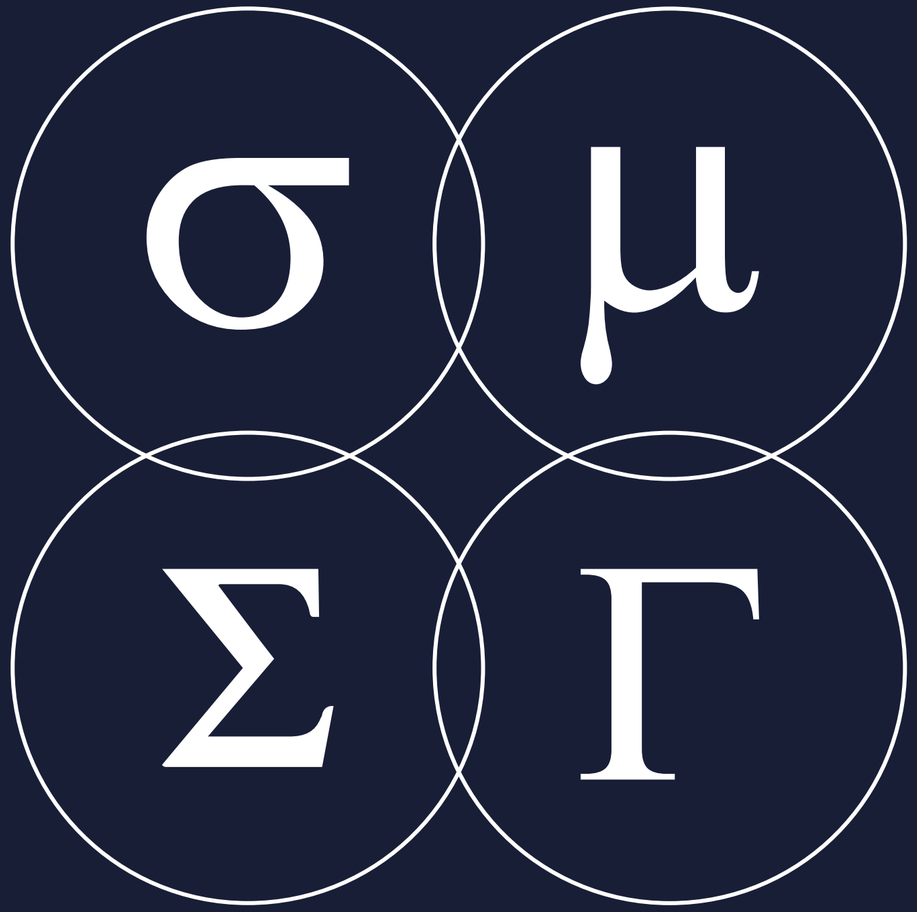
భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad



MA2540 APPLIED STATISTICS

A STATISTICAL STUDY ON ATTENDANCE IN IIT HYDERABAD

Overview

INTRODUCTION

About Our Data

- This study at IITH explores factors influencing student attendance, using survey data to understand engagement and success, including course load, class timing, and demographics.
- We distributed a Google Form questionnaire to IIT Hyderabad students to collect sample data for statistical analysis.
- To boost participation, we sent frequent reminders and ensured survey anonymity. This effort yielded 259 valid responses, forming a strong dataset for our study.

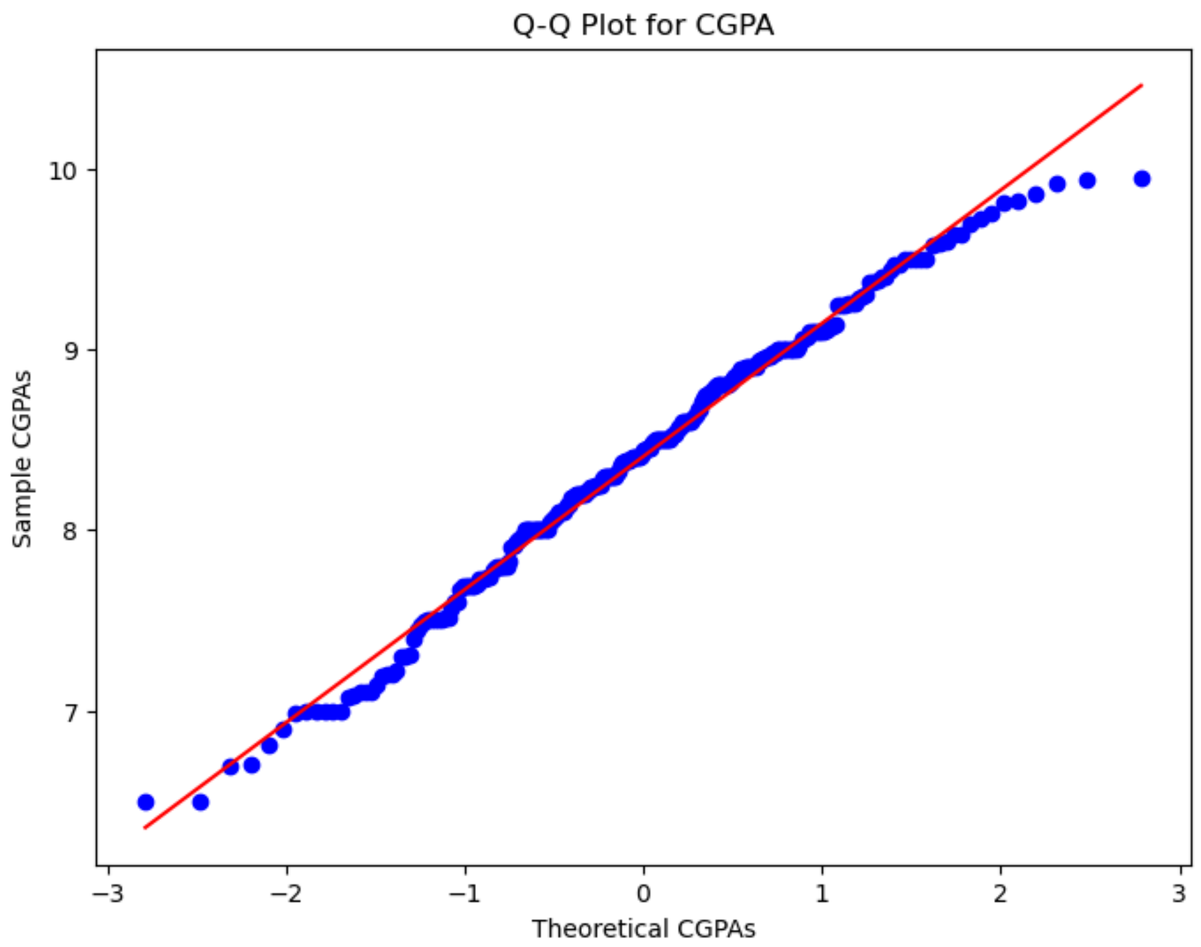
Preprocessing

- Around 50 responses from Google Forms were discarded due to inconsistencies, like CGPA values exceeding 10 or reaching 10 after the 3rd semester, which is unlikely.
- Certain responses showed localized inconsistencies, leading to the substitution of erroneous values with the mean derived from validated data points.
- The initial dataset of 300 responses was refined to a final dataset of about 260 authentic responses after thorough data cleaning and preprocessing.

