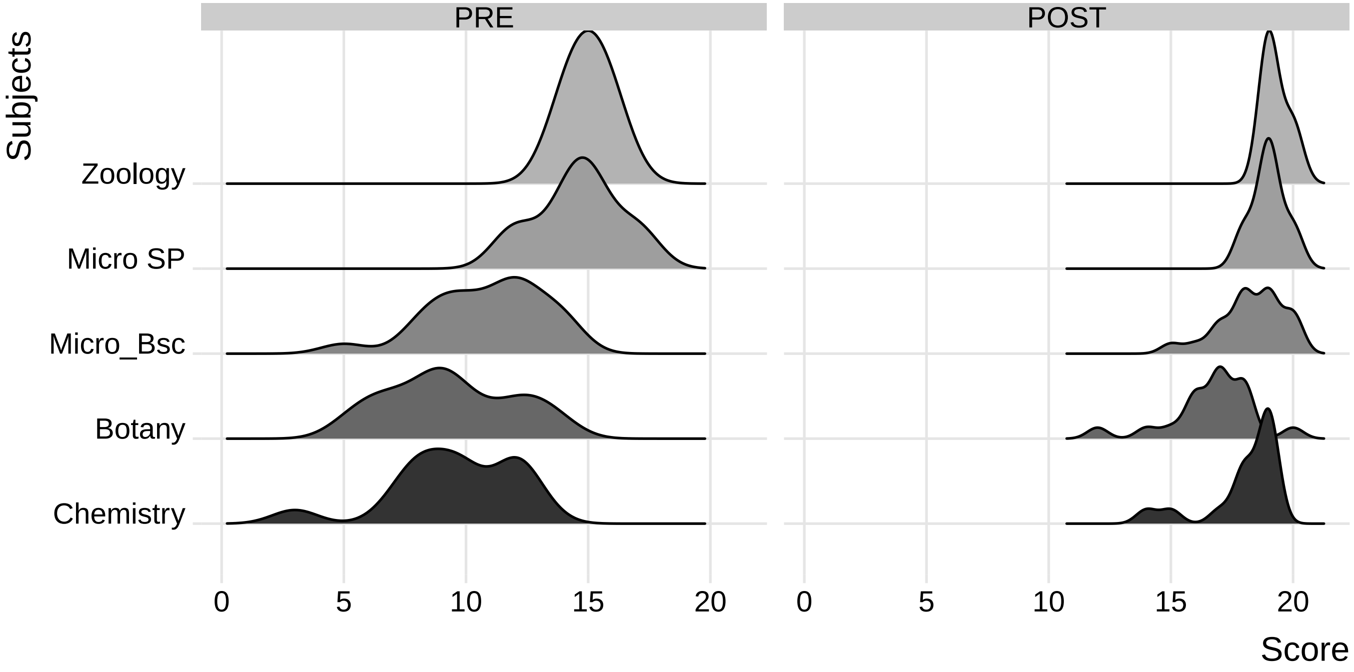

Figure 0.6: Subjectwise Ridgeline Plots

Observation :-

From several statistical measures and data vizualization techniques, applied on the given data, it can be suspected that the awareness programme had a major impact on the test scores of the students. Also their respective discplines may have an impact on their performance. The significance of these claims are verified in the next part of this report using conformatory analysis tools.

Normality Test :-

To check whether both the scores follow normal distribution, we make their respective qqplots and also the qqplot for the difference in the scores.

