A Statistical Study on Attendance in IIT Hyderabad

Exploring Attendance Trends through Statistical Analysis

ఐఐటీ హైదరాబాద్
आई आई टी हैदराबाद
IIT Hyderabad

MA2540 Applied Statistics

Collaboratively done by:

Abhishek Jha (MA22BTECH11002)
Ananya S Reddy (MA22BTECH11004)
Mayank Parasramka (AI22BTECH11018)
Pranay Jain (AI22BTECH11020)
Yasir Usmani (AI22BTECH11031)

Under the Guidance of

Prof. Sameen Naqvi

Contents

1 Overview

Abstract

This study investigates how often students attend classes at IITH. Attendance is really important for understanding how engaged students are and how well they're doing in their studies. Using a dataset collected through a survey conducted this semester, this research explores various factors influencing attendance, including number of courses, timing of classes, and student demographics.

Data collection

In order to gather sample data for our statistical analysis, we created a questionnaire and distributed it as a Google Form among the students of IIT Hyderabad. To encourage maximum participation, we sent out frequent reminders to complete the survey. Additionally, we ensured the anonymity of the survey to alleviate any privacy concerns that might deter participation. Ultimately, we received **259** valid responses, providing a robust dataset for our study.

The survey contained followed questions:

- Gender
- Branch
- Programme
- Current year
- How many courses do you have in current semester?
- How many hours of classes do you have in a week?
- How frequently do you attend classes?
- You are most likely to attend classes during (Morning,Afternoon,Evening)
- How many courses do you have that has mandatory or random attendance, surprise quiz, marks on interaction etc?
- How likely are you to attend 2 classes that has gap in between?
- How frequently have u attended classes before launching of attendance app?
- How frequently have u attended classes after launching of attendance app?
- When do you wake up usually?
- If your peers aren't attending a class, will you attend that class?
- Do you care to attend classes after getting low score in exams?
- What is your current CGPA?

2 Preprocessing

2.1 Data Cleaning

During the preprocessing phase of the data obtained via Google Forms (over 300 responses), approximately 50 responses were deemed unsuitable for analysis due to inconsistencies, such as instances where reported Cumulative Grade Point Average (CGPA) exceeded 10 or equaled 10 at the end of 3rd semester which is highly unlikely. Furthermore, in some instances, specific responses exhibited localized inconsistencies, prompting the replacement of erroneous values with the mean derived from validated data points. Consequently, the initial dataset comprising 300 responses was refined to a final dataset comprising 259 authentic responses after meticulous data cleaning and preprocessing procedures.

2.2 Shapiro-Wilk Test for Normality for the number of courses

Given Data:

$$5, 3, 7, 5, 5, 10, ...$$

Step 1: Sort the data in ascending order. Sorted Data:

$$1, 1, 1, 2, ..., 10$$

Step 2: Compute the coefficients using a table of Shapiro-Wilk coefficients. The Shapiro-Wilk coefficients for $n = 200$ are:

$$a_1 = 0.5773, a_2 = 0.4810, a_3 = 0.4194, ..., a_{100} = 0.0001$$

Step 3: Calculate the test statistic W.

$$W = \frac{(\sum(a_i \cdot (x_{(n+1-i)} - x_i))^2)}{(\sum(x_i - \bar{x})^2)}$$

Where, $\bar{x}$ is the sample mean
Calculating the sample mean ($\bar{x}$):

$$\bar{x} = \frac{(1 + 1 + ... + 10)}{200} = 6.2$$

Calculating the sum of squared differences:

$$\sum(x_i - \bar{x})^2 = 7892.25$$

Calculating W:

$$W = \frac{(\sum(a_i \cdot (x_{(n+1-i)} - x_i))^2)}{7892.25} = 0.9812$$

Step 5: Compare the calculated W value with the critical value from the Shapiro-Wilk table to determine if the data is normally distributed or not. The critical value for $n = 200$ and $\alpha = 0.05$ is 0.9757.
Since the calculated W value (0.9812) is greater than the critical value (0.9757), we fail to reject the null hypothesis that the data is normally distributed.
Therefore, based on the Shapiro-Wilk test, the given dataset can be considered as coming from a normal distribution.

2.3 Shapiro-Wilk Test for Normality for the CGPA of students

When Shapiro-Wilk test is done for CGPA, by the method used in the previous Subpart, we obtain the test-statistic to be around 0.98 with a p-value of about 0.049. This gives us a indication that the data may be belonging to a normal distribution.
Futhermore, Q-Q' plot was also used as a method to verify the normality.

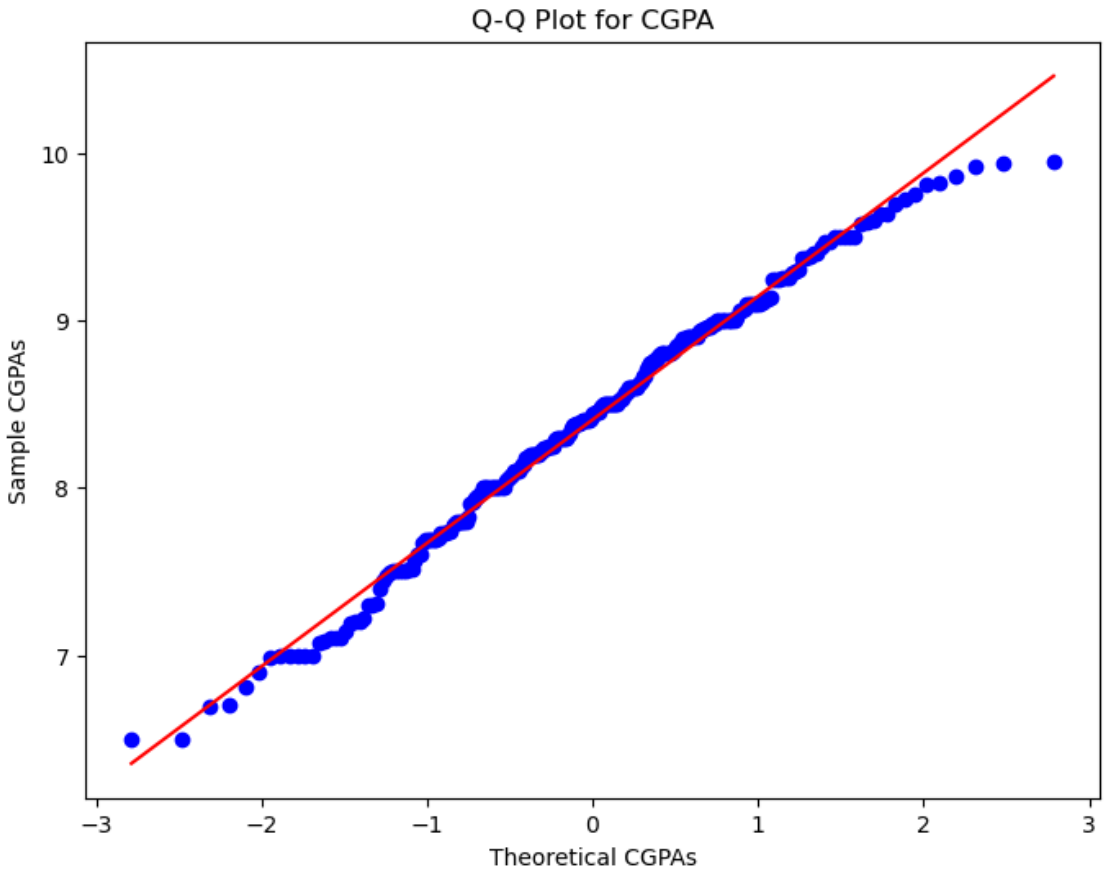


Figure 1: Q-Q' Plot for CGPA

The datapoints lie on the linear line approximately, which leads us to conclusion that the CGPA of students is likely to belong to a normal distribution.

Therefore, based on the Shapiro-Wilk test, the given dataset can be considered as coming from a normal distribution.