NORMALITY
TESTS

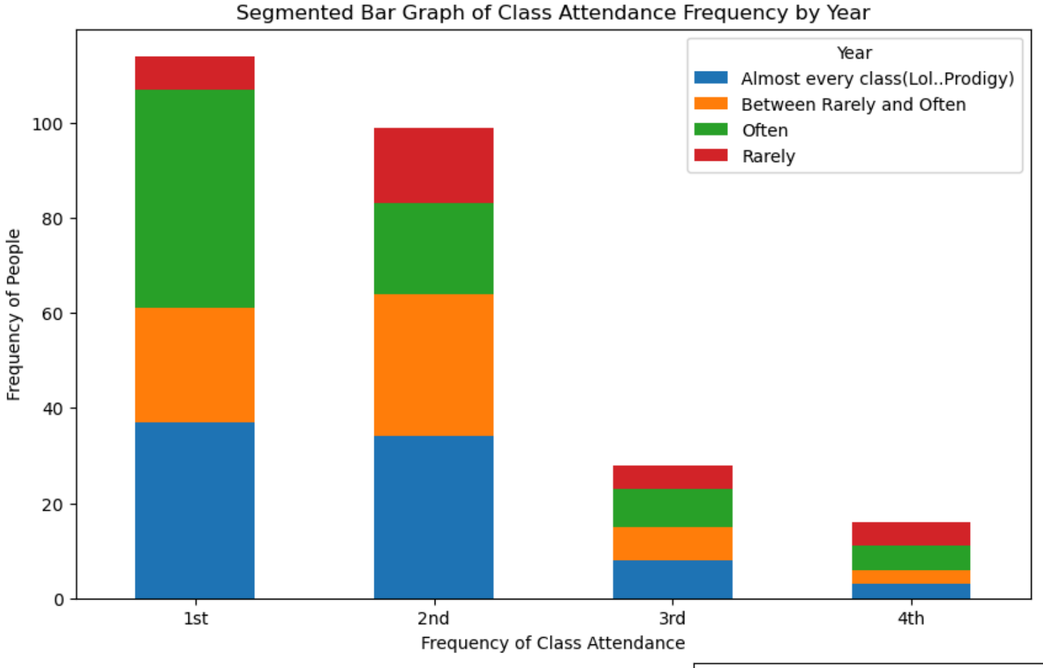
Shapiro – Wilk Test

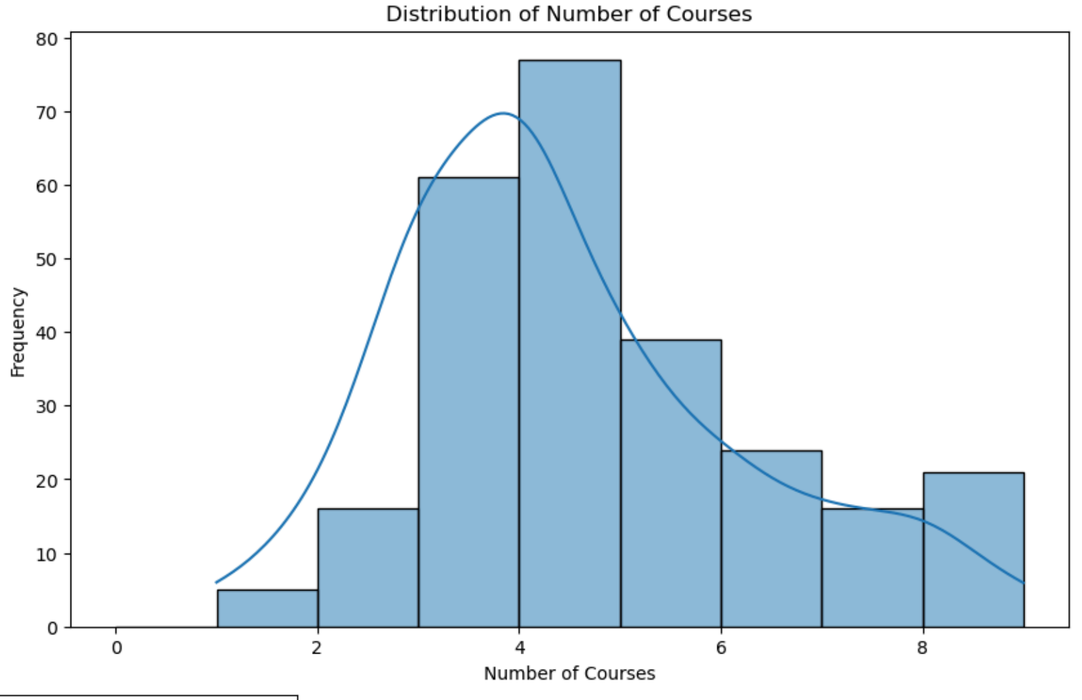
After conducting the Shapiro-Wilk test and observing a test-statistic of approximately 0.98 with a p-value of about 0.049, we verified our inference using a Q-Q' plot. The close alignment of datapoints with the linear line suggests that the CGPA of students likely follows a normal distribution. Hence, the dataset can be considered to originate from a normal distribution.

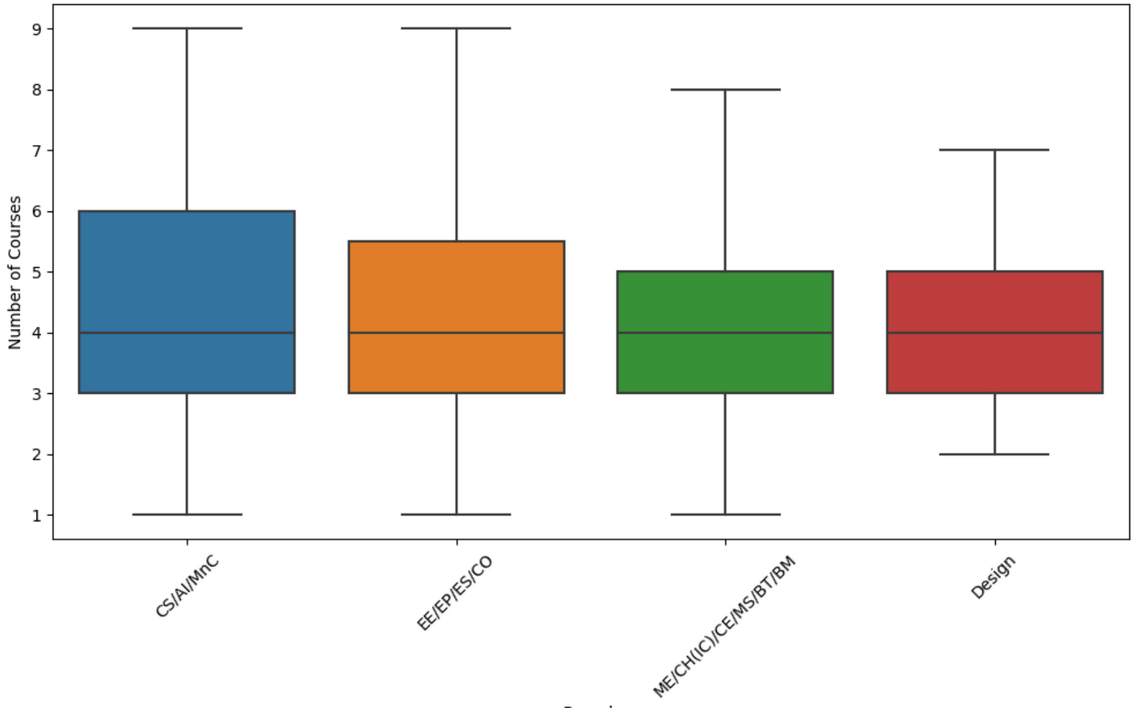


Q – Q’ Plot for CGPA

VISUALISATIONS

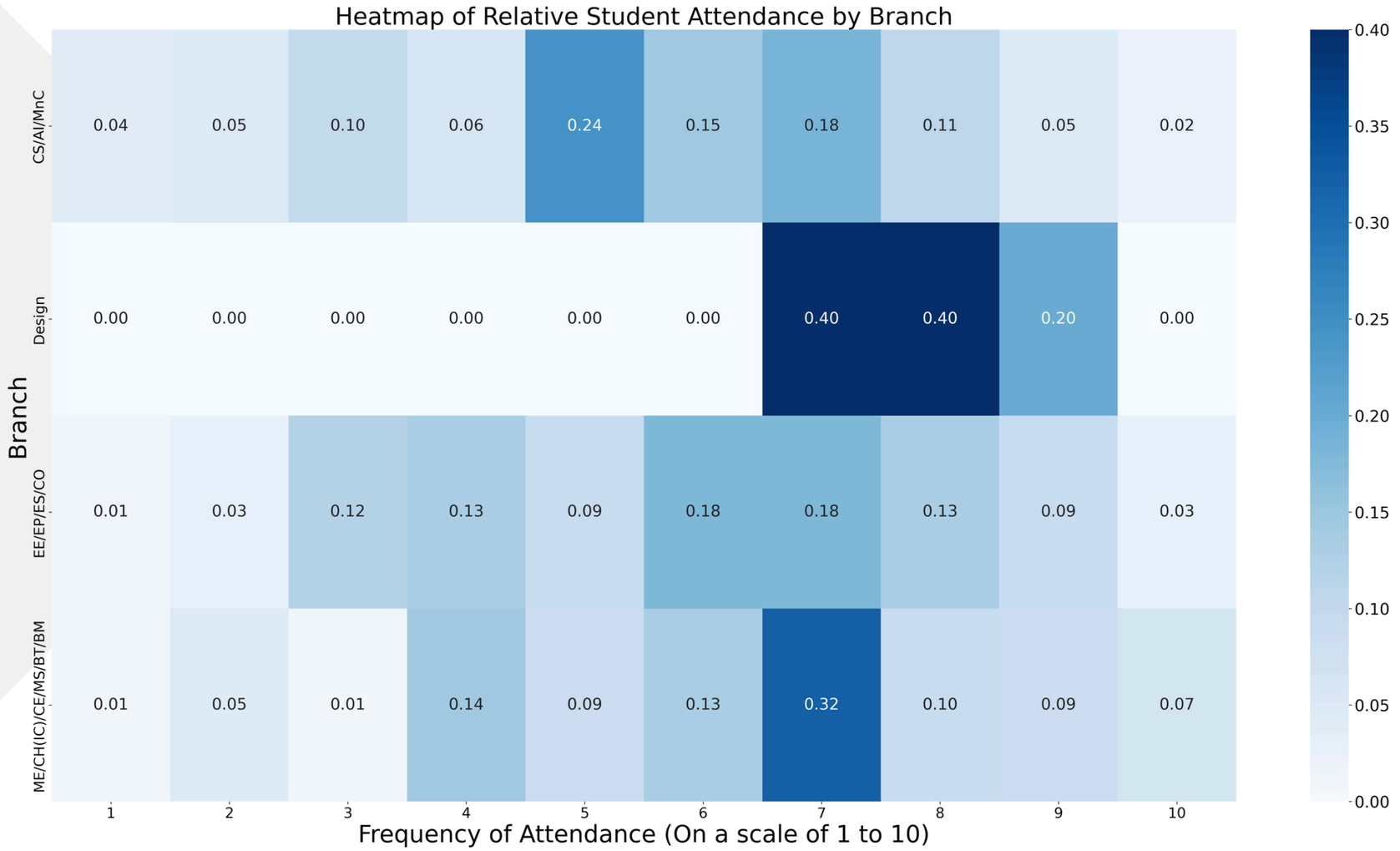

Segmented Bar Graph of Class Attendance Frequency by Year
Year
Almost every class(Lol..Prodigy)
Between Rarely and Often
Often
Rarely
Frequency of People
0
20
40
60
80
100
1st
2nd
3rd
4th
Frequency of Class Attendance


Distribution of Number of Courses
Frequency
0
10
20
30
40
50
60
70
80
0
2
4
6
8
Number of Courses