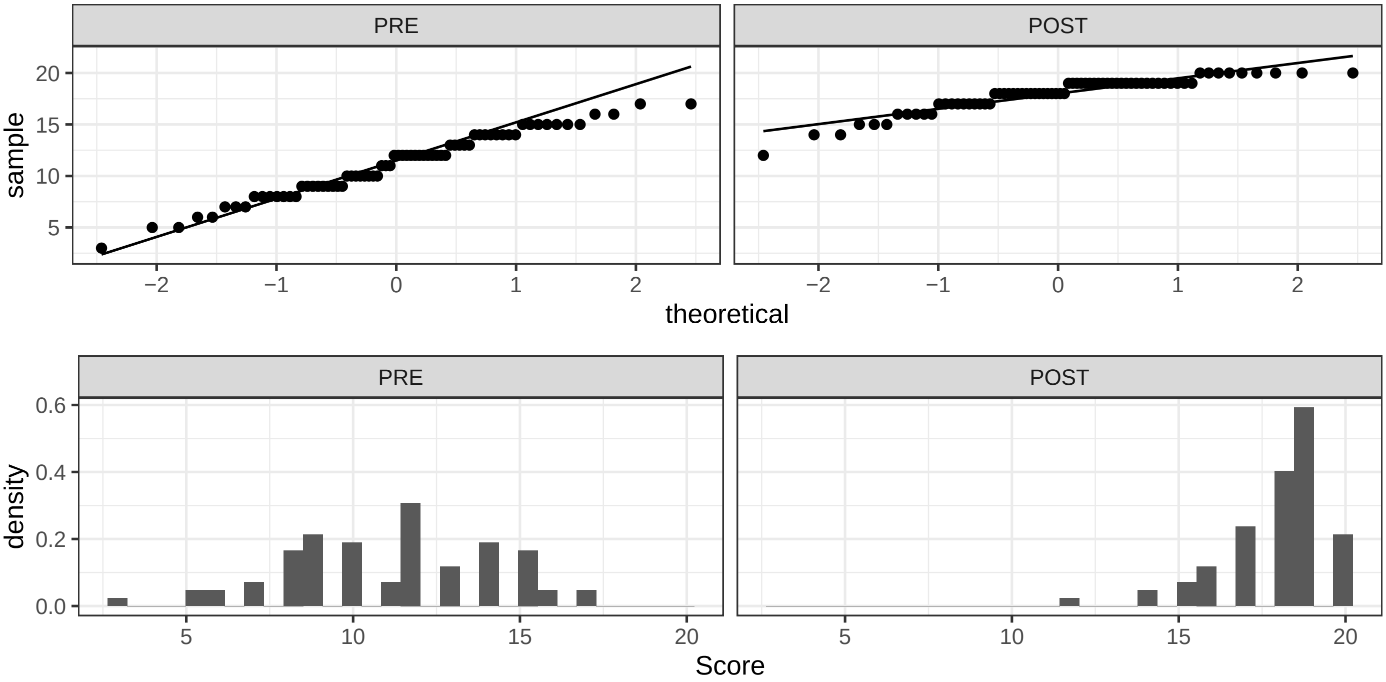

Figure 0.7: qqplot and histogram for both Scores

As the quantiles fall alongside the reference line, hence our assumption of normality may be assumed true.

Inferential Data Analysis

Model

Since there are 72 students and the data consists of the PRE.TEST and POST.TEST scores of the same individual and from the QQplots in the previous section, we can assume both the scores jointly follow a Bivariate Normal Distribution with unknown parameters. Let X : PRE.TEST scores, Y : POST.TEST scores then, our model is :-

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim BN\left(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho\right)$$

Test of Hypothesis

From the exploratory data analysis, we did in the previous segment, we hypothesize some facts about the given sample data and check their significance using inferential tools.

From figure (0.2), we can observe that PRE.TEST scores had greater variance, compared to POST.TEST scores. Hence, we set the following hypothesis :-

$$\mathscr{H}_o : \frac{\sigma_X^2}{\sigma_Y^2} = 1$$
$$\mathscr{H}_1 : \frac{\sigma_X^2}{\sigma_Y^2} > 1$$

Using R we test our hypothesis :-

```
F test to compare two variances

data:  A$PRE.TEST and A$POST.TEST
F = 3.6511, num df = 71, denom df = 71, p-value =
7.062e-08
alternative hypothesis: true ratio of variances is greater than 1
95 percent confidence interval:
 2.464515      Inf
sample estimates:
ratio of variances
          3.651135
```

Since the p-value for the Test :-

```
[1] 7.061674e-08
```

is very less than $\alpha = 0.05$, we can easily reject our $\mathscr{H}_o$ with level of significance α. Hence POST.TEST scores has a significantly lower variance than PRE.TEST scores.

From figure (0.2), we can also note a difference in the means of PRE.TEST and POST.TEST scores. Hence we now hypothesize the following :-

$$\mathscr{H}_o : \mu_Y = \mu_X$$
$$\mathscr{H}_1 : \mu_Y > \mu_X$$

Then the above is a paired t-test :-

```
Paired t-test

data:  A$POST.TEST and A$PRE.TEST
t = 23.294, df = 71, p-value < 2.2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 6.396006      Inf
sample estimates:
mean of the differences
               6.888889
```

In this case also the p-value is :-

```
[1] 2.815859e-35
```

very less than $\alpha = 0.05$ hence our suspect was right and we can reject the null hypothesis. Hence mean POST.TEST score is significantly higher than mean PRE.TEST score.