3 Visualisations

3.1 Segmented Bar Graph for years vs frequency

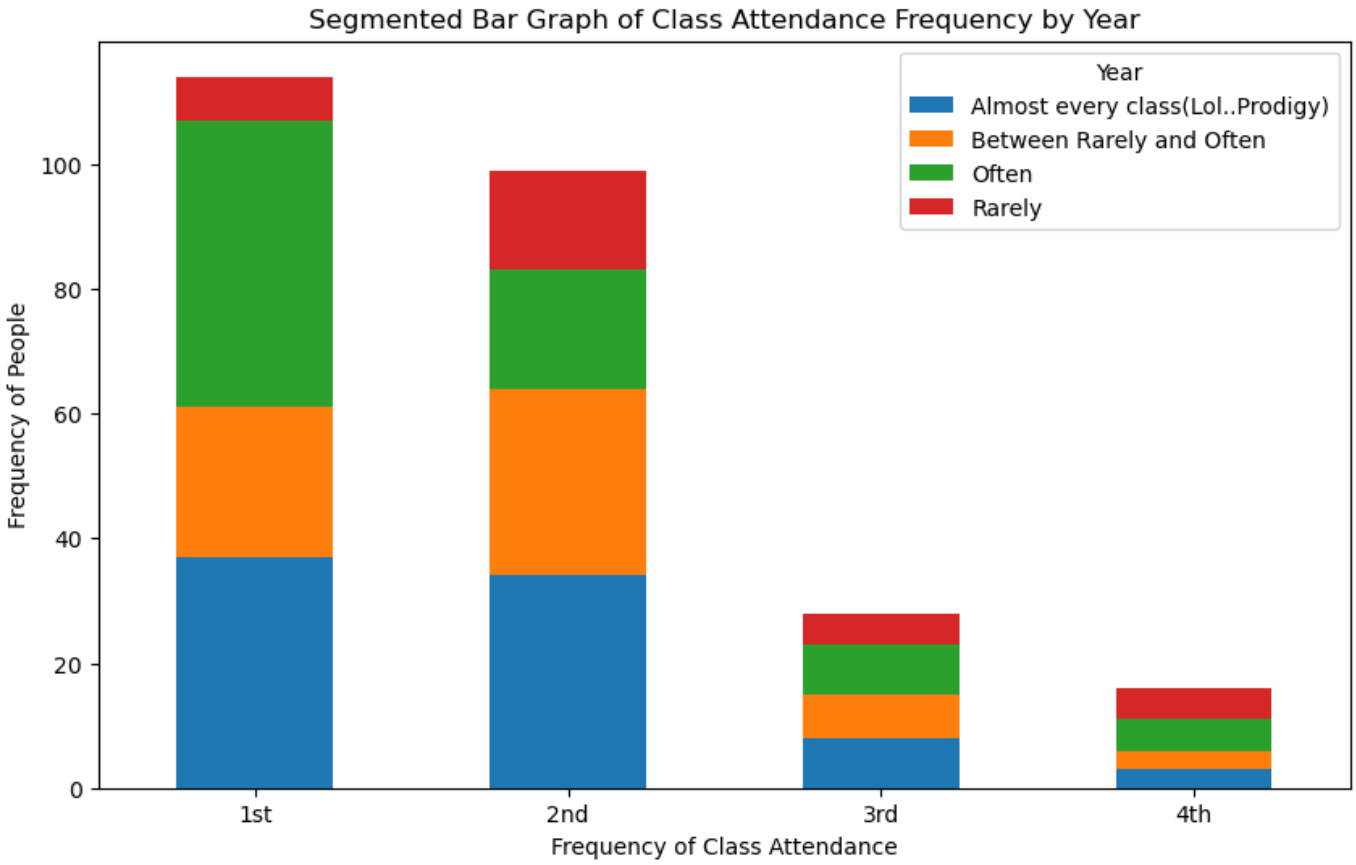


Figure 2: Bar Graph for years vs frequency

3.1.1 Insights:

Year-wise Attendance Trends:

- Among all the students, the proportion of students attending most of the classes is equally high in all the years.
- The 'Often' and 'Almost Every Class' proportion of 1st year is high as compared to other years. This can be deduced by the sincerity of the 1st year students towards their studies.
- A good proportion of 2nd and 4th year students tend to attend very less classes when compared to other years. This can be due to hectic schedule in 2nd year which leads to people missing classes,

3.2 Visuals:

Some of the graphs indicating the distribution of various inputs taken by us, and some of the illustrations of box plot, histogram are plotted here:

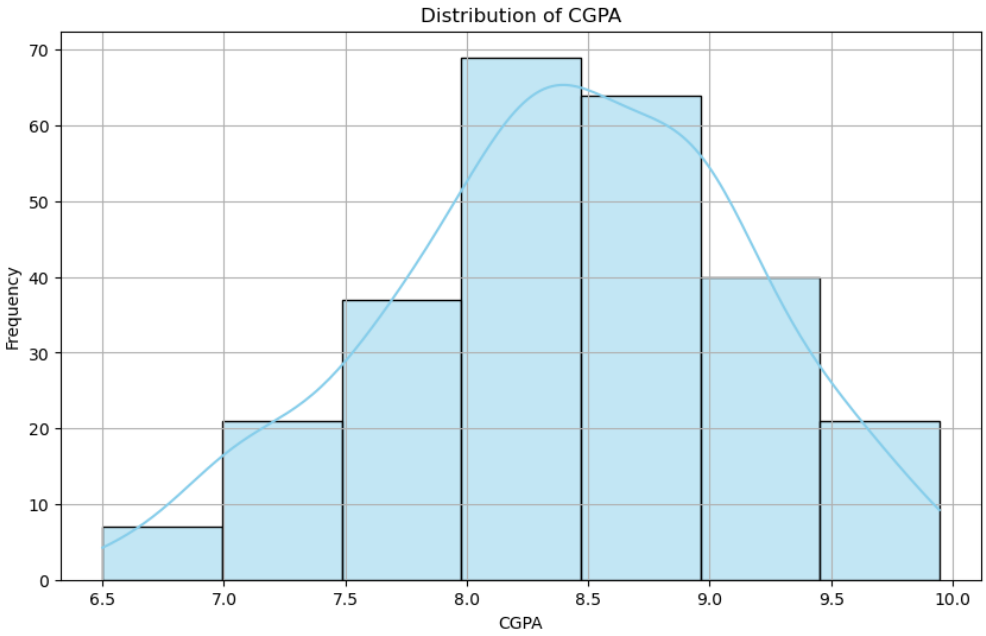


Figure 3: Graph for distribution of CGPA

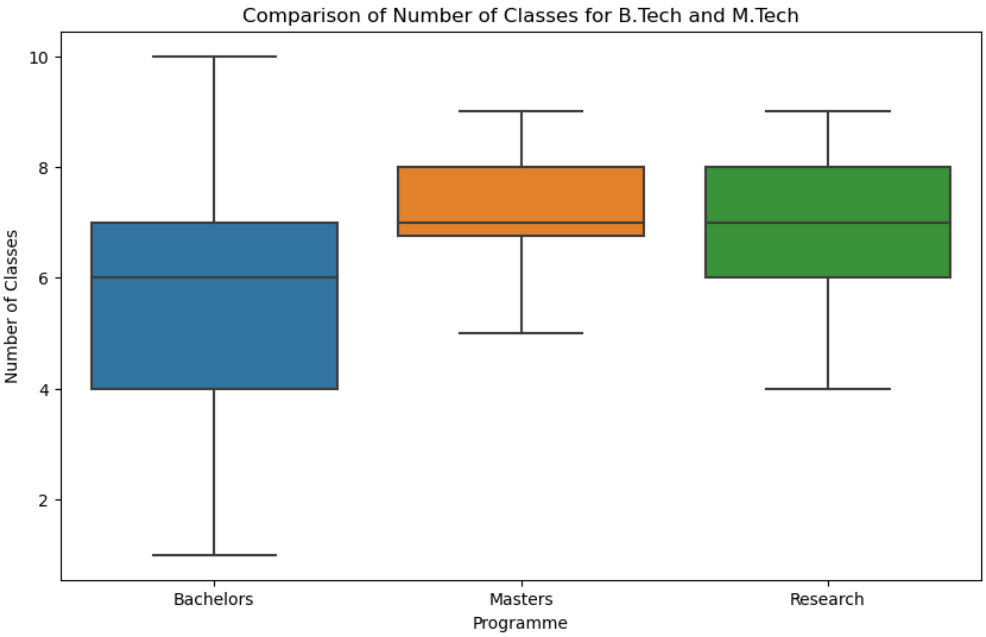


Figure 4: Box Plot of comparison of number of classes across various streams

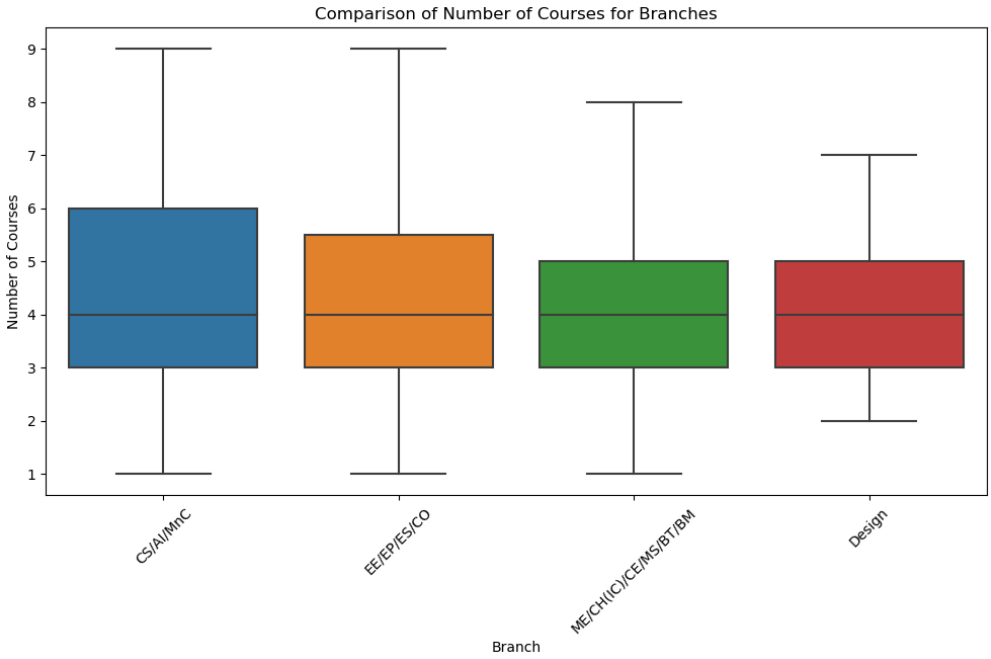


Figure 5: Box Plot of comparison of number of classes across various branches

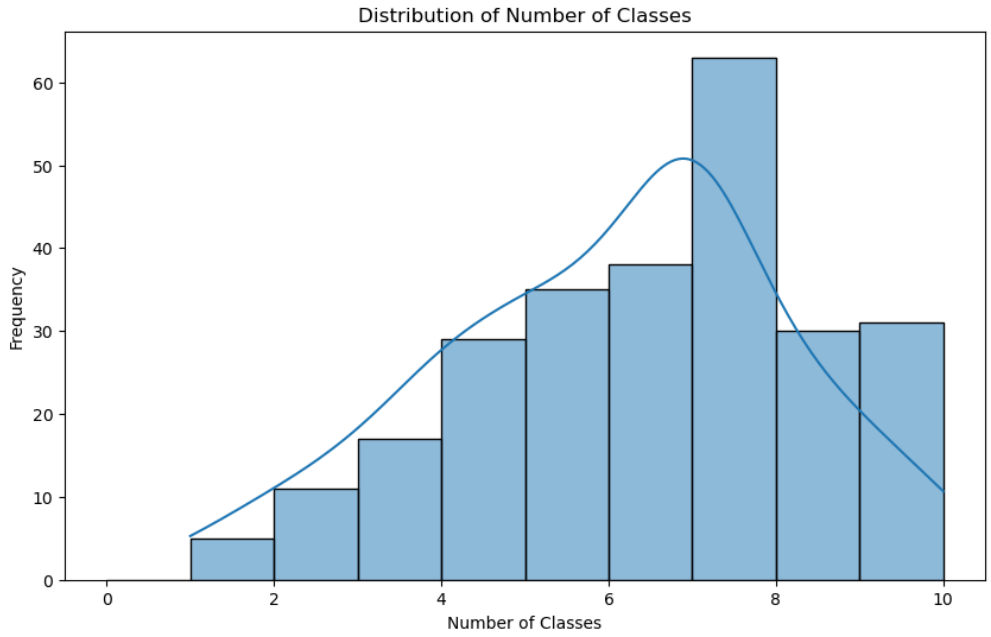


Figure 6: Bar Graph for distribution of number of classes attended by students

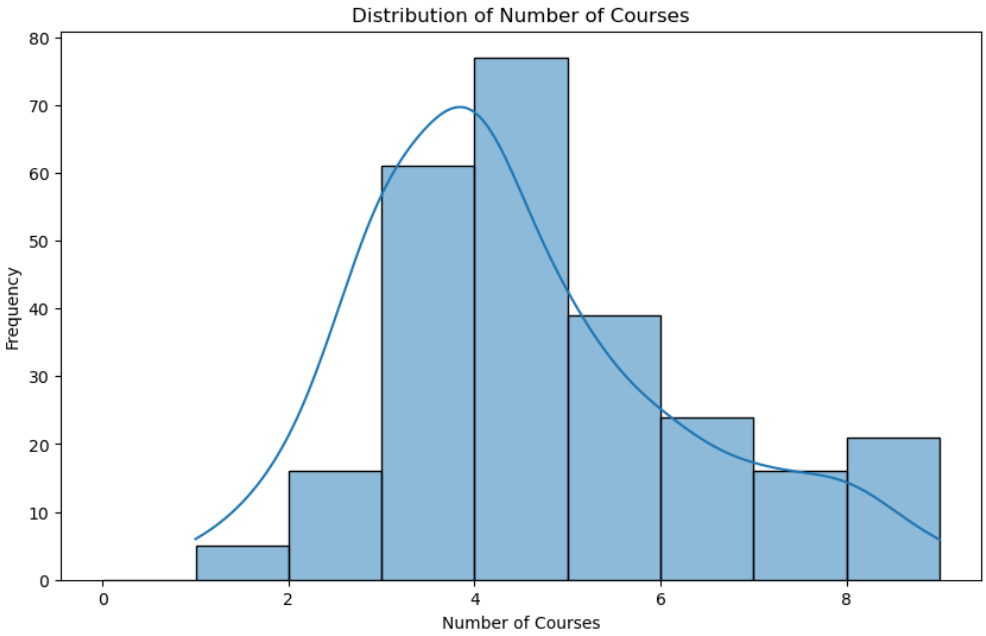


Figure 7: Bar Graph for distribution of number of courses

3.3 Heatmap of Relative Student Attendance by Branch

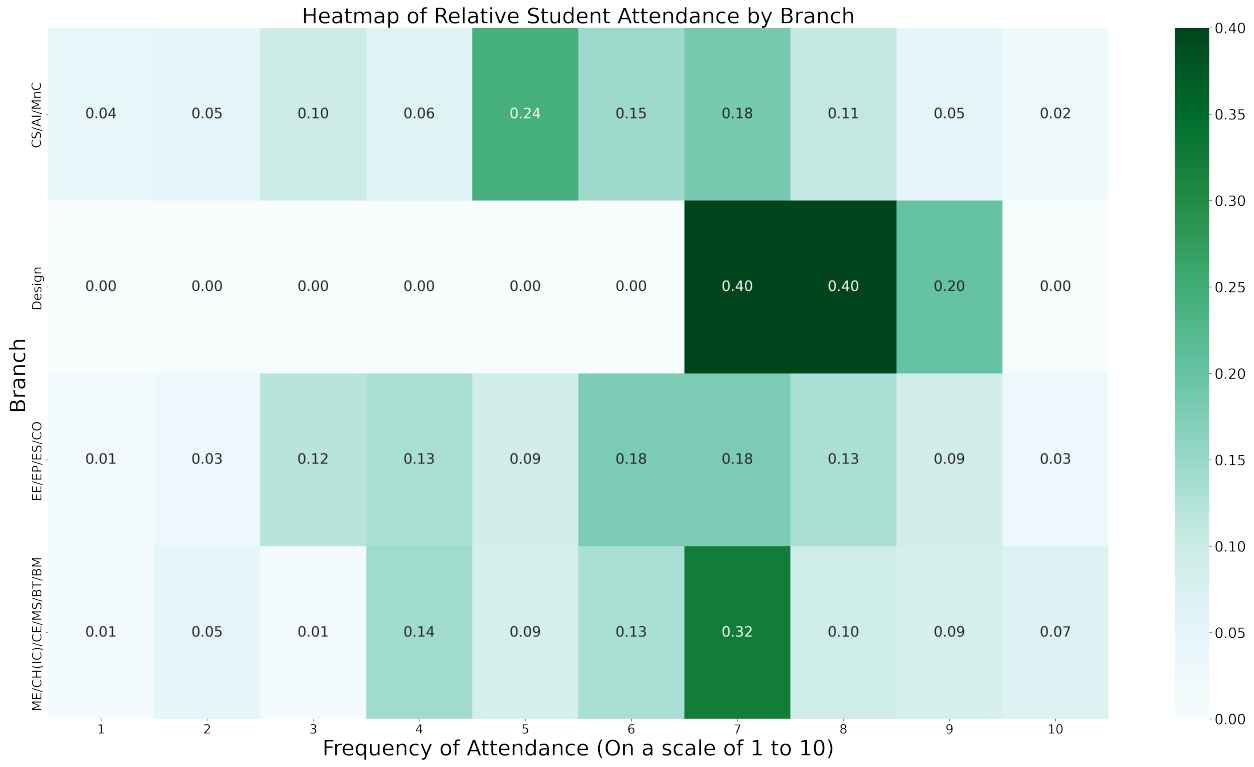


Figure 8: Heatmap for Branch vs Frequency

3.3.1 Description:

The heatmap visualizes the relative attendance of students from different branches across various classes, represented on a scale of 1 to 10. Each cell in the heatmap represents the proportion of students from a specific branch attending classes, normalized by the total number of students in that branch. The color intensity of each cell indicates the attendance frequency, with darker shades representing higher attendance rates and lighter shades indicating lower attendance rates.

3.3.2 Insights:

1. **Branch-wise Attendance Trends**:
    - **Design**: Highest proportion attending most classes.
2. **Branch-wise Attendance Distribution**:
    - **CS/AI/MnC**: 5-7 (50% attendance).
    - **Design**: 7-9 (70%+ attendance).
    - **EE/ES/EP/CO**: 6-8 (65% attendance), some 3-4 (30% attendance).
    - **ME/CH(IC)/CE/MS/BT/BM**: 7-10 (70%+ attendance).
3. **Variability in Attendance Among The frequency scale**:
   Variance:
    - **CS/AI/MnC:** 0.005
    - **Design:** 0.029
    - **EE/EP/ES/CO:** 0.004
    - **ME/CH(IC)/CE/MS/BT/BM:** 0.008

   Design shows highest variability.

4 Correlations

4.1 Correlation between no. of total course registered and attendance:

The correlation coefficient is a measure of the strength and direction of the linear relationship between two variables.

Definition

The sample correlation coefficient, often denoted by r_{xy}, is calculated using the following formula:

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

where:

- n : number of data points
- X, Y : variables
- $Var(X), Var(Y)$: Variance of X and Variance of Y
- $Cov(X,Y)$: Covariance between X,Y
- $\sum XY$: sum of the products of the corresponding values of X and Y
- $\sum X, \sum Y$: sum of the values of X and Y, respectively
- $\sum X^2, \sum Y^2$: sum of the squares of the values of X and Y, respectively

The correlation coefficient ranges from -1 to 1:

- $r_{xy} = 1$ indicates a perfect positive linear relationship.
- $r_{xy} = -1$ indicates a perfect negative linear relationship.
- $r_{xy} = 0$ indicates no linear relationship between the variables.

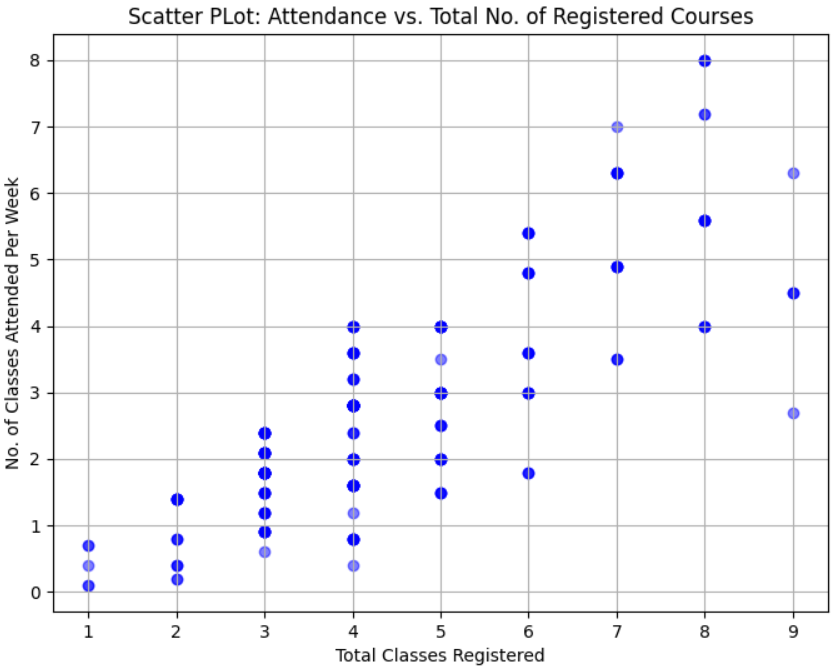


Figure 9: Scatter Plot of Attendance vs. Number of Courses Registered

On calculating we get the Correlation Coefficient for the two random variables, $r_{xy} \approx 0.8$ This implies a strong linear relationship between the two observed random variables. So we can conclude that for this sample a student taking more no. of courses is more likely to attend all the classes.

4.2 Correlation between Attendance and Academic performance:

The data for attendance is available to us as the no. of classes a student attends in a week (on a scale of 1 to 10). We also have metrics(CGPA) for assessing the academic performance of a student.

Now we want to see if there exists some relationship between the given two variables. We use Pearson correlation coefficient to find how the two dataset correlates.