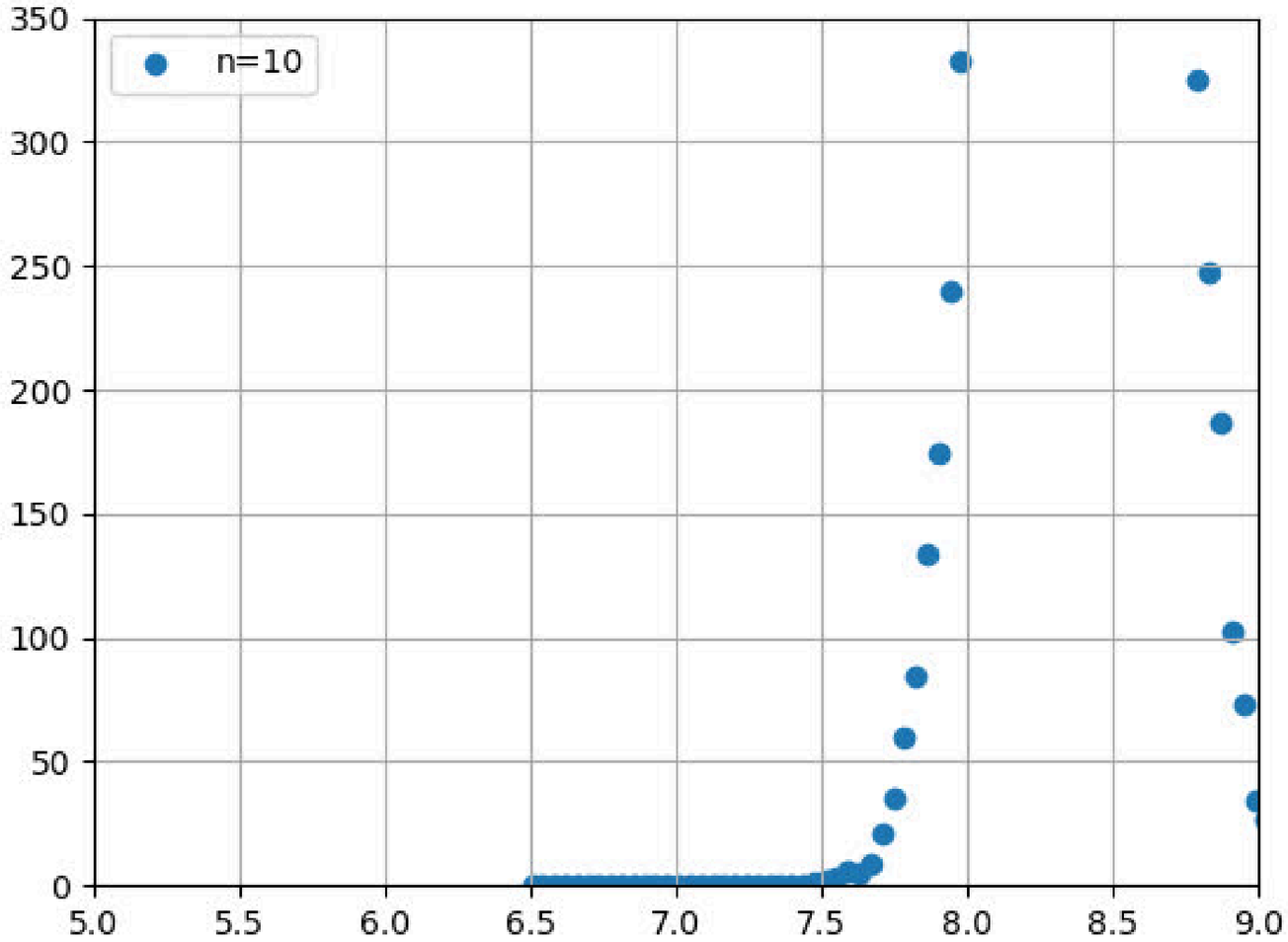

Number of Courses
1
2
3
4
5
6
7
8
9
CS/AI/MnC
EE/EP/ES/CO
ME/CH(IC)/CE/MS/BT/BM
Design

Heatmap

The heatmap shows branch-wise attendance across classes, scaled from 1 to 10. Dark green shades indicate higher attendance, light green shades lower.

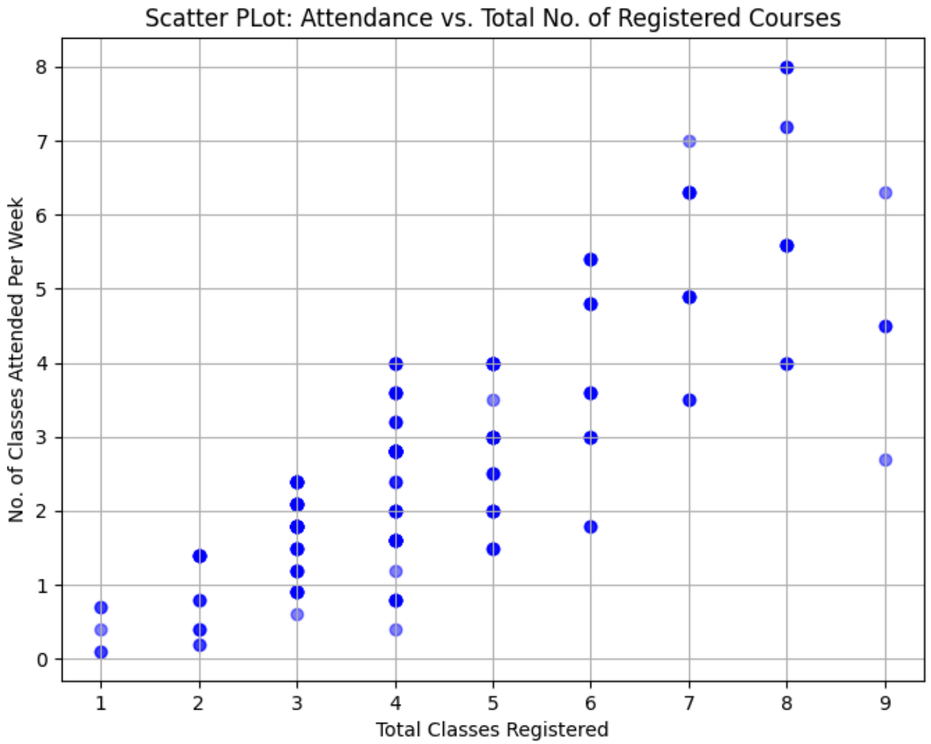


Central Limit Theorem



CORRELATIONS

Correlation Between No. Of Total Course Registered And Attendance

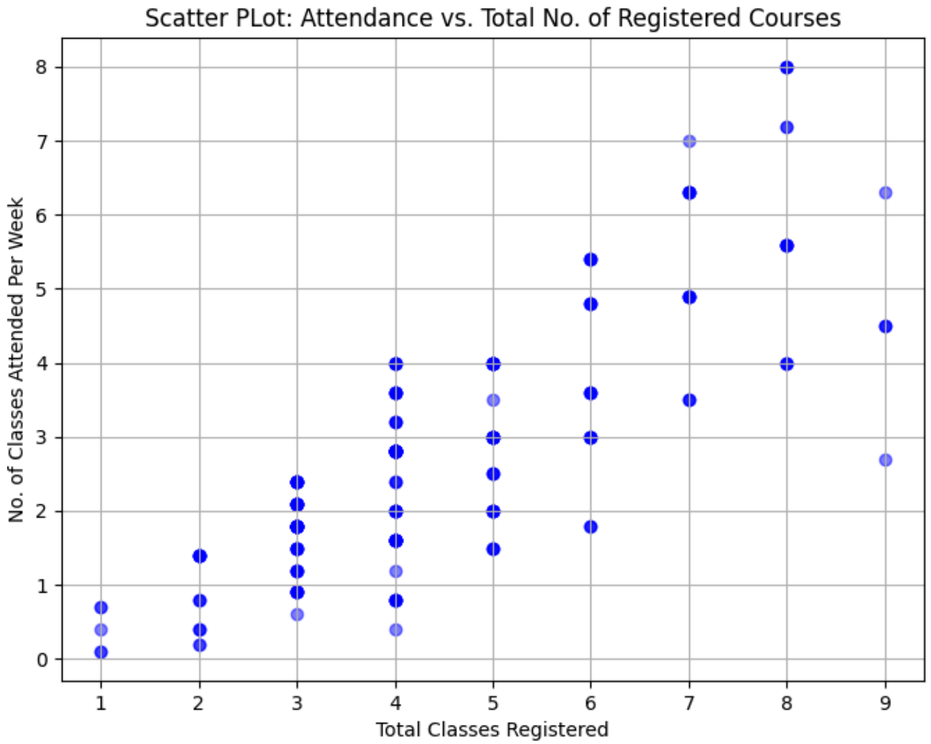


Sample Correlation Coefficient

$$r_{xy} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

ΣX	:	**Sum of values of X**
Var(X)	:	**Variance of X**
Cov(X, Y)	:	**Covariance of X and Y**

Correlation Between No. Of Total Course Registered And Attendance



The correlation coefficient ranges between -1 and 1:

$r_{xy} = 1$ - indicates a perfect positive linear relationship

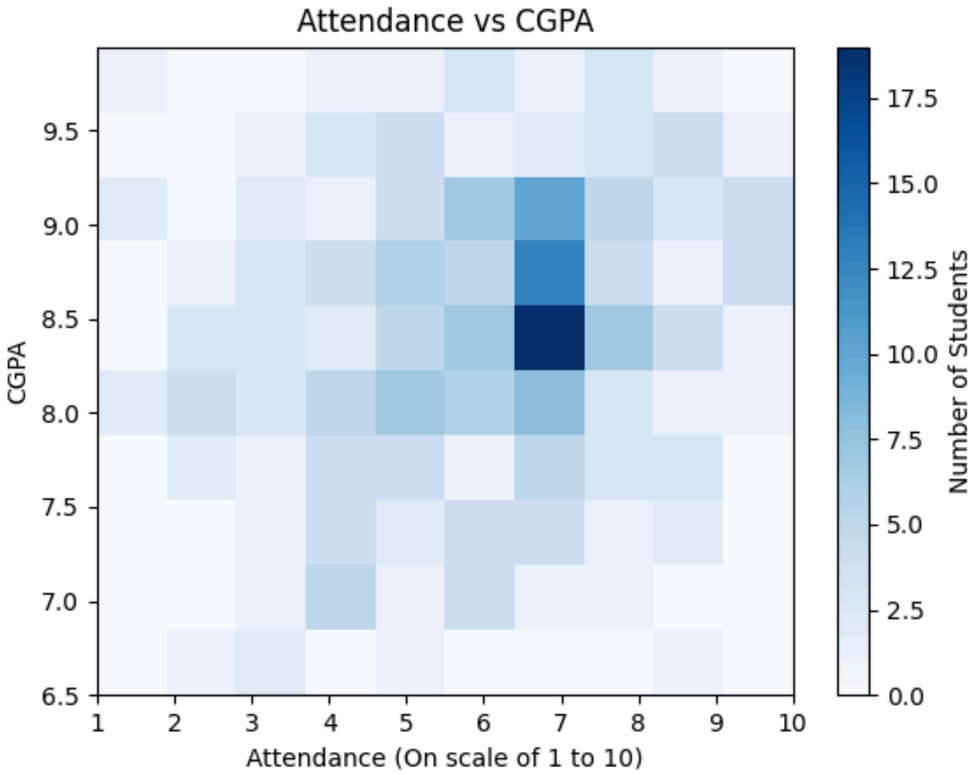
$r_{xy} = 0$ - indicates no linear relationship between X,Y

$r_{xy} = -1$ - indicates a perfect negative linear relationship

Conclusion:

$r_{xy} = 0.8$ - **Indicates a strong linear relationship between Attendance and No. of course Registered**

Correlation Between Attendance And Academic Performance

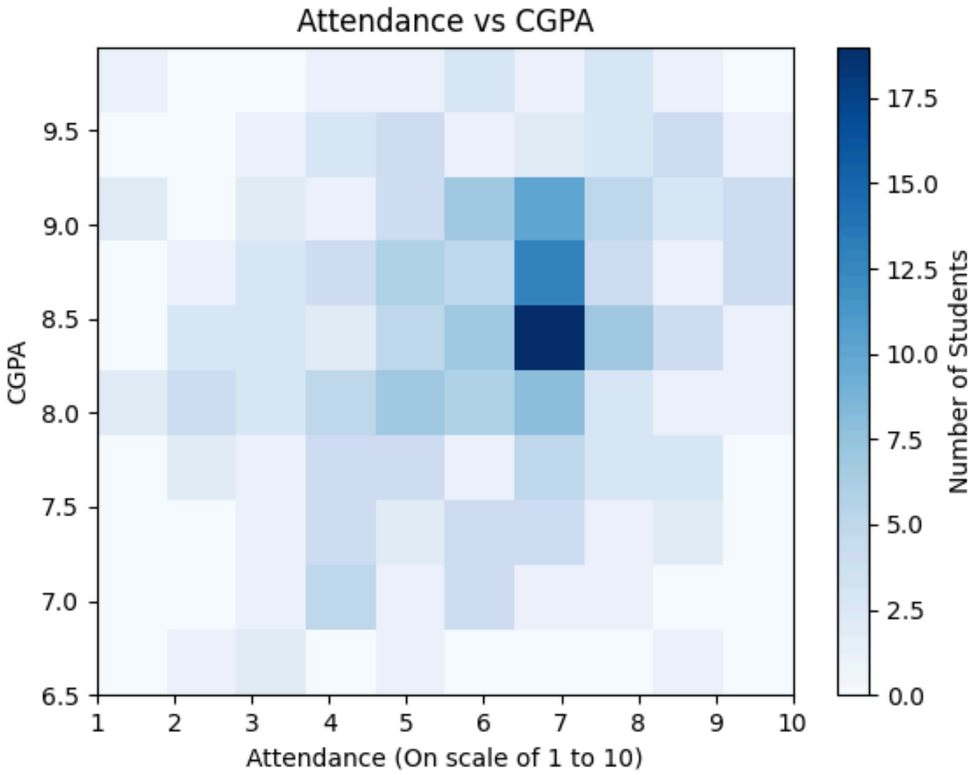


Sample Correlation Coefficient

$$r_{xy} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

ΣX	:	**Sum of values of X**
Var(X)	:	**Variance of X**
Cov(X, Y)	:	**Covariance of X and Y**

Correlation Between Attendance And Academic Performance



The correlation coefficient ranges between -1 and 1:

$r_{xy} = 1$ - indicates a perfect positive linear relationship

$r_{xy} = 0$ - indicates no linear relationship between X,Y

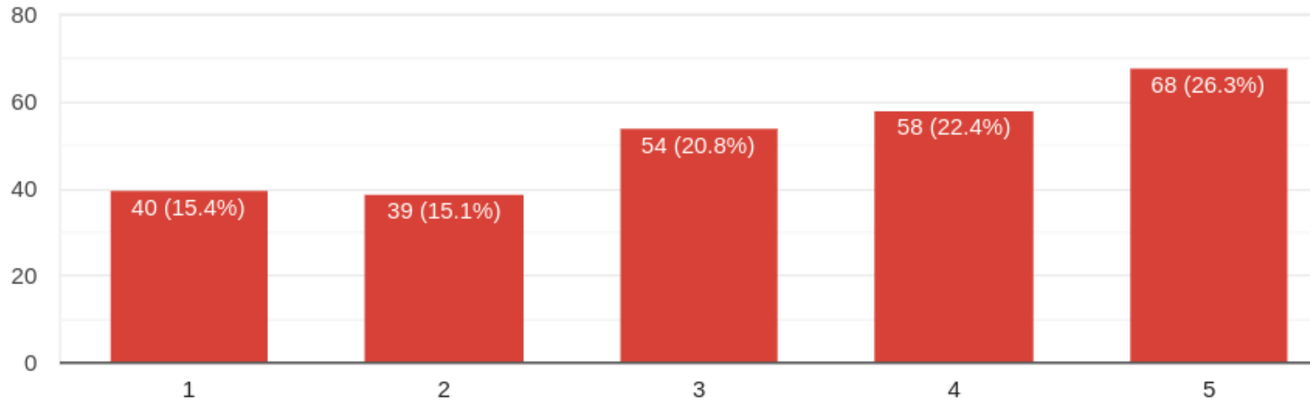
$r_{xy} = -1$ - indicates a perfect negative linear relationship

Conclusion:

$r_{xy} = 0.20$ - **Almost negligible correlation coefficient**
Non-linear correlation between Attendance and CGPA

CONFIDENCE INTERVALS

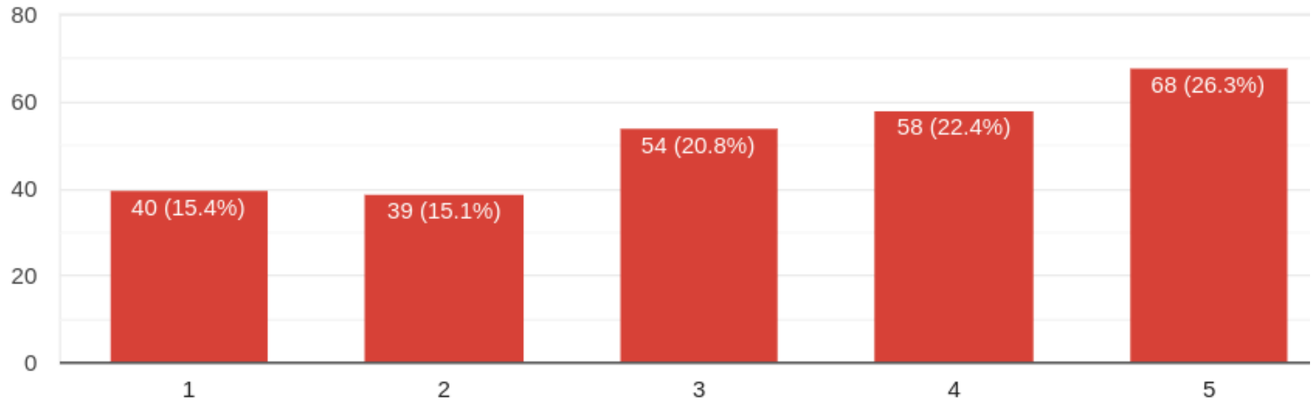
Confidence Interval for mean attendance in classes with gap in between.



CI of the mean when sigma is unknown is given by,

$$\bar{x} \pm t_{\alpha/2,n-1}\left(\frac{S}{\sqrt{n}}\right)$$

Confidence Interval for mean attendance in classes with gap in between.



We're 95% confident that students with time gaps between classes have an attendance likelihood between 3.119 and 3.461 on a 1 to 5 scale.