To check if the mean values are significantly higher for each group or not, we perform Subjectwise t-test for unequal variance :-

```
   Subjects     p-values
1 Chemistry 9.558381e-11
2    Botany 6.778763e-11
3 Micro_Bsc 1.643017e-13
4  Micro SP 3.460692e-05
5   Zoology 1.055050e-05
```

As we can clearly see that for each group the p-values are very less than $\alpha = 0.05$ hence we can reject the null hypothesis that the values of mean are significantly equal hence we can state on the light of the given data that the mean POST.TEST scores are significantly higher than the mean PRE.TEST scores for each subject.

To check homoskedasticity for different groups we apply F-test for each subjects :-

```
   Subjects    p-values
1 Chemistry 0.033882473
2    Botany 0.045926733
3 Micro_Bsc 0.009034524
4  Micro SP 0.004843715
5   Zoology 0.117876951
```

Since for most of the Subjects the p-values are very less than $\alpha = 0.05$ hence we reject the null hypothesis that the Scores are Homoskedasticity across different subjects. We also an readily state that the score variances are significantly reduced for most of the subjects.

For measuring subjectwise mean score differences (mean of the differences between PRE.TEST and POST.TEST scores) , we can use pairwise t-test for unequal variance. The p-values for all the subject combinations are given below :-

```
                Botany    Chemistry    Micro SP  Micro_Bsc
Chemistry 0.5658940283           NA          NA         NA
Micro SP  0.0118923100 6.787685e-04          NA         NA
Micro_Bsc 1.0000000000 5.658940e-01 0.006254561         NA
Zoology   0.0009524723 2.106536e-05 1.000000000 0.00013246
```

from the matrix we can say that among students of Microbiology , students persuing Bsc degree did much improvement in the tests compared to students with Specialization in Microbiology where as they had less mean PRE.TEST scores than the later. Whereas Students from Chemistry and Botany did more or less same improvement.

For much stronger conclusions, we perform Non-Parametric tests (Since they make no Parametric Assumptions) to check if the differences are uniform across Subjects and Years or not. We perform Wilcoxon signed rank test :-

```
Wilcoxon signed rank test with continuity correction

data:  A$POST.TEST and A$PRE.TEST
V = 2628, p-value = 7.511e-14
alternative hypothesis: true location shift is greater than 0
```

And even from this test, we can claim that there is a significant difference in the PRE.TEST and POST.TEST scores. (The p-value being very small from $\alpha = 0.05$)

The p-values of the Wilcoxon Signed Rank Test for individual subjects are :-

```
   Subjects     p-values
1 Chemistry 3.403613e-04
2    Botany 6.748512e-05
3 Micro_Bsc 3.033740e-05
4  Micro SP 2.864688e-03
5   Zoology 1.006838e-02
```

Hence for every subject the p-value given by the test is less than $\alpha = 0.05$ hence for every subject or discipline, the POST.TEST scores are significantly high compared to PRE.TEST scores.

Conclusion

On the basis of confirmatory and exploratory analysis done on the data , it can be said that the **awareness programme significantly improved the overall result of the students** by both **increasing their average marks** and **decreasing the variance of test scores (increased consistency)**. Secondly, among different subject specializations , **students of biological sciences** (especially **Microbiology specialization** and **zoology** students) **did better than others** in both the tests. This is quite obvious as they learn about these diseases in their curriculum. Whereas chemistry students did highest improvement among all. Between years Microbiology students of Bachelor level scored less than those with specialization in the subject which is quite obvious due to a higher level of awareness amongst them. But after the awareness programme, both Bsc and Msc students scored almost same which indicates the effectiveness of the material.

Discussion

This study clearly indicates that students of **non-biological streams are less aware** than Students of Biological Sciences. This may be due to the fact that majority of this things are included in the curriculum of some biological streams. **So more emphasis should be given to the rest part** i.e those with non-biological streams even for students from Arts and other disciplines etc. Organizations conducting such awareness programmes should focus more on the segment who are ignorant about the facts regarding HIV/AIDS. However on the light of the given data we can say that the **overall performance of Botany students was quite dissatisfactory** this may be a wrong conclusion due to small sample size, so it's recommended to look upon any specific reason behind this.