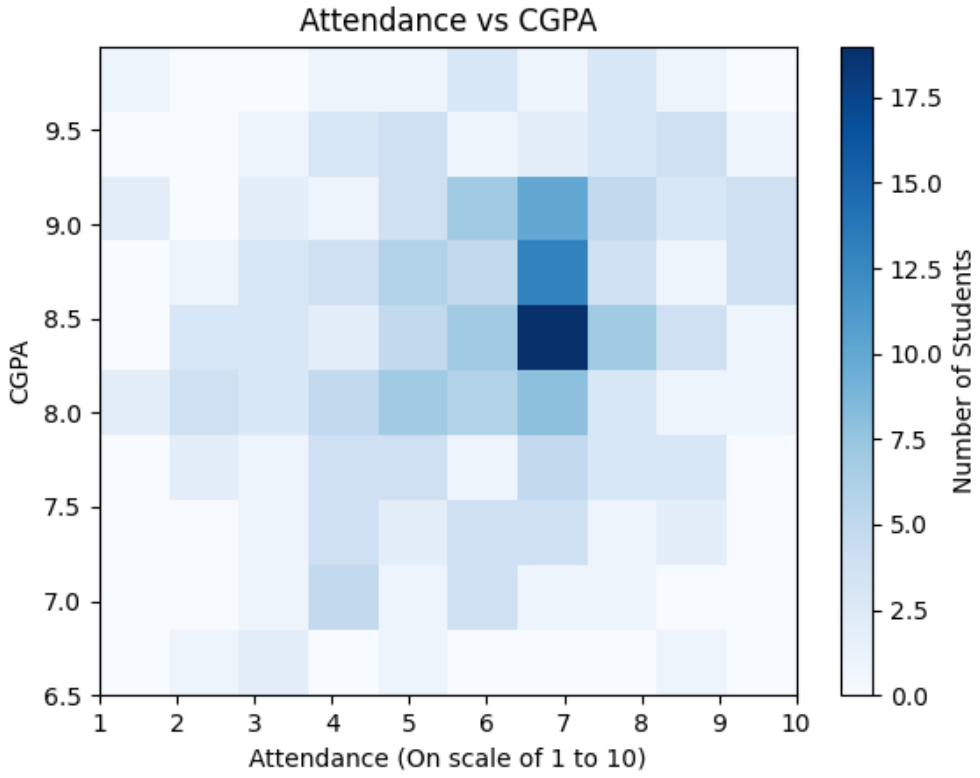


Figure 10: 2D Histogram of Attendance vs CGPA

From our calculations we get the correlation coefficient as following:

$$r_{xy} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}} = 0.20$$

Since a correlation coefficient of 0.2 is considered to be negligible correlation we infer that there is no relationship between attendance and cgpa. Hence, we can say that more attendance does not imply better academic performance and vice-versa.

4.3 Correlation between no. of courses with mandatory attendance vs mean no. of classes attended by the students:

The data for attendance is available to us as the no. of classes a student attends in a week (on a scale of 1 to 10). We also have the number of courses students have which have mandatory attendance set by the instructor.

Now we want to see if there exists some relationship between the given two variables. We use Pearson correlation coefficient to find how the two dataset correlates.

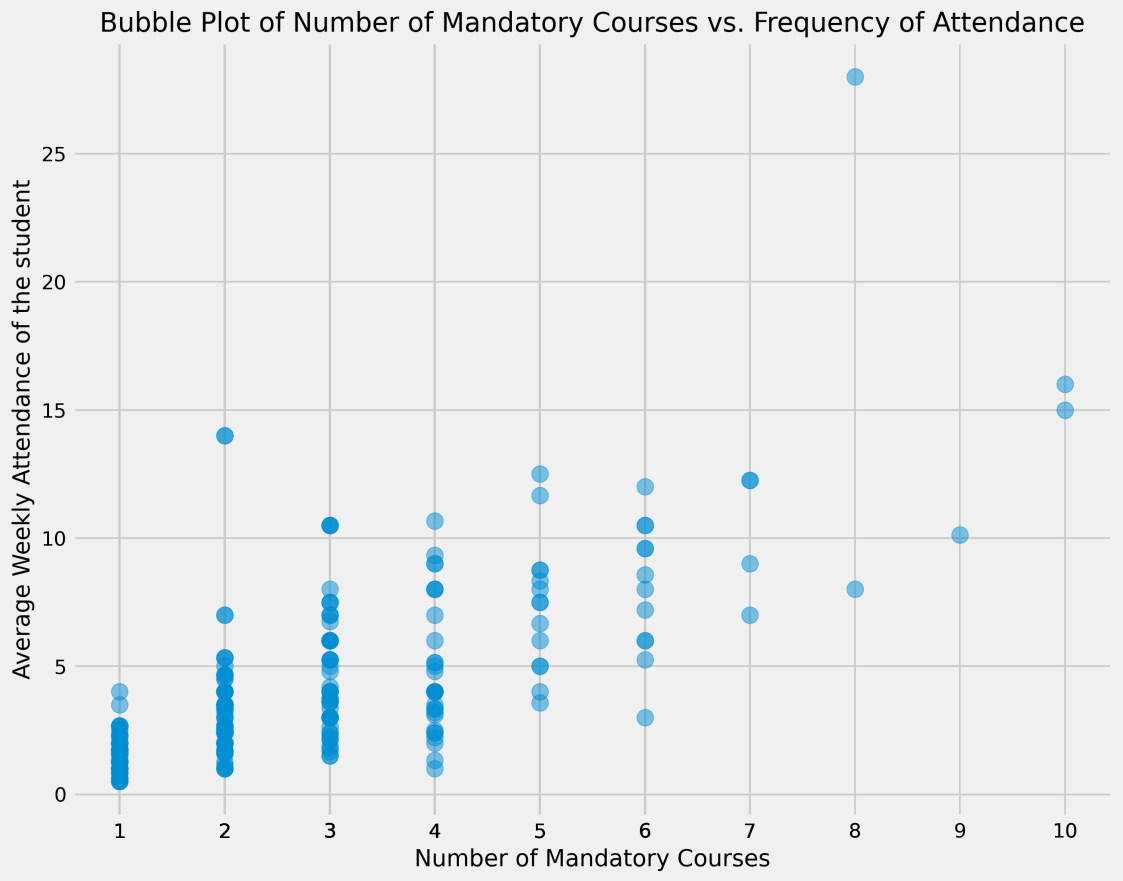


Figure 11: Scatter Plot of No. of Mandatory Courses vs Attendance

From our calculations we get the correlation coefficient as following:

$$r_{xy} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}} = 0.71$$

With a correlation coefficient as high as 0.71, it's clear that there's a strong link between the number of mandatory courses and student attendance.

5 Factors influencing Attendance due to courses

This is shown by calculating confidence intervals for different variations.

5.1 Confidence Interval for Mean (σ is unknown)

We aim to construct a confidence interval to estimate the population mean of students attending two consecutive classes with a gap in between.

The formula for calculating the confidence interval for the population mean (μ) when the population standard deviation (σ) is unknown is given by:

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

Where:

- $\bar{x}$ is the sample mean,
- $t_{\alpha/2,n-1}$ is the critical value from the t-distribution with $n-1$ degrees of freedom corresponding to the desired level of confidence (α is the significance level),
- S is the sample standard deviation,
- n is the sample size.

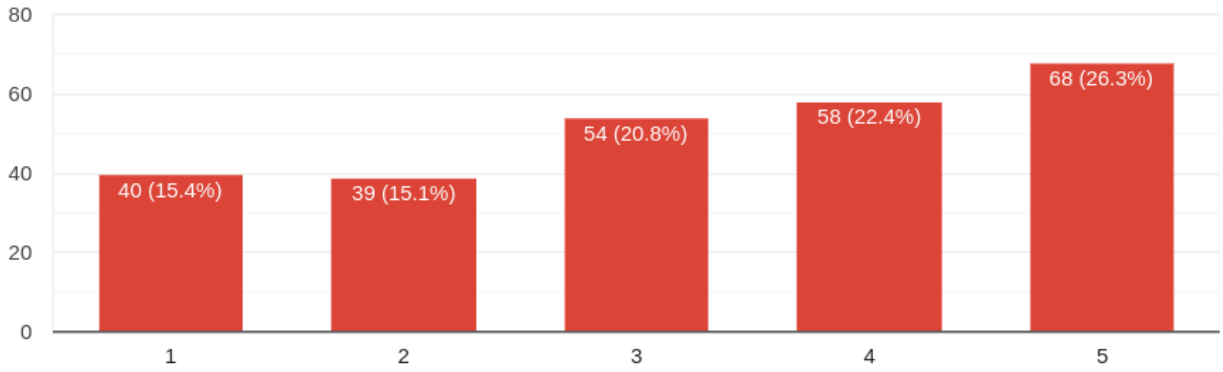


Figure 12: Rating percentages of students on their likeliness of attending two classes with a gap in between

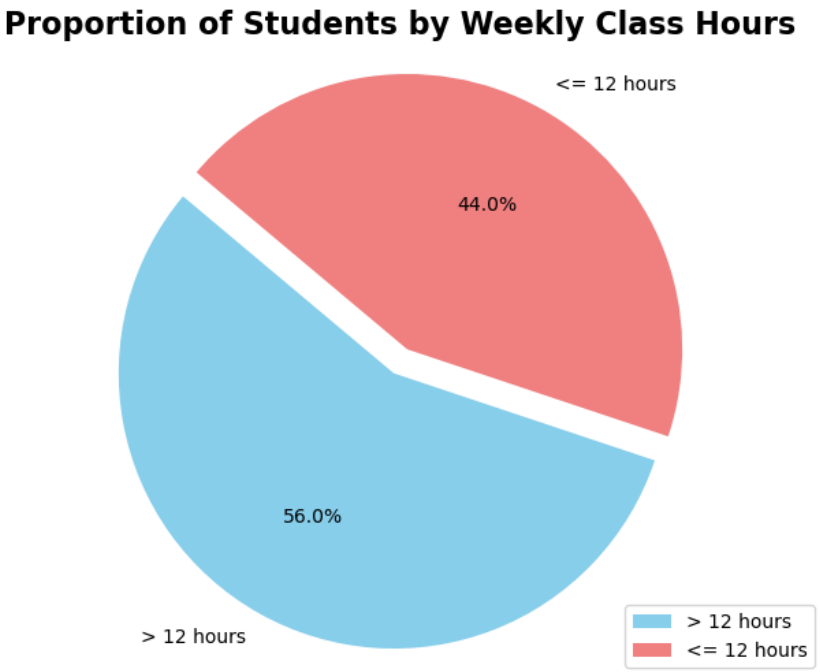
We have,
Sample mean ($\bar{x}$) = 3.29, S = 1.40, n = 259, and $t_{0.025,258} \approx 1.969$.