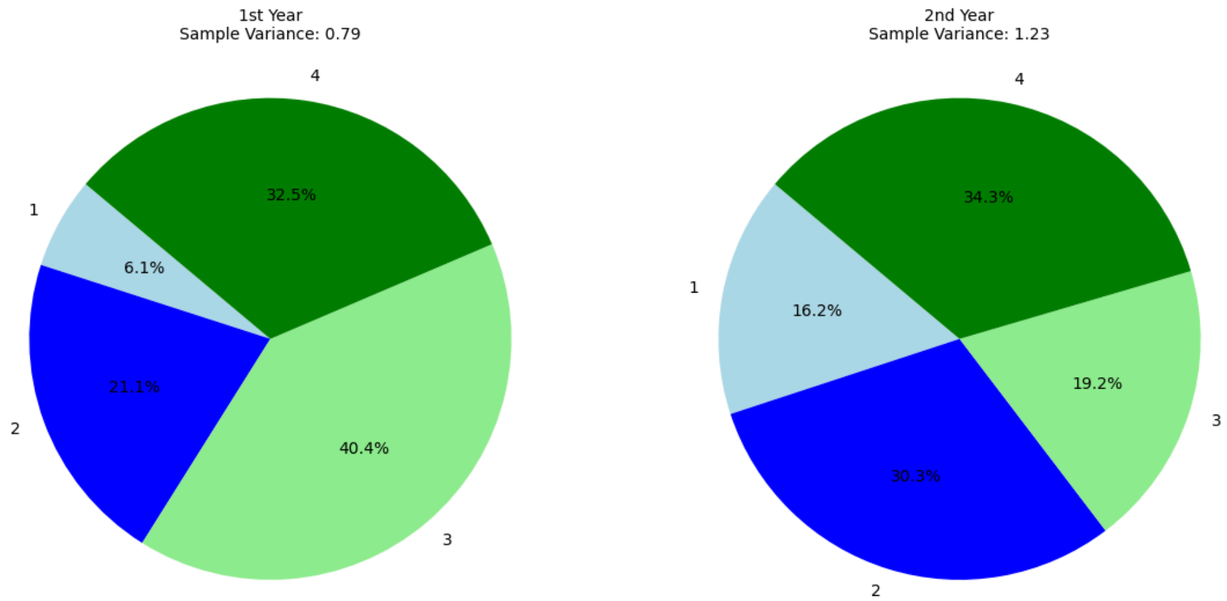
CI of the mean when sigma is unknown is given by,

$$\bar{x} \pm t_{\alpha/2,n-1}\left(\frac{S}{\sqrt{n}}\right)$$

On solving we obtain,

$$(L, U) = (3.119, 3.461)$$

Confidence Interval for Variance Ratio of 1st and 2nd Year Students' Class Attendance

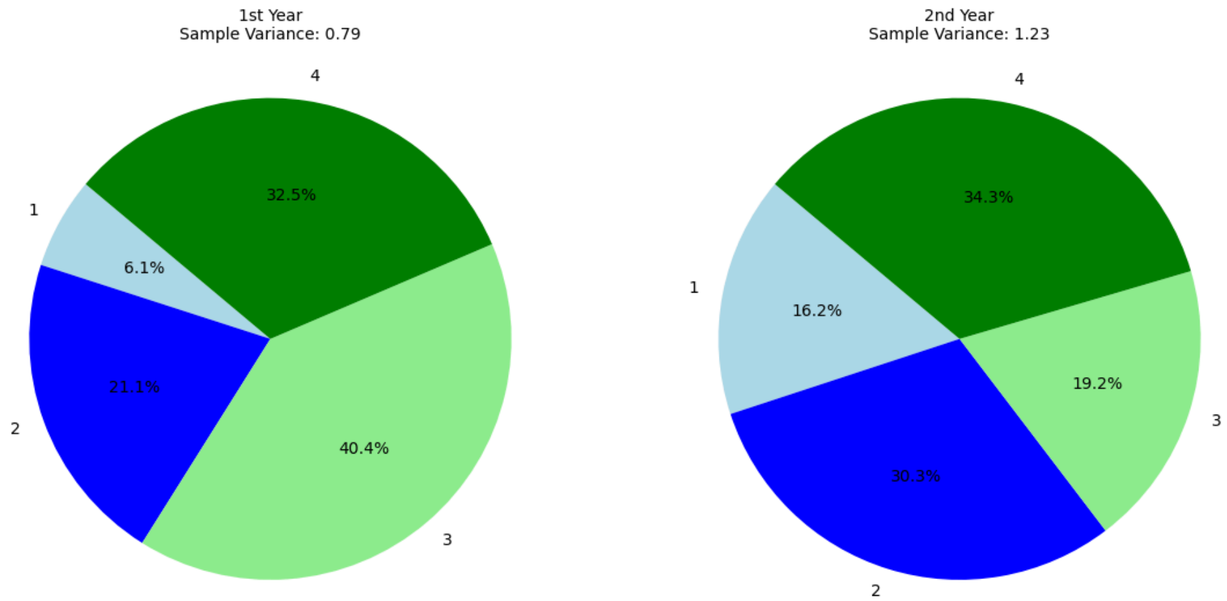


The Frequency of Attendance - 1st year vs 2nd year

CI for ratio of population variance is given by,

$$\left(\frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{\alpha/2, n_1-1, n_2-1}}, \frac{S_1^2}{S_2^2} \cdot F_{\alpha/2, n_2-1, n_1-1} \right)$$

Confidence Interval for Variance Ratio of 1st and 2nd Year Students' Class Attendance



The Frequency of Attendance - 1st year vs 2nd year

CI for ratio of population variance is given by,

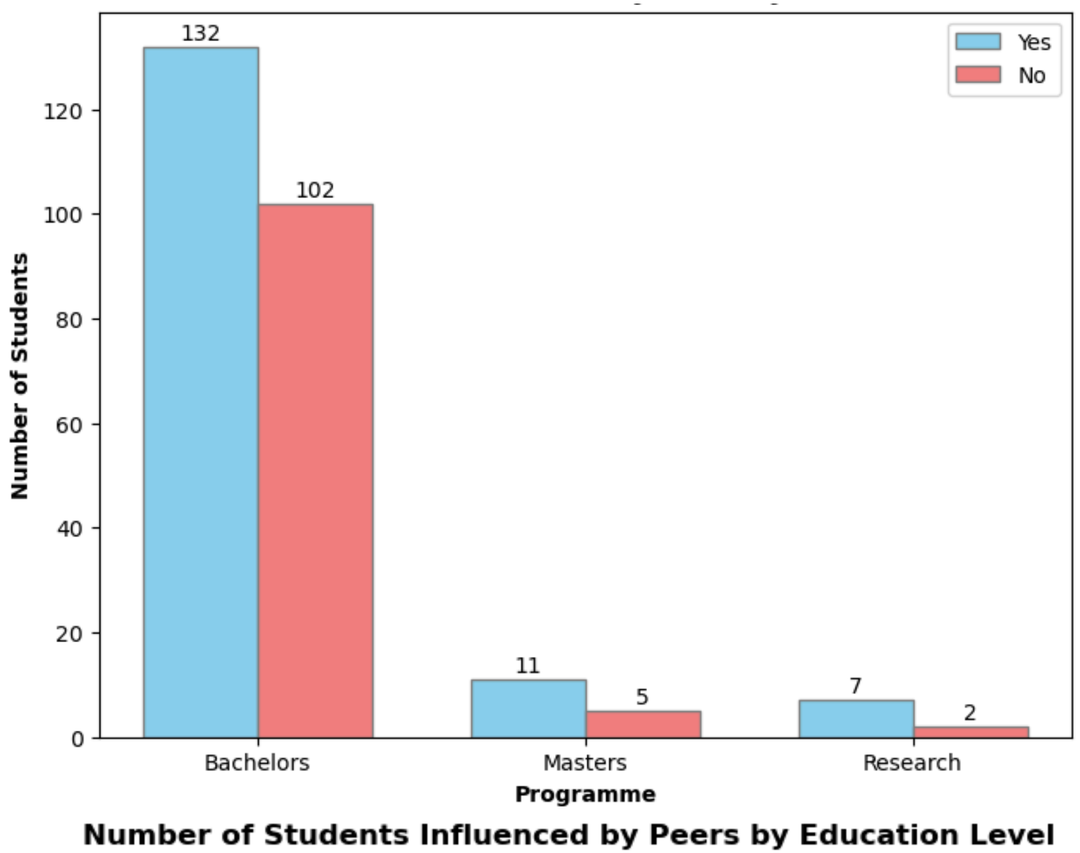
$$\left(\frac{S_1^2}{S_2^2} \cdot \frac{1}{F_{\alpha/2, n_1-1, n_2-1}}, \frac{S_1^2}{S_2^2} \cdot F_{\alpha/2, n_2-1, n_1-1} \right)$$

On solving we obtain,

$$(L, U) = (0.646, 1.039)$$

We are 95% confident that ratio of variances of frequency of 1st and 2nd year students attending the classes lies in this interval.

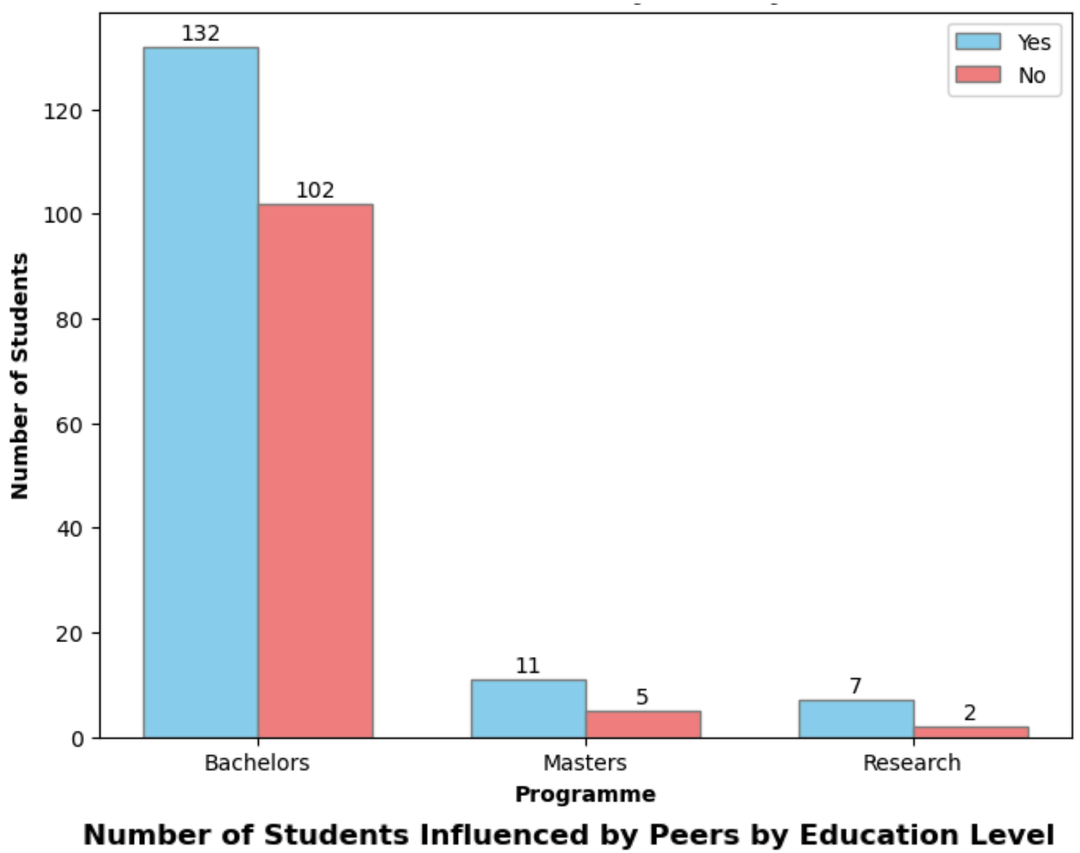
Confidence Interval for Proportional Disparity in Attendance: UG vs. PG Students