Appendix

R Codes

Reading the data frame :-

```
A <- read.table("C:\\Users\\Sukanta\\Desktop\\SEE_Project\
\AIDS_final.dat", header = T , sep = "\t")
```

Mean values for both tests :-

```
mean(A$PRE.TEST)
mean(A$POST.TEST)
```

Subjectwise mean values :-

```
name <- unique(A$Subjects) #_Distinct Subjects
means <- matrix(rep(0,length(name)*2),ncol = 2) #_A matrix to store the values
for(i in 1:length(name))
{
        means[i,1] <- mean(subset(A , Subjects %in% name[i])$PRE.TEST)
        means[i,2] <- mean(subset(A , Subjects %in% name[i])$POST.TEST)
}
#_forming the data frame to store corresponding values_
groups <- data.frame(name , means , means[,2]-means[,1])
colnames(groups) <- c("Subjects","PRE_mean","POST_mean","Diff")
groups
```

Subjectwise variances :-

```
var_s <- matrix(rep(0,length(name)*2),ncol = 2)
for(i in 1:length(name))
{
        var_s[i,1] <- var(subset(A , Subjects %in% name[i])$PRE.TEST)
        var_s[i,2] <- var(subset(A , Subjects %in% name[i])$POST.TEST)
}
groups_var <- data.frame(name , var_s)
colnames(groups_var) <- c("Subjects","PRE_var","POST_var")
groups_var
```

Figure 1 :-

```
library(ggplot2)
library(ggpubr)
library(magrittr)

ggplot(A ,aes(PRE.TEST,POST.TEST, color = Subjects)) +
           geom_point() +
                facet_wrap(~Subjects) +
                 scale_color_grey(start=0.2,end=0.6) +
                  theme_bw()
```

Correlation between PRE.TEST and POST.TEST scores for different Subjects :-

```
cor(A$PRE.TEST,A$POST.TEST)
corr_s <- matrix(rep(0,length(name)),ncol = 1)
for(i in 1:length(name))
{
#_Subjectwise correlation
   corr_s[i,1] <- cor(subset(A , Subjects %in% name[i])
    $PRE.TEST,subset(A , Subjects %in% name[i])$POST.TEST)
}
groups_corr <- data.frame(name , corr_s)
colnames(groups_corr) <- c("Subjects","Correlation")
groups_corr
```

Figure 2 :-

```
#_The data.frame is reshaped with combining both PRE.TEST and
#_POST.TEST scores and adding an extra column with two factor
#_variables "PRE" and "POST"_

A1 <- A[-5]
A2 <- A[-4]
colnames(A1)[4] <- "Score"
colnames(A2)[4] <- "Score"
AIDS1 <- rbind(A1,A2)
#_The Factor variables are stored in "Indicator" column_
Indicator <- c(rep("PRE",72),rep("POST",72))
AIDS <- cbind(AIDS1,Indicator)

#_the factors are ordered as "PRE","POST"_
AIDS$Indicator <- factor(AIDS$Indicator , levels = c("PRE","POST"))

#_the Subjects are ordered accordingly_
AIDS$Subjects <- factor(AIDS$Subjects , levels=
                                   "Zoology"))

#_The boxplots for PRE.TEST and POST.TEST scores are drawn on Y axis
#_containing Scores_
plot1 <- ggboxplot(AIDS , x = "Indicator" , y = "Score") +
                  theme_bw()

#_The Jittered plots along with mean sq.error for PRE.TEST and
#_POST.TEST scores are drawn on Y axis containing Scores_
plot2 <- ggline(AIDS, x = "Indicator", y = "Score" , add =
c("mean_se","jitter")) +
theme_bw()

ggarrange(plot1 , plot2 , ncol = 2 , nrow = 1)
```

Testwise Boxplots for each Subject :-

```
#_Here the boxplot has been drawn w.r.t two factors namely the
#_Subjects and "PRE" or "POST"_
ggboxplot(AIDS , x = "Subjects", y = "Score", color = "Indicator") +
 theme_bw()
```

Testwise Jittered Plots for each Subject :-

```
#_Here the Jitter Plot has been drawn w.r.t two factors namely the
#_Subjects and "PRE" or "POST"_
ggline(AIDS, x = "Subjects", y = "Score", color = "Indicator", add =
c("mean_se","jitter")) +
 theme_bw()
```

Ridgeline Plot for both Scores :-

```
library(ggridges)
ggplot(AIDS, aes(x = Score, y = Indicator, fill = Indicator)) +
                geom_density_ridges() +
                 theme_ridges() +
          theme(legend.position = "none") +
           scale_fill_grey(start = 0.2 , end= 0.7)
```