Substituting the given values into the formula, we find the confidence interval for the population mean of students attending two consecutive classes with a gap in between:

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

Hence, we are 95% confident that the likeliness of a student attending classes that has time gap in between lies in the interval **3.119** and **3.461** on a scale of **1** to **5**, in the increasing order of willingness of attending that class.

5.2 Confidence Interval for Proportion



a

Figure 13: Proportions of students having classes lesser and greater than 12 hours

To calculate the confidence interval (CI) for the proportion of students having classes less than 12 hours $\hat{p}$ with $(1-\alpha)100\%$ confidence, we use the formula:

$$\hat{p} \pm z_{\alpha/2} \left(\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

Where:

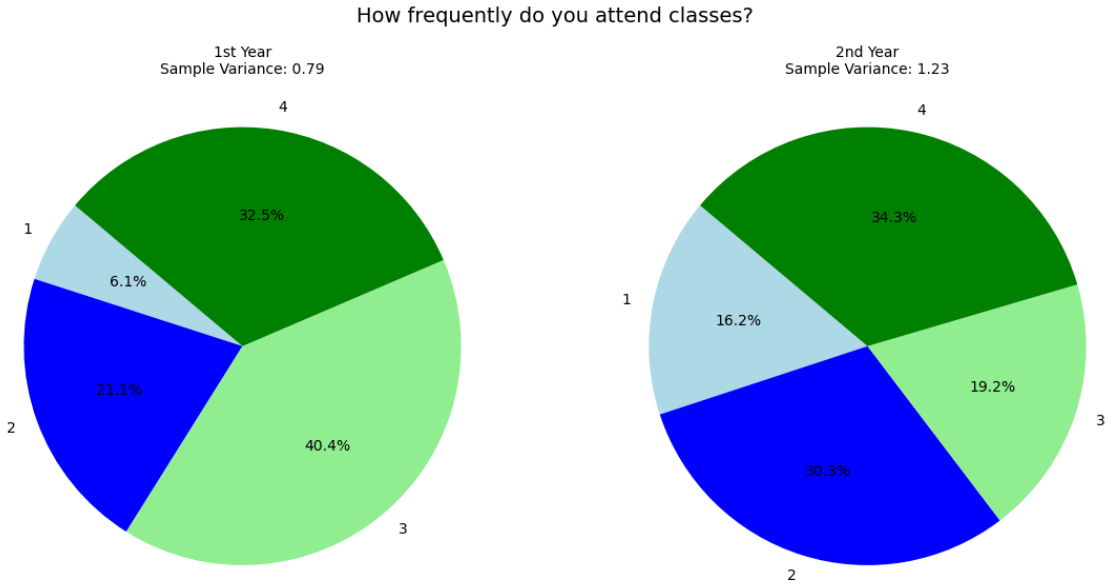
- $\hat{p}$ is the sample proportion of students having classes less than or equal to 12 hours.
- $z_{\alpha/2}$ is the Z-score corresponding to our desired confidence level.
- n is the sample size (here total number of students).

Given that our sample has $n = 259$ and $\hat{p} = 0.44$, where $np \geq 5$ and $n(1-p) \geq 5$ for the proportion, we can proceed with the analysis.
Let's take $\alpha = 0.05$, indicating a confidence level of 95%.
On calculating, we obtain the confidence interval as

$$(L, U) = (0.3795, 0.5004)$$

Thus, we are 95% confident that between **37.9%** and **50%** of students have classes less than or equal to 12 hours in a week.

5.3 Confidence Interval for Ratio of two population variances



a

Figure 14: Proportions of students from 1st and 2nd year on the frequency of classes they attend.

To calculate the confidence interval (CI) for the variance of students from 1st and 2nd year on the frequency of classes they attend i.e. $(\frac{\sigma_1^2}{\sigma_2^2})$ with $(1-\alpha)100\%$ confidence, we use the formula:

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

Where:

- S_1^2 and S_2^2 are the sample variances of students from the 1st and 2nd year on their frequency of attending classes, respectively.
- n_1 and n_2 are the sample sizes of students from the 1st and 2nd year, respectively.

We have,
$n_1 = 114$, $n_2 = 98$, $S_1^2 = 0.79$, and $S_2^2 = 1.23$ and $\alpha = 0.05$ for calculating a confidence interval with 95% confidence level.

After obtaining the values of $F_{0.025,\,113,\,97}$ and $F_{0.025,\,97,\,113}$ and substituting the given values in the above formula, we obtain the confidence interval as

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

So, we are 95% confident that ratio of variances of frequency of 1st and 2nd year students attending the classes lies in the interval $(0.646, 1.039)$.

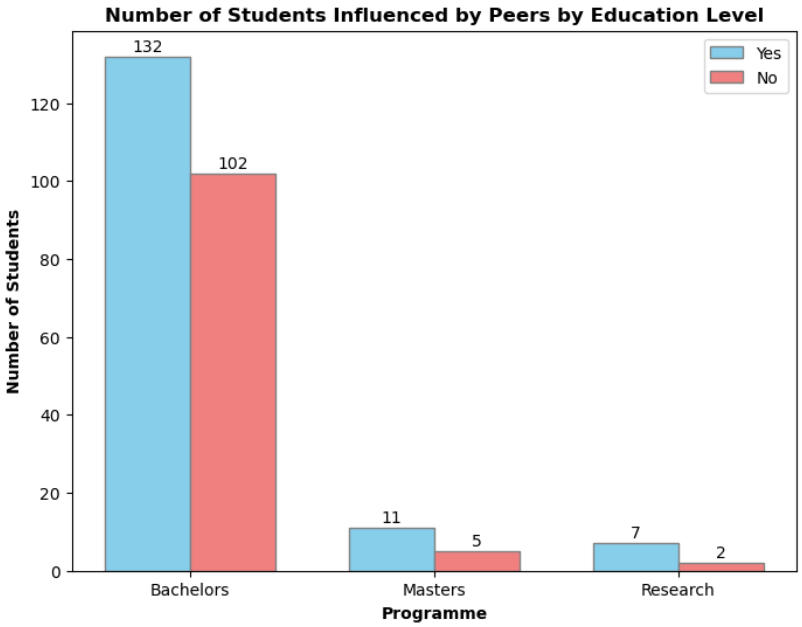
5.4 Confidence Interval for Difference in Proportions

To compute the confidence interval for the disparity in proportions concerning attendance dependency on peers between undergraduate (UG) and postgraduate (PG) students, we use the formula:

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

Where:

- $\hat{p}_1$ is the sample proportion of UG students with attendance depending on peers.
- $\hat{p}_2$ is the sample proportion of PG students with attendance depending on peers.
- n_1 is the sample size of UG students.
- n_2 is the sample size of PG students.
- $z_{\alpha/2}$) is the Z-score corresponding to your desired confidence level.

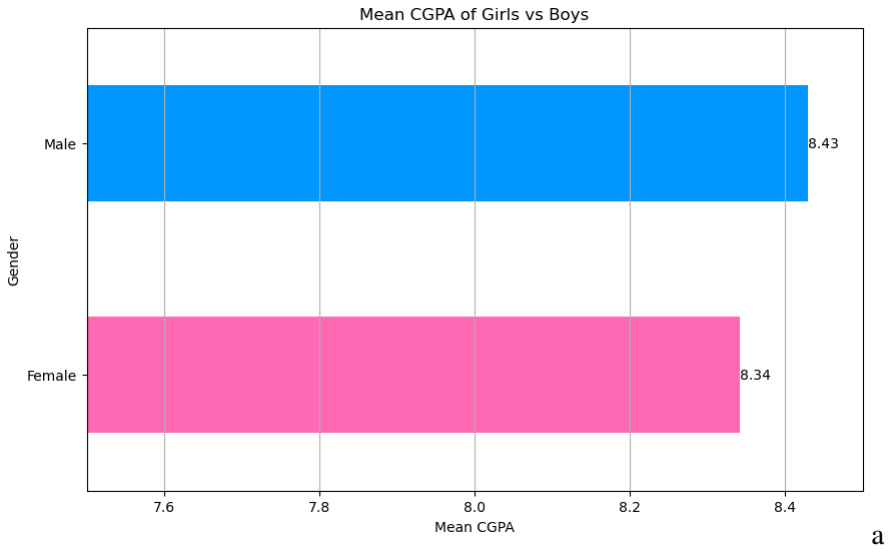


Considering the values of $\hat{p}_1 = 0.56$, $\hat{p}_2 = 0.69$, $n_1 = 234$, and $n_2 = 16$ from the data, we calculate the CI for the difference in proportions. Given that the sample size $n_1 = 234$, $n_2 = 16$, we can verify that $n_1\hat{p}_1 \geq 5$, $n_1(1-\hat{p}_1) \geq 5$, $n_2\hat{p}_2 \geq 5$, $n_2(1-\hat{p}_2) \geq 5$.
Let's take $\alpha = 0.01$, indicating a confidence level of 99%˙
On calculating, we obtain the confidence interval as

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

5.5 Confidence Interval for Difference in Mean

We aim to calculate a confidence interval to estimate the difference between the mean CGPA of female and male students.



a

Figure 15: Horizontal Bar Graph showing mean CGPA of female and male students.

The formula for calculating the CI for the difference in population means (μ_1) and (μ_2) is given by:

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

Where:

- $r = \dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$
- $\bar{x}_1$ and $\bar{x}_2$ are the sample means of the female and male CGPA, respectively.
- s_1 and s_2 are the sample standard deviations of the female and male CGPA, respectively.
- n_1 and n_2 are the sample sizes of the female and male students, respectively.
- $t_{\alpha/2,r}$ is the critical value from the t-distribution table corresponding to the desired confidence level and degrees of freedom (which is generally the smaller of $n_1 - 1$ and $n_2 - 1$).