Confidence Interval for Difference in Proportions

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \left(\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right)$$

Confidence Interval for Proportional Disparity in Attendance: UG vs. PG Students



Number of Students Influenced by Peers by Education Level

Confidence Interval for Difference in Proportions

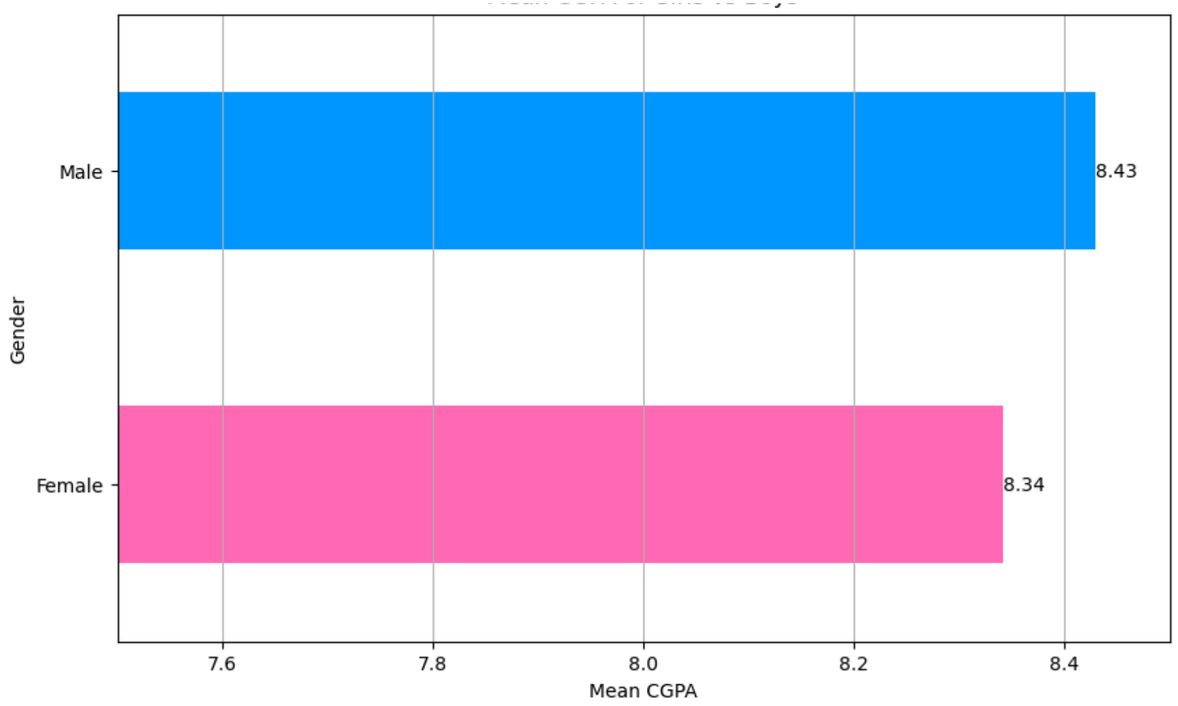
$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \left(\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right)$$

On solving we obtain,

$$(L, U) = (-0.799, 0.539)$$

With confidence level 99%, we can say that proportions of UG student effected by peers is same those of PGs.

Confidence Interval for Difference in mean CGPA: Male vs Female

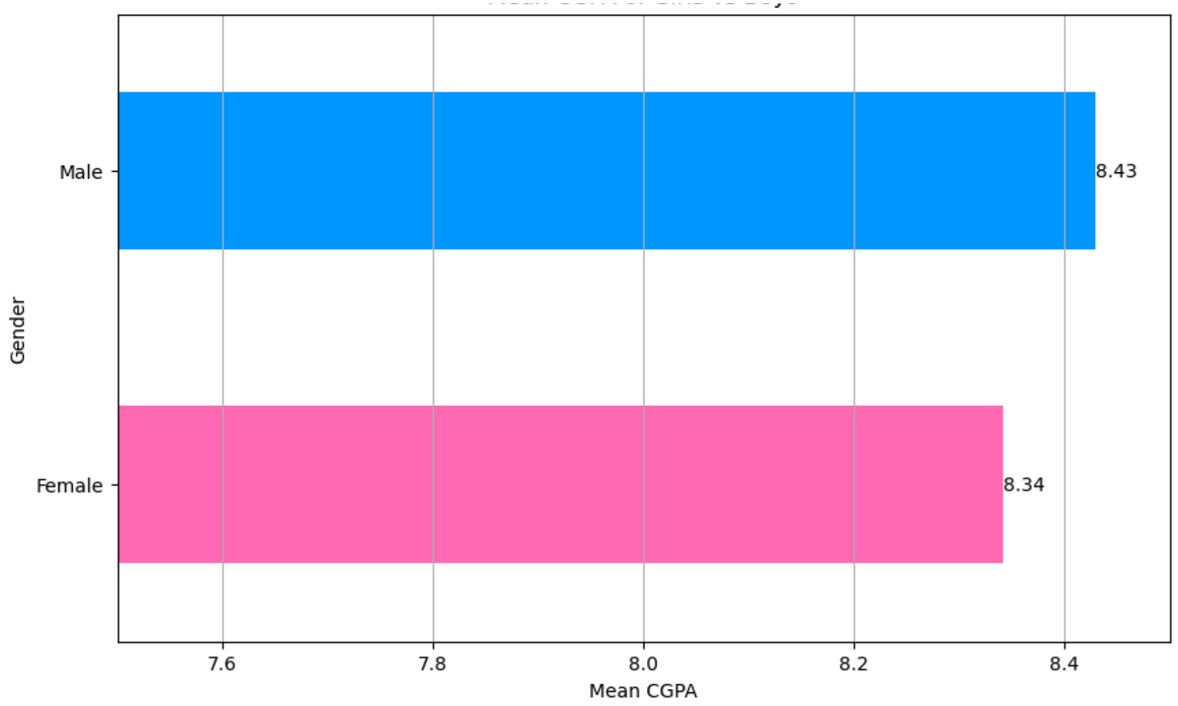


Mean CGPA- Male vs Female

CI for Difference in Population means is given by,

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,r} \left(\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right)$$

Confidence Interval for Difference in mean CGPA: Male vs Female



Mean CGPA- Male vs Female

CI for Difference in Population means is given by,

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2,r} \left(\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right)$$

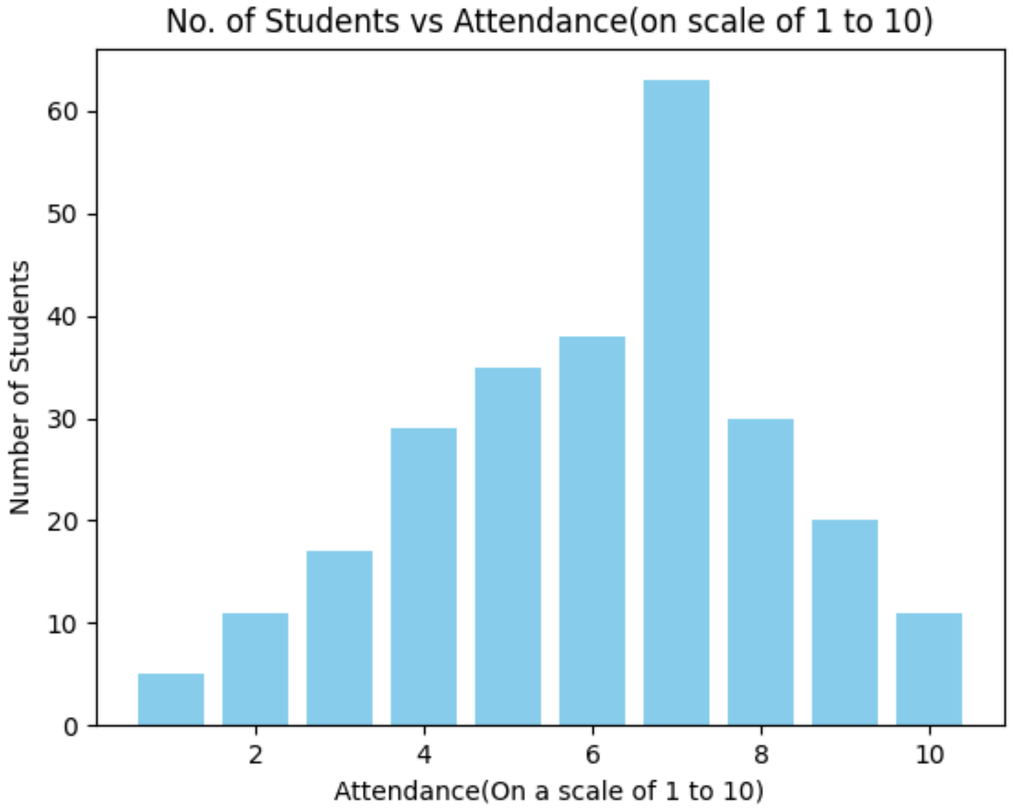
On solving we obtain,

$$(L, U) = (-0.34, 0.16)$$

With 99% confidence, we can say that difference between the mean CGPA of female and male students lies in this interval.