Ridgeline Plots for both Scores, subjectwise :-

```
library(ggridges)
ggplot(AIDS, aes(x = Score, y = Indicator, fill = Indicator)) +
 geom_density_ridges() +
  theme_ridges() +
   theme(legend.position = "none") +
    scale_fill_grey(start = 0.2 , end = 0.7)
```

QQplot and Histograms for checking normality assumptions for both test scores :-

```
diff <- A$POST.TEST-A$PRE.TEST
#_By subtracting PRE.TEST scores from POST.TEST , we get the
#_increase(or decrease) in the scores for each individual
AI <- data.frame(A[,-c(4,5)] , diff)

qp <- ggplot(AIDS , aes(sample = Score)) +
           stat_qq() +
                stat_qq_line() +
                 facet_wrap(~Indicator) +
                  theme_bw()

hg <- ggplot(AIDS , aes(Score)) +
           geom_histogram()+
                facet_wrap(~Indicator) +
                 theme_bw()
```

```
qq_diff <- ggplot(AI , aes(sample = diff)) +
                        stat_qq() +
                         stat_qq_line() +
                          theme_bw()

hg_diff <- ggplot(AI , aes(diff)) +
                        geom_histogram() +
                         theme_bw()

ggarrange(qp , hg , qq_diff , hg_diff , ncol = 2 , nrow = 2)
```

Testing Homoskedasticity for PRE.TEST and POST.TEST Scores :-

```
vtest1 = var.test(A$PRE.TEST , A$POST.TEST , ratio = 1 , alternative
                 = "greater")
vtest1
vtest1$p.value
```

Testing if the mean values of both tests are equal or not :-

```
ttest1 = t.test(A$POST.TEST , A$PRE.TEST , alternative = "greater" ,
                  paired = T)
ttest1
ttest1$p.value
```

Checking if Mean PRE.TEST and POST.TEST scores are equal :-

```
p_val <- matrix(rep(0,length(name)),ncol = 1)

for(i in 1:length(name))
{
test = t.test(subset(A , Subjects %in% name[i])$POST.TEST ,
subset(A , Subjects %in% name[i])$PRE.TEST ,
alternative = "greater" , paired =T)
p_val[i,1] <- test$p.value
}

p_val <- data.frame(name , p_val)
colnames(p_val) <- c("Subjects","p-values")
p_val
```

Checking Subjectwise Homoskedasticity :-

```
p_val <- matrix(rep(0,length(name)),ncol = 1)

for(i in 1:length(name))
{
test = var.test(subset(A , Subjects %in% name[i])$POST.TEST
, subset(A , Subjects %in% name[i])$PRE.TEST
,alternative = "less",conf.level = 0.99)
p_val[i,1] <- test$p.value
}

p_val <- data.frame(name , p_val)
colnames(p_val) <- c("Subjects","p-values")
p_val
```

Pairwise t-test to check Subjectwise mean differences of both scores :-

```
pair.test <- pairwise.t.test(AI$diff, AI$Subjects
, pool.sd = FALSE)
pair.test$p.value
```

Signed rank test to check whether the two test scores have same mean or not :-

```
rank.test <- wilcox.test(A$POST.TEST , A$PRE.TEST , paired = T ,                    al
rank.test
```

Subjectwise signed rank test :-

```
p_val <- vector(length = length(name))
for(i in 1:length(name))
{
        A_sub = subset(A , Subjects %in% name[i])
        rank.test1 = wilcox.test(A_sub$POST.TEST , A_sub$PRE.TEST
                , paired = T , alternative = "greater")
        p_val[i] = rank.test1$p.value
#_p-values stored in the vector
#_p_val
}
z <- data.frame(name,p_val)
colnames(z) <- c("Subjects","p-values")
z
```

References

The following books were used as a reference to the analysis done in the report :-

- Fundamentals of Statistics, Goon A.M., Gupta M.K., Dasgupta B.
- Extending the Linear Model with R, Faraway Julian J.
- ggplot2: Elegant Graphics for Data Analysis, Wickham Hadley.
- STHDA. [Online]. Available from: http://www.sthda.com/english/wiki/one-way-anova-test-in-r
- The Grammar of Graphics, Leland Wilkinson, D. Wills, D. Rope, A. Norton, R. Dubbs.
- Wikipedia [Online]. Available from: https://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test

Acknowledgment

I would like to express my sincere gratitude to my respected instructor **Prof.Atanu Ghosh** for their guidance and constant supervision as well as for providing all the necessary information regarding the project & also for their support in completing this project.