We have,
Sample means ($\bar{x}_1$) = 8.34 and ($\bar{x}_2$) = 8.43, s_1 = 0.66, s_2 = 0.76, n_1 = 71, n_2 = 188, r = 144 and $t_{0.005,144} \approx 2.6109$.
Considering $\alpha = 0.01$, indicating a confidence level of 99%.
On calculating from the formula, the CI for the difference between the mean CGPA of female and male students comes out to be:

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

This outcome indicates a distinction in the mean CGPAs between female and male students, with females slightly trailing behind. The wide width of the interval suggests that it lacks precision.

6 Wake-up time analysis

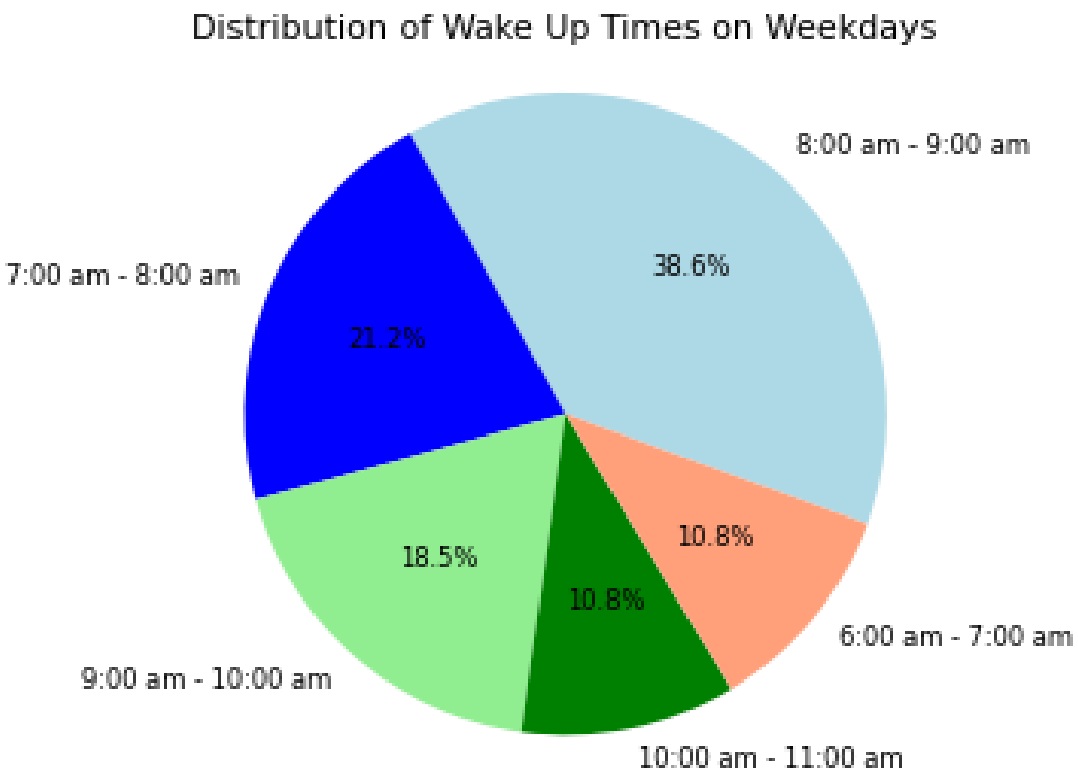
6.1 Confidence Interval for Population Variance

To compute the confidence interval for estimation of usual wake-up times of students in college on weekdays i.e. (σ^2) with $(1-\alpha)100\%$ confidence using the formula:

$$\left(\frac{(n-1)\cdot S^2}{\chi^2_{\alpha/2,\,n-1}}, \frac{(n-1)\cdot S^2}{\chi^2_{1-\alpha/2,\,n-1}} \right)$$

Where:

- S^2 is the sample variance,
- n is the sample size,



a

Figure 16: Proportions of wake-up times of students on weekdays.

We have,
$S^2 = 4560.15$ (min^2), $n = 259$ from the data and $\alpha = 0.05$ to calculate 95% confidence interval.

From $n = 259 \rightarrow df = n - 1 = 258$ Substituting the values in above equation, we obtain the confidence interval of population variance as

$$(L, U) = (3865.29, 5462.01)$$

Now, to calculate the confidence interval for population standard deviation, we can obtain it by using:

$$\left(\sqrt{\frac{(n-1) \cdot S^2}{\chi^2_{\alpha/2,\, n-1}}}, \sqrt{\frac{(n-1) \cdot S^2}{\chi^2_{1-\alpha/2,\, n-1}}} \right)$$

i.e. for the 95% confidence interval of standard deviation:

$$(L, U) = (\sqrt{3865.29}, \sqrt{5462.01})$$

$$(L, U) = (62.17, 73.91)$$

6.2 Point Estimate of Population Mean using Method of Moments (MME)

To estimate the population mean (μ) of wake up times of students of our college using the method of moments. The point estimate of μ is obtained as follows:

$$\hat{\mu} = \bar{X}$$

Where:

- $\hat{\mu}$ is the point estimate of the population mean,
- $\bar{X}$ is the sample mean.

Now we have the data as from Figure 16:

Xi	Time Interval	Percentage of Population
0	6:00 am - 7:00 am	10.8%
1	7:00 am - 8:00 am	21.2%
2	8:00 am - 9:00 am	38.6%
3	9:00 am - 10:00 am	18.5%
4	10:00 am - 11:00 am	10.8%

Table 1: Percentage of population waking up between different time intervals

Substituting and obtaining the point estimate of mean from this data, we get:

$$\bar{X} = \frac{(0 \times 10.8) + (1 \times 21.2) + (2 \times 38.6) + (3 \times 18.5) + (4 \times 10.8)}{100}$$

$$\bar{X} = 1.971$$

$$\hat{\mu} = 1.971$$

On converting this to actual time format, we have the point estimated wake up time of students using **MME** as **7:58 am**.

6.3 Hypothesis Testing for Genderwise Mean Wake-Up Times

From the sample data, we test the hypothesis that for the entire IIT-H student population the average wake-up time of the male and female students is the same or not. We first compute the sample standard deviation of the wake time of both male and female students.
We Get,

$$S^2_{female} = 1.08 \tag{1}$$

$$S^2_{male} = 1.14 \tag{2}$$

$$\frac{S^2_{female}}{S^2_{male}} = \frac{1.08}{1.14} = 0.95 \tag{3}$$

Since $0.25 \leq \frac{S^2_{female}}{S^2_{male}} \leq 4$, we can assume that the population variances are equal.

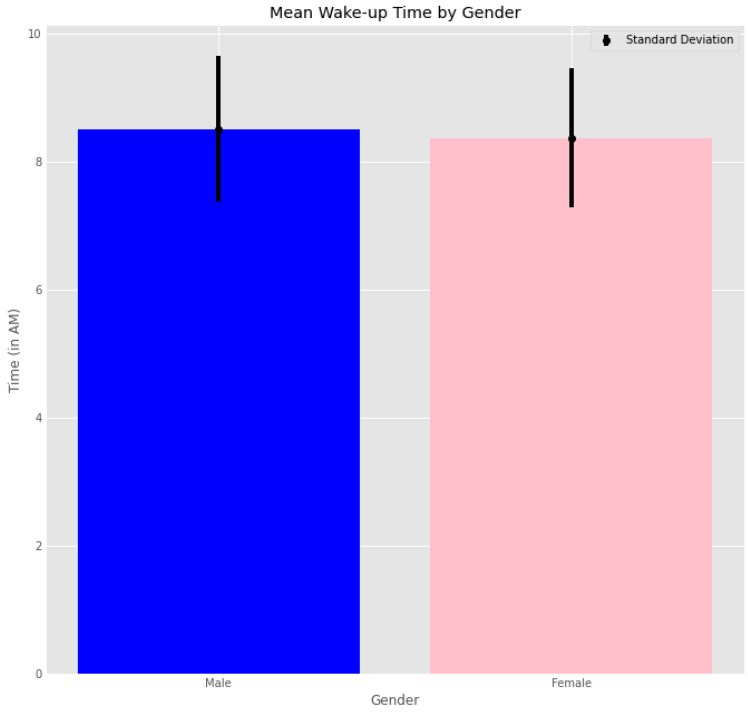


Figure 17: Bar Graph of Wake-up time of IITH Students:

Hence,

$$\sigma^2_{female} \approx \sigma^2_{male} \tag{4}$$

Now writing the hypothesis,

Null Hypothesis (H_0): $\mu_{female} - \mu_{male} = 0$
Alternate Hypothesis (H_1): $\mu_{female} - \mu_{male} \neq 0$

On computation the sample mean wake time of both female and male students are:

$$\bar{X}_{female} = 8.37, \bar{X}_{male} = 8.51 \tag{5}$$

Given that the population variances are approximately equal and the sample size $n = 259$. Assuming that the sleep patterns of male and female students are independent, let's perform the **Two-Tailed Hypothesis Test**, for the equal variances case.

- Set the significance level α: Let's assume $\alpha = 0.05$.
- The sample size of female students are $n_{female} = n_1 = 71$, and the sample size of male students are $n_{male} = n_2 = 188$.
- Calculating S_p:

$$S_p = \sqrt{\frac{(n_1 - 1)S^2_{female} + (n_2 - 1)S^2_{male}}{n_1 + n_2 - 2}} = \sqrt{\frac{70 \times 1.08 + 187 \times 1.14}{71 + 188 - 2}} = 1.06 \tag{6}$$

  Where S_p is the combined sample standard deviation.

- Calculating the test statistic:

$$t^* = \frac{(\bar{X}_{female} - \bar{X}_{male}) - 0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{-0.14}{1.06 \times \sqrt{\frac{1}{71} + \frac{1}{188}}} = -0.95 \tag{7}$$

- Find the degrees of freedom $df = n_1 + n_2 - 2 = 71 + 188 - 2 = 257$.
- Finding the critical value $t_{\frac{\alpha}{2},df} = t_{0.025,257}$ from the t-distribution table:

$$t_{\frac{\alpha}{2},df} = t_{0.025,257} = 1.965 \tag{8}$$

- Compare the test statistic with the critical value:
  - If $|t^*| \geq t_{\frac{\alpha}{2},df}$, we reject the null hypothesis.
  - If $|t^*| < t_{\frac{\alpha}{2},df}$, we fail to reject the null hypothesis.