HYPOTHESIS TESTING

Hypothesis Testing For Average Student Attendance



Null Hypothesis:

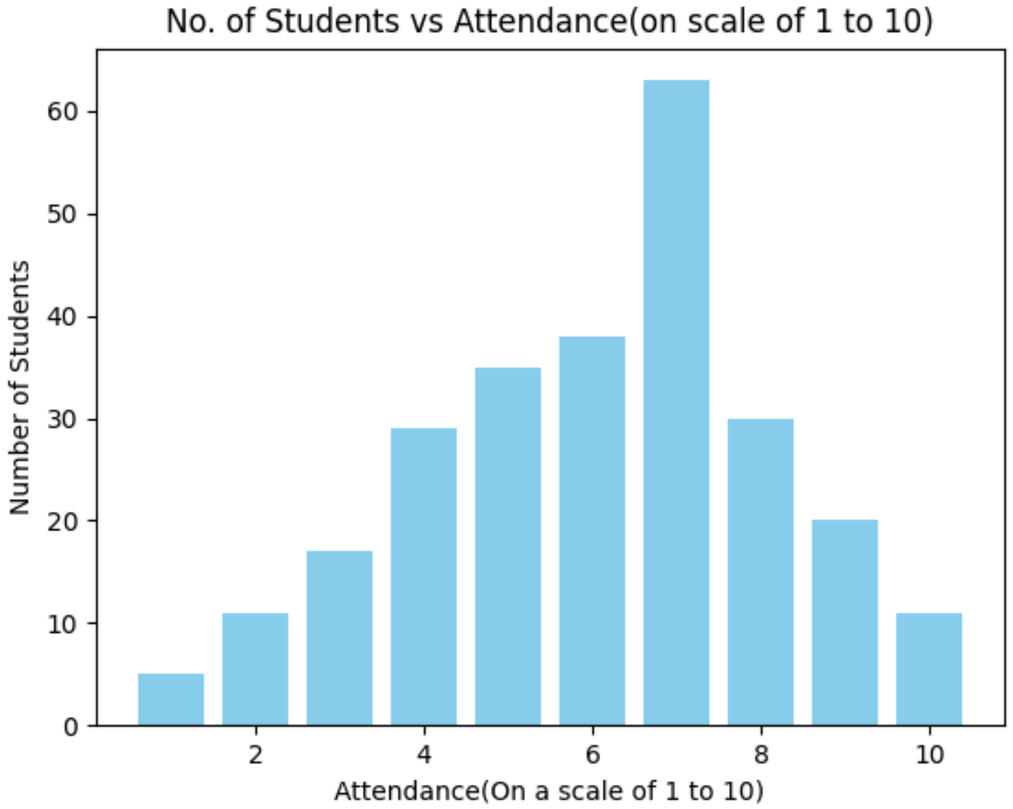
$$H_0 : \mu \leq 0.5$$

Alternate Hypothesis:

$$H_a : \mu > 0.5$$

μ is the average student attendance

Hypothesis Testing For Average Student Attendance



Conclusion:

Since our test statistic,

$$t^* = 8.06 \geq t_{\alpha,df} = 1.65$$

we reject the null hypothesis (H0)

Hence, we conclude that an average student attends more than 50% of the classes.

Hypothesis Testing for Influence of Exam Scores on Attendance

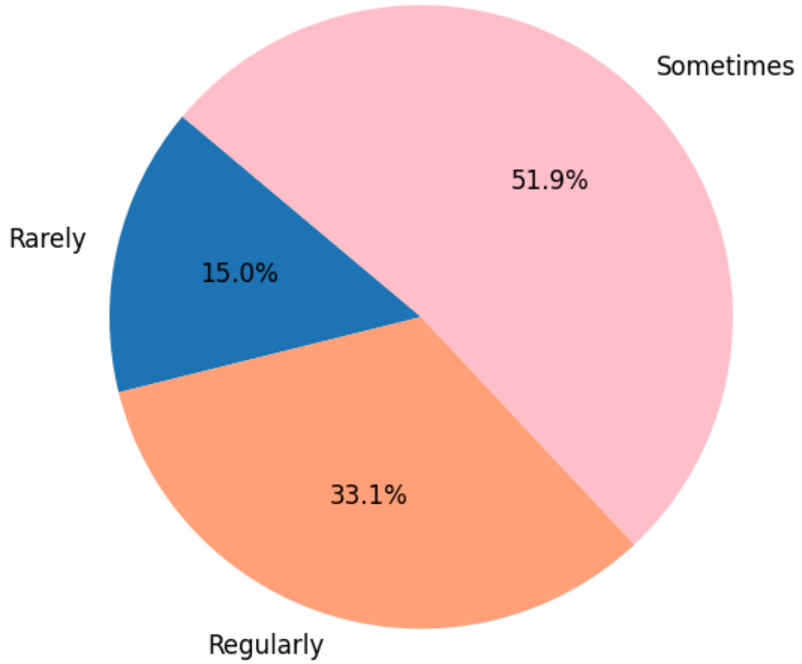


Fig: Percentage of Students who attend more classes after getting a low score

Null Hypothesis:

$$H_0 : p \leq 0.33$$

Alternate Hypothesis:

$$H_a : p > 0.33$$

p is the proportion of students who gets influenced after getting low score in exams

Hypothesis Testing for Influence of Exam Scores on Attendance

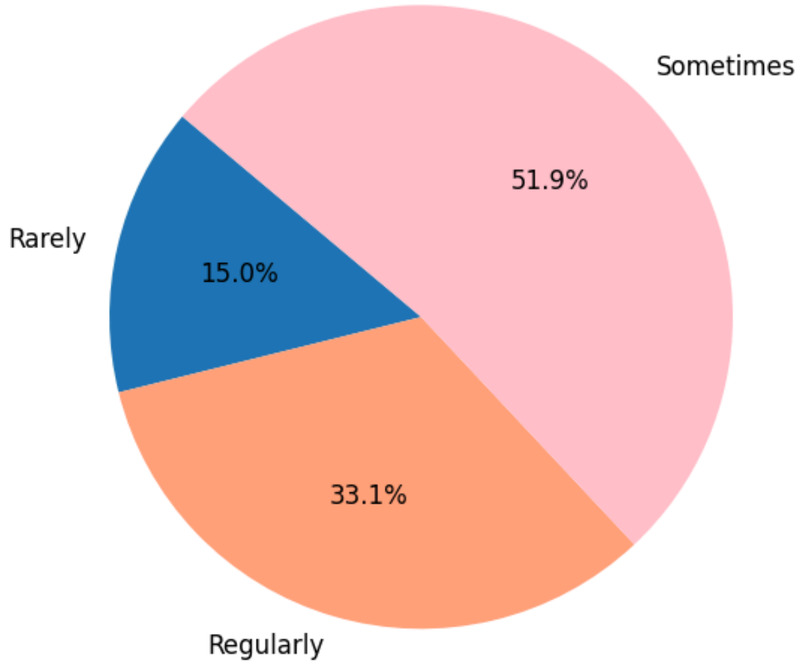


Fig: Percentage of Students who attend more classes after getting a low score

Conclusion:

Since our test statistic,

$$z^* = 2.19 \geq z_\alpha = 1.625$$

we reject the null hypothesis

Hence, we can conclude that more than 33% of the IIT-H student population attends more classes after getting a low score in previous quiz/mid-sem

Hypothesis Testing for Variation in Number of Courses of Freshers & Seniors

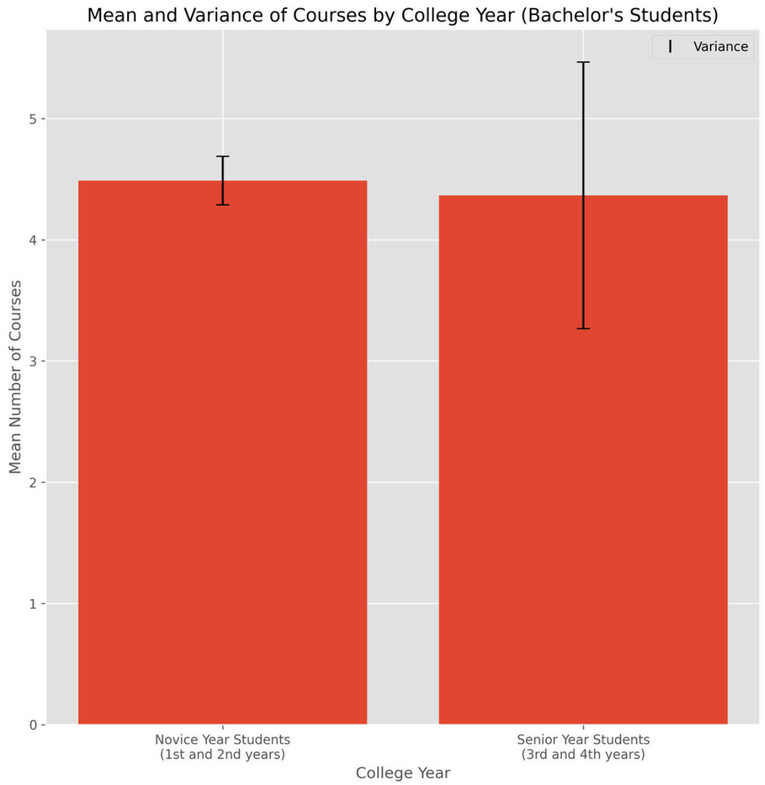


Fig: Mean no. of courses by college year

Null Hypothesis:

$$H_0 : \sigma^2_{\text{novices}} \geq \sigma^2_{\text{seniors}}$$

Alternate Hypothesis:

$$H_a : \sigma^2_{\text{novices}} < \sigma^2_{\text{seniors}}$$

sigma is the variance in no. of courses taken

Hypothesis Testing for Variation in Number of Courses of Freshers & Seniors

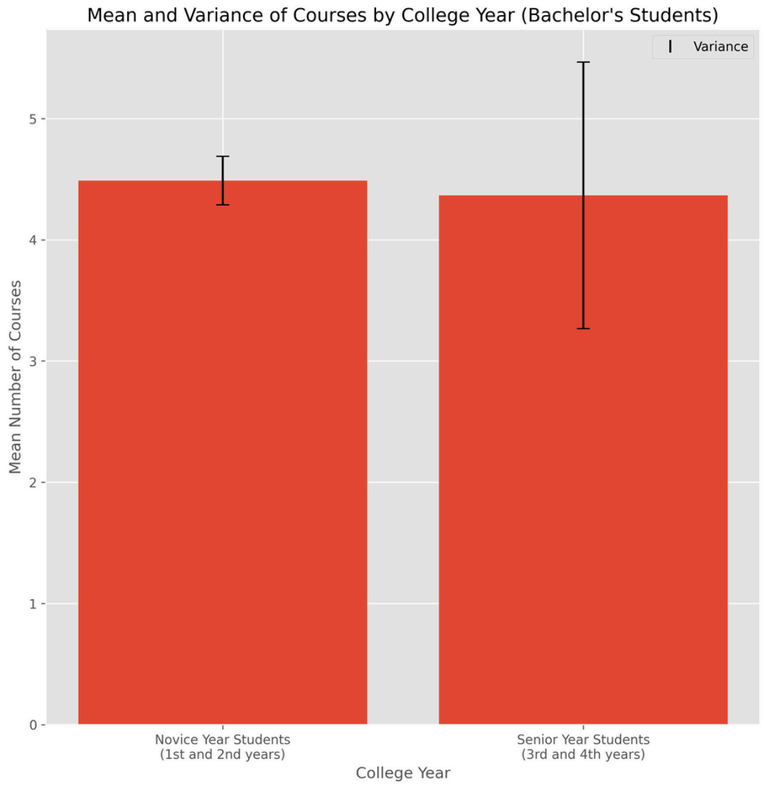


Fig: Mean no. of courses by college year

Conclusion:

We get,

$$F^* = 0.18 \leq F_{1-\alpha, df_1, df_2} = 0.81$$

so we reject the null hypothesis

Thus, we conclude that senior year (3rd and 4th Year) students have more variation in number of courses taken compared to the novice year (1st and 2nd Year) students

Hypothesis Testing for Dependent Samples: Attendance app effect

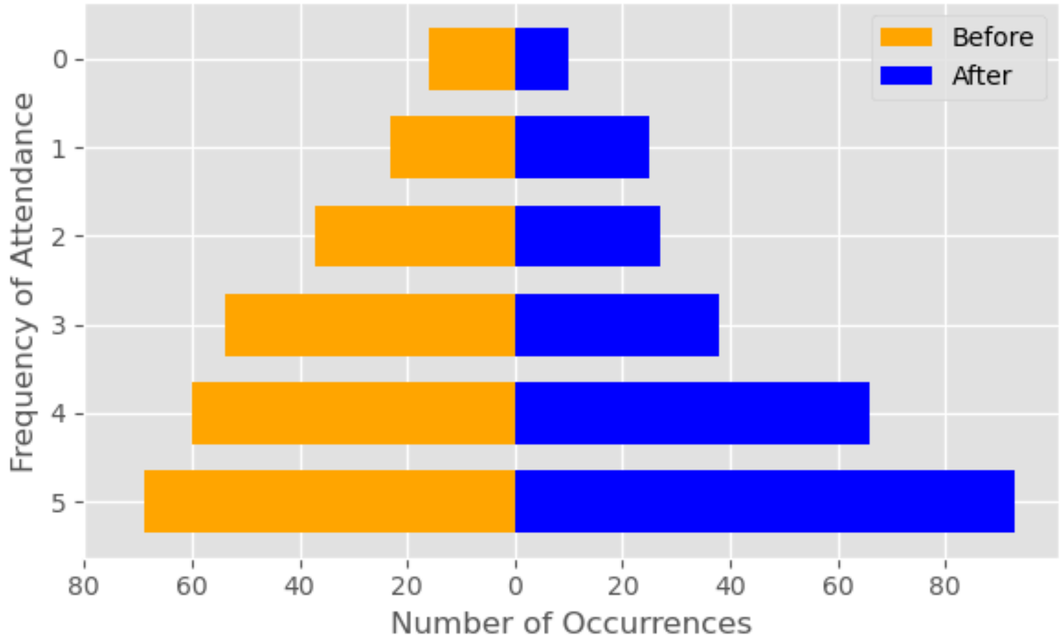


Fig: Bar graph showing the difference in attendance before and after launch of app

Null Hypothesis:

$$H_0 : \mu_D \leq 0$$

Alternate Hypothesis:

$$H_1 : \mu_D > 0$$

μD is mean difference for dependent samples

Hypothesis Testing for Dependent Samples: Attendance app effect

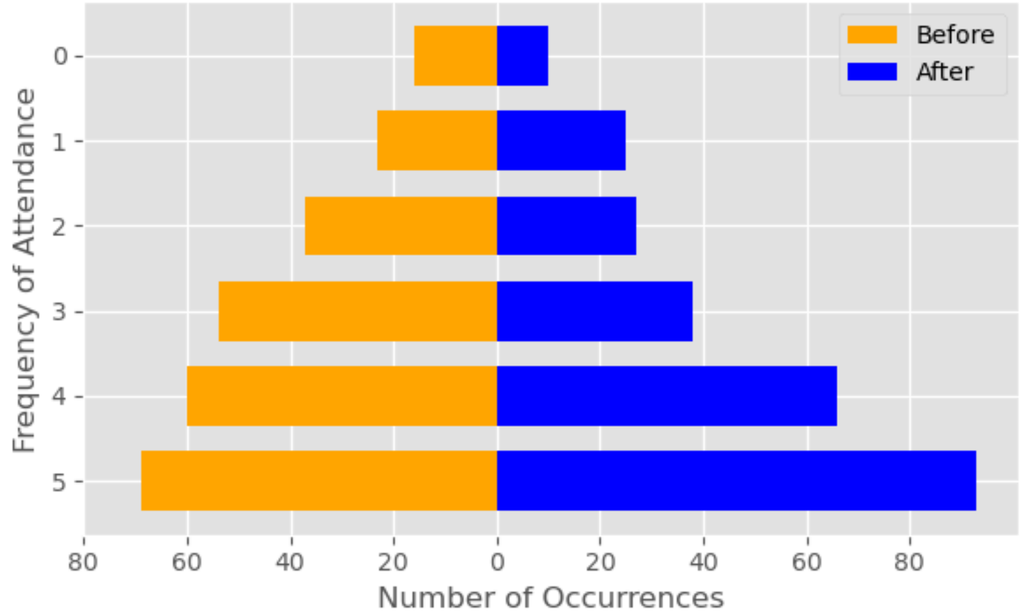


Fig: Bar graph showing the difference in attendance before and after launch of app

Conclusion:

$t^* > t_{\alpha,df}$ - **We can reject the null hypothesis**

Therefore, we conclude that there is an increase in attendance after the launch of the attendance app.

CHI SQUARED
GOODNESS FIT TEST

Chi Squared Test to check if peers influence people going to class

Category	Frequency
Peer affecting	109
Peer not affecting	150

Null Hypothesis:

$$p_1 = p_2 = 1/2$$

Alternate Hypothesis:

$$p_1 \neq 1/2$$

Where p1 and p2 represent the probabilities of individual going to classes depending on peers.

Chi Squared Test to check if peers influence people going to class

Category	Frequency
Peer affecting	109
Peer not affecting	150

We calculate :

$$\chi^2 = \sum_{i=1}^{2} \frac{(O_i - E_i)^2}{E_i}$$

Where Ei is the expected frequency and Oi is the original frequency.

We obtain :

$$\chi^2 > \chi^2_{1,0.05}$$

Therefore, we reject the null hypothesis.

Chi-Squared Test for gender-wise preference time of attending classes

Gender	Morning	Afternoon	Evening
Male	123	115	43
Female	57	33	14

Null Hypothesis:

There is no association between Gender and class time preferences.

Alternate Hypothesis:

There is association between Gender and class time preferences.

Chi-Squared Test for gender-wise preference time of attending classes

Gender	Morning	Afternoon	Evening
Male	123	115	43
Female	57	33	14

We calculate :

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Where Ei is the expected frequency and Oi is the original frequency.

We obtain :

$$\chi^2 < \chi^2_{2,0.05}$$

Therefore, we fail to reject the null hypothesis.

CONCLUSIONS

Conclusions

- The correlation coefficient between the **number of courses** registered and **attendance** indicates a **high correlation**, with a value of approximately 0.8.
- The correlation between **attendance and CGPA is negligible**, with a very small correlation coefficient of around 0.2.
- The introduction of **attendance app** has been **successful**, as evidenced by increased class attendance compared to before, as supported by hypothesis testing.

Conclusions

- The variation in the **number of courses** taken by **first and second-year** students is considerably lower compared to the variation observed in **third and fourth-year** students.
- More than **33% of students** are influenced by their low exam marks and are **motivated to attend more classes**.
- **Peers** significantly **influence** a student's attendance in class.
- The **average wake-up time is 7:58 am**, which is conducive to establishing a healthy routine for attending classes throughout the day.

THANK YOU!

Scan this code to access the GitHub repository.