Given that the calculated test statistic t^* is calculated using Equation 25, let's compare it with the critical value from the t-distribution table. Now,

$$t_{\frac{\alpha}{2},df} = 1.965 \tag{9}$$

and,

$$|t^*| = 0.95 \tag{10}$$

Hence,

$$|t^*| < t_{\frac{\alpha}{2},df} \tag{11}$$

Therefore, we cannot reject the null hypothesis. There is a significant risk of rejecting the hypothesis that the mean wake-up time of male and female students of IIT-H is equal.

7 Hypothesis Testings

7.1 Hypothesis Testing for Dependent Samples

We conduct a hypothesis test to determine whether the attendance app has led to an increase in attendance among students. We use the test for **paired data** and set up the research hypothesis:

Null Hypothesis (H_0): $\mu_D \leq 0$
Alternate Hypothesis (H_1): There is an increase in the mean attendance after the launch of the attendance app. $\mu_D > 0$

Given the sample data, We obtain the sample difference of mean to be $\bar{D} = 0.30$, the Sample variance of differences in data $S_D = 1.02$ and we have $n = 259$ where n is the number of paired observations, let's perform the hypothesis test.

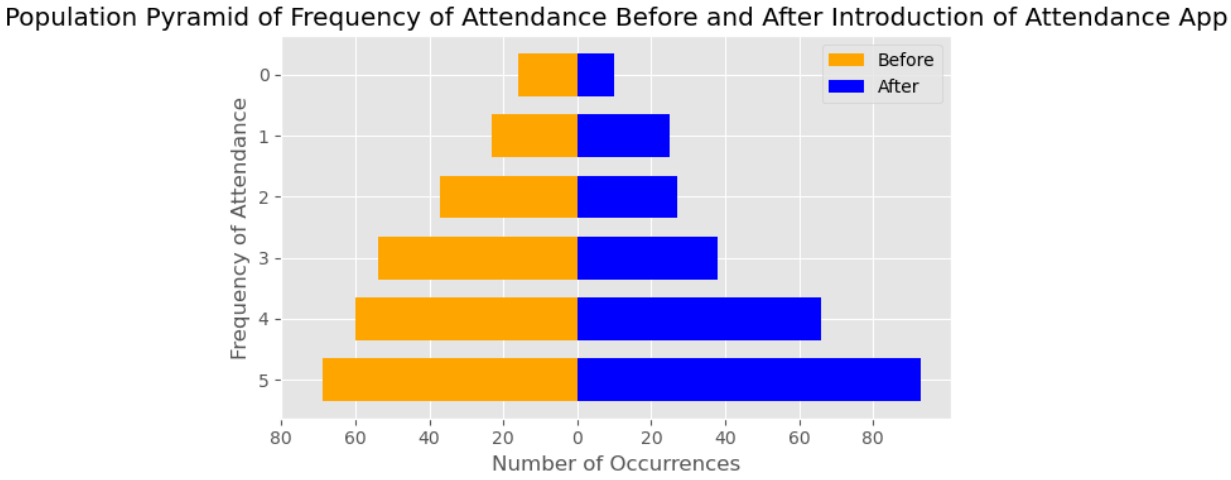


Figure 18: Bar graph showing the difference in attendance before and after launch of app.

- Set the significance level α. Let's assume $\alpha = 0.05$.
- Calculate the test statistic:

$$t^* = \frac{\bar{D} - D_0}{\frac{S_D}{\sqrt{n}}} = \frac{0.3 - 0}{\frac{1.02}{\sqrt{259}}} = 4.742 \tag{12}$$

- Find the degrees of freedom ($df = n - 1 = 259 - 1 = 258$) for the t-distribution.
- Determine the critical value from t-distribution table:

$$t_{\alpha,df} = t_{0.05,258} = 1.650 \tag{13}$$

- Compare the test statistic with the critical value(s):
  - If $t^* \geq t_{\alpha,df}$, we reject the null hypothesis.
  - If $t^* < t_{\alpha,df}$, we fail to reject the null hypothesis.

Now, we have the calculated test statistic t that is calculated using Equation 13, let's compare it with the critical value from the t-distribution table.

$$t_{\alpha,df} = 1.650 \tag{14}$$

and,

$$t^* = 4.742 \tag{15}$$

Hence,

$$t^* > t_{\alpha,df} \tag{16}$$

As we have the result $t^* > t_{\alpha,df}$, we can reject the null hypothesis H_0. Therefore, we conclude that there is sufficient evidence to suggest that the population mean difference μ_D is greater than 0, indicating an increase in attendance after the launch of the attendance app.

P-Value Calculation:
To calculate the p-value, we use the cumulative distribution function (CDF) of the t-distribution with degrees of freedom $df = 258$ and the test statistic $t^* = 4.742$. Since it's a right-tailed test, we find the probability of obtaining a test statistic greater than t^*.

The p-value can be calculated as follows:

$$\text{p-value} = P(T > 4.742)$$

Using t-distribution table, we get:

$$p - value \approx 0$$

Since the p-value is very less than the significance level $\alpha = 0.05$, it favors the alternate hypothesis and thus, we reject the null hypothesis. Therefore, we conclude that there is sufficient evidence to suggest that the population mean difference μ_D is greater than 0, indicating an increase in attendance after the launch of the attendance app.

7.2 Hypothesis Testing For Variation in Number of Courses

From the sample data, we test the hypothesis that for the Bachelors student population, an average 1st and 2nd year (Novices) student has very less variation in the number of courses taken compared to the 3rd and 4th year (Senior) students. We can set up the research hypothesis as:

Null Hypothesis (H_0): $\sigma_{novices} \geq \sigma_{seniors}$
Alternate Hypothesis (H_1): $\sigma_{novices} < \sigma_{seniors}$

Given that the calculated sample variances $S^2_{novices} = 0.20$, $S^2_{seniors} = 1.10$ and the sample size $n = 234$, let's perform the **One-Tailed Hypothesis Test**.

- Set the significance level α: Let's assume $\alpha = 0.05$.
- Calculating the test statistic:

$$F^* = \frac{S^2_{novices}}{S^2_{seniors}} = \frac{0.20}{1.10} = 0.18 \tag{17}$$

  Where S is the sample standard deviation.

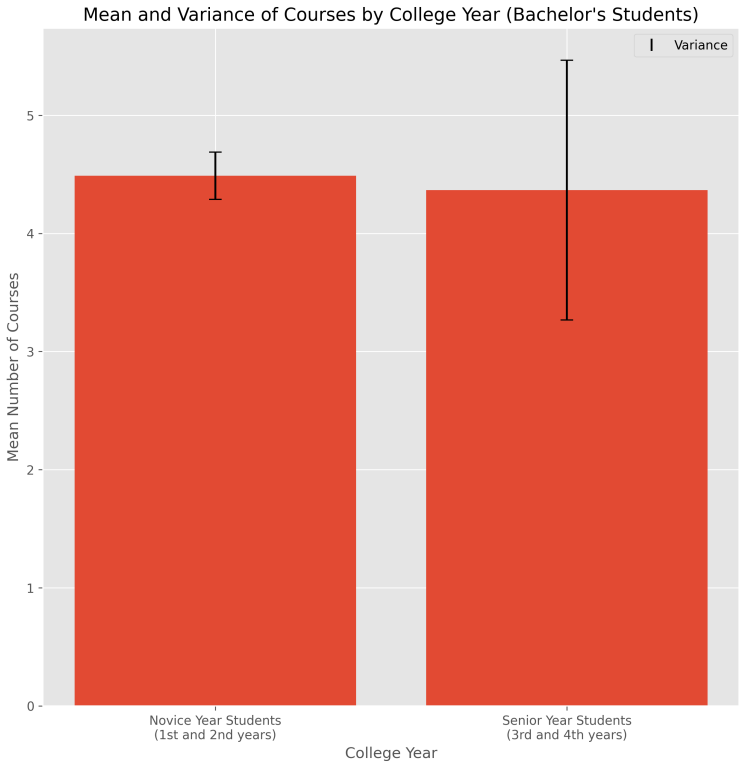


Figure 19: Bar Graph of College year and mean number of courses

- Find the degrees of freedom $df_1 = n - 1 = 234 - 1 = 233$, $df_2 = n - 1 = 234 - 1 = 233$.
- Finding the critical value $F_{1-\alpha,df_1,df_2} = F_{0.05,233,233}$ from the F-distribution table:

$$F_{1-\alpha,df_1,df_2} = F_{0.95,233,233} = 0.81 \tag{18}$$

- Compare the test statistic with the critical value:
  - If $F^* \leq F_{1-\alpha,df_1,df_2}$, we reject the null hypothesis.
  - If $F^* > F_{1-\alpha,df_1,df_2}$, we fail to reject the null hypothesis.

Given that the calculated test statistic F^* is calculated using Equation 25, let's compare it with the critical value from the F-distribution table.
Now,

$$F_{1-\alpha,df_1,df_2} = 0.81 \tag{19}$$

and,

$$F^* = 0.18 \tag{20}$$

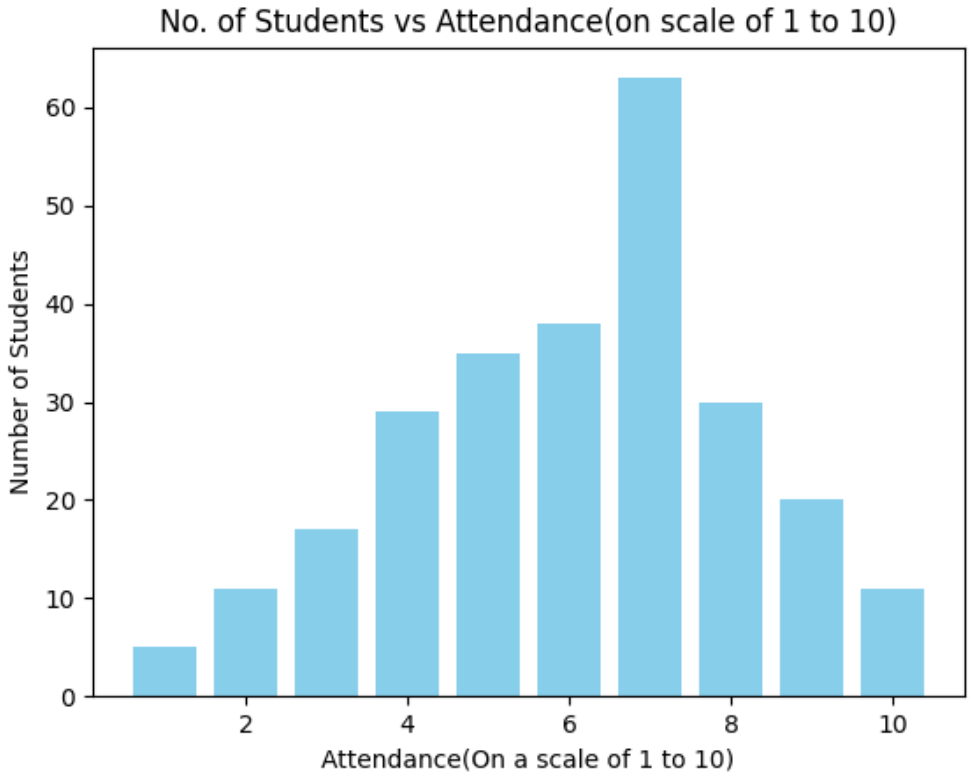
Hence,

$$F^* \leq F_{1-\alpha,df_1,df_2} \tag{21}$$

Therefore, we reject the null hypothesis and conclude the 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. This is also evident from the fact that senior year students opt for more or less electives whereas the novice year students have a fixed curriculum of courses.

7.3 Hypothesis Testing For Average Student Attendance

From the sample data, we test the hypothesis that for the whole student population, an average student attends at least 50% of the classes.



We can use the **one-mean t-test** and set up the research hypothesis:

Null Hypothesis (H_0)**:** $\mu \leq 0.5$
Alternate Hypothesis (H_1)**:** $\mu > 0.5$

where μ is the average student attendance

Given that the calculated sample mean $\bar{X} = 0.605$ and the sample size $n = 260$, let's perform the hypothesis test.

- Set the significance level α. Let's assume $\alpha = 0.05$.
- Calculate the test statistic:

$$t^* = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{0.605 - 0.5}{\frac{0.21}{\sqrt{260}}} = 8.06 \tag{22}$$

  Where $S = 0.21$ is the sample standard deviation.

- Find the degrees of freedom $df = n - 1 = 260 - 1 = 259$.

- Finding the critical value $t_{\alpha,df}$ from the t-distribution table:

$$t_{\alpha,df} = t_{0.05,259} = 1.65 \tag{23}$$

- Compare the test statistic with the critical value:
  - Here $t^* = 8.06 \geq t_{\alpha,df} = 1.65$, so we reject the null hypothesis.

Therefore, we reject the null hypothesis and conclude that an average student of IIT-H attends more than 50% of the classes.

7.4 Hypothesis Testing For Influence of Exam Scores

Since we have data about how a students attendance is affected after his/her performance in exams, we can build a hypothesis that in response to receiving low scores on exams, a significant portion of students will exhibit a change in behavior by increasing their attendance in classes.

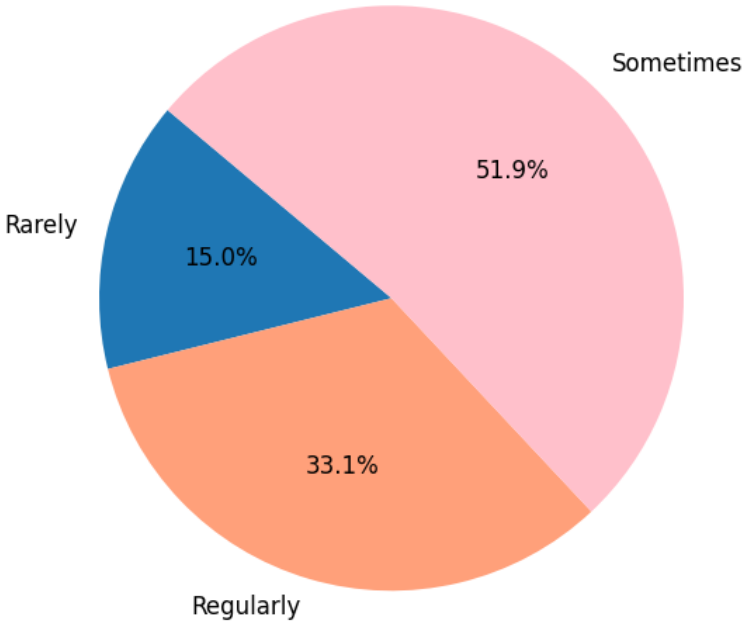


Figure 20: Percentage of Students who attend more classes after getting a low score

We can use the **test for one proportion** and set up the research hypothesis:

Null Hypothesis (H_0)**:** $p \leq 0.33$
Alternate Hypothesis (H_1)**:** $p > 0.33$

where p is the proportion of students who gets influenced after getting low score in exams.// Given that the sample size $n = 259$, $np_0 = 86 \geq 5$ and $n(1 - p_0) = 174 \geq 5$ let's perform the hypothesis test.

- Set the significance level α. Let's assume $\alpha = 0.05$.
- Calculate the test statistic:

$$z^* = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.394 - 0.33}{\sqrt{\frac{0.33(1-0.33)}{260}}} = 2.19 \tag{24}$$

- Finding the critical value z_α from the z-distribution table:

$$z_\alpha = z_{0.05} = 1.645 \tag{25}$$

- Compare the test statistic with the critical value:
  - Here $z^* = 2.19 \geq t_{\alpha,df} = 1.645$, so we reject the null hypothesis.

Thus, we reject the null hypothesis and conclude that more than 33% of the IIT-H student population starts attending classes more frequently after getting a low score in previous quiz/mid-sem.

8 Chi-squared Tests

8.1 Chi-squared Test for Association

Now, let's perform the chi-squared test to determine if there is an association between gender and the time of day preference for attending the classes.
Null Hypothesis (H_0): There is no association between gender and the time of day preference for attending the classes.
Alternative Hypothesis (H_a): There is an association between gender and the time of day preference for attending the classes.

Table 2: Contingency Table: Gender vs Time of Day Preference

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

- Using the observed frequencies and the formula $E_{ij} = \frac{(Row\ Total_i) \times (Column\ Total_j)}{Grand\ Total}$, we calculate the expected frequencies.
- After calculating the expected frequencies, we get:

$$E_{male,Morning} \approx 131.3$$

$$E_{male,Afternoon} \approx 108.0$$

$$E_{male,Evening} \approx 41.6$$

$$E_{female,Morning} \approx 48.6$$

$$E_{female,Afternoon} \approx 40.0$$

$$E_{female,Evening} \approx 15.4$$

- Using the formula $\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$, we calculate the chi-squared test statistic.
- After calculating the chi-squared test statistic, we get:

$$\chi^2 \approx 3.827$$

- With $df = (2-1) \times (3-1) = 2$, and at a significance level of 0.05, the critical value from the chi-squared distribution table is approximately 5.991.

Conclusion: Since the calculated chi-squared test statistic (3.827) is less than the critical value (5.991), we fail to reject the null hypothesis. Based on the data and the chi-squared test, we do not have sufficient evidence to conclude that there is an association between gender and the time of day preference for attending the classes. Therefore, we accept the null hypothesis.

8.2 Chi-Square Goodness of Fit Test

Null Hypothesis (H_0): Peers don't affect people going to class ($p_1 = p_2 = \frac{1}{2}$)
Alternative Hypothesis (H_a): Peers affect people going to class (p_1 and p_2 are not both equal to $\frac{1}{2}$)

Table 3: Observed Frequencies

Category	Frequency
Peer affecting	109
Peer not affecting	150

- The expected frequency for each category, given the null hypothesis, is:

$$E = n \times p$$

  where n is the total number of observations and p is the probability under H_0.

  - Expected frequency for "Peer affecting": $E_1 = 259 \times \frac{1}{2} = 129.5$
  - Expected frequency for "Peer not affecting": $E_2 = 259 \times \frac{1}{2} = 129.5$

- Now, we calculate the Chi-Square statistic:

$$\chi^2 = \frac{(109 - 129.5)^2}{129.5} + \frac{(150 - 129.5)^2}{129.5}$$

$$\chi^2 = 3.2476 + 3.2476$$

$$\chi^2 = 6.4952$$

- Since we have two categories, the degrees of freedom (df) is $k - 1 = 2 - 1 = 1$.
- Let $\alpha = 0.05$ for a 95% confidence level.
- Using a Chi-Square distribution table, we find the critical value corresponding to $\chi^2 = 6.4952$ and $df = 1$.
- Using the chi-square distribution table (or a statistical software), the critical value for df=1 and $\alpha = 0.05$ is 3.841.

Interpretation: With a critical value of 3.841, we reject the null hypothesis. Therefore, we have sufficient evidence to conclude that there is a significant difference in the proportion of people attending class depending on whether peers affect them or not. Peers do seem to have an influence on people going to class.

9 Conclusions:

- A strong correlation coefficient of approximately 0.8 exists between the number of courses registered and students' attendance, suggesting that students who register for more courses tend to attend classes more regularly.
- Conversely, the correlation between attendance and CGPA is minimal, with a coefficient around 0.2. This implies that attending more classes may not necessarily directly translate to higher academic performance, as measured by CGPA.
- The implementation of an attendance monitoring system has yielded positive results, as indicated by the observed increase in class attendance compared to the period before its introduction. This finding is supported by hypothesis testing.
- Notably, there is a significant discrepancy in the variability of course enrollment between students in their first and second years compared to those in their third and fourth years.
- More than a third of students report being influenced by their performance on exams to increase their class attendance, indicating the motivational impact of academic outcomes on attendance behavior.
- Social factors also play a significant role, with peer influence identified as a contributing factor to a student's attendance patterns in class.
- Additionally, the average wake-up time of 7:58 am aligns well with establishing a healthy routine conducive to attending classes throughout the day, potentially contributing to improved academic engagement and performance.