

Statistics on further studies of students in IITH

Kola Akshitha	Mouktika Cherukupalli
Avinash Malothu	Shreyas Wankhede
Pradeep Mundlik	Dhatri Reddy
Sumeeth Kumar	Sathvika Marri

May 1, 2023

1 Introduction

This project is based on further studies of students studying at IITH. We have used statistics to deduce few conclusions from the given data, assuming that the data is random sample population.

We used sampling, more specifically volunteer sampling for collection of data through mail from students at IITH but only a few of them volunteered to respond. The data was collected from 114 students, and it is diverse with data from different years of UG.

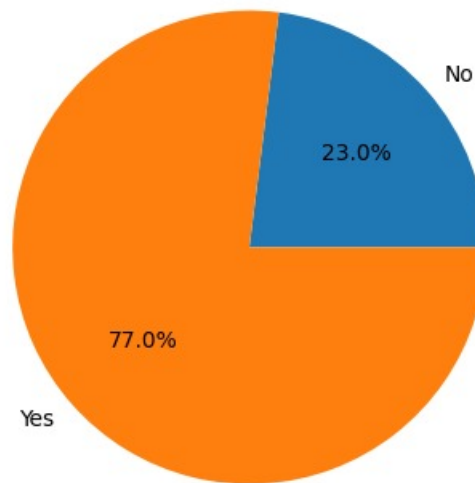
These are the questions asked in the survey

1. Which department?
2. Gender
3. Annual family income
4. CGPA
5. Interested in further studies (if yes)
 - After how many years of work experience?
 - Which degree?(Masters/ PhD/ MBA)
 - Location
 - In which department(Same as bachelors or different?)

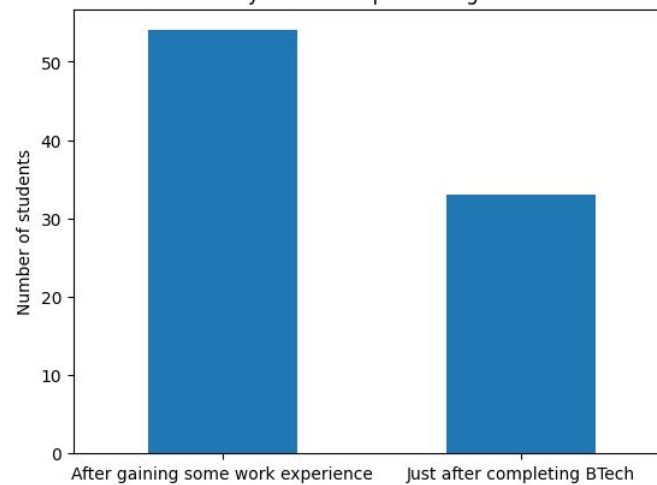
2 Data Visualisation

2.1 Frequency plots

Pie Chart of Students interested in pursuing higher studies

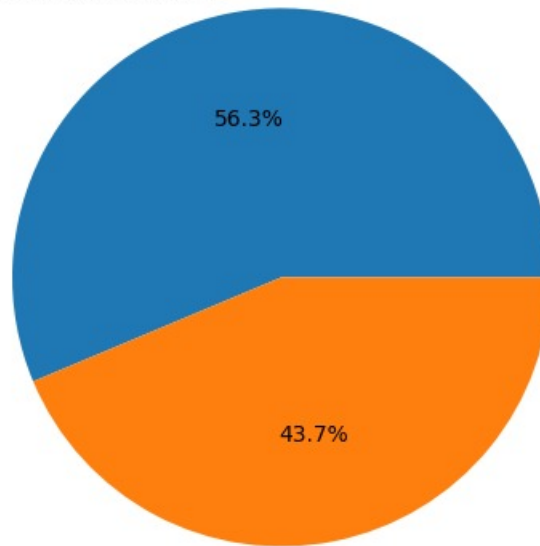


When do you want to pursue higher studies

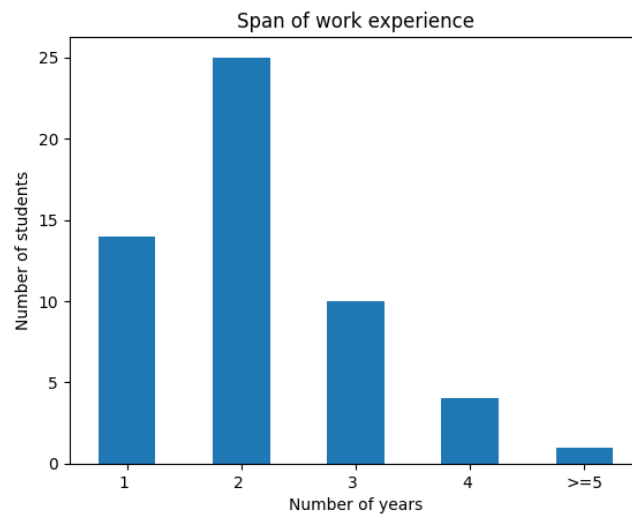


Which department are you willing to continue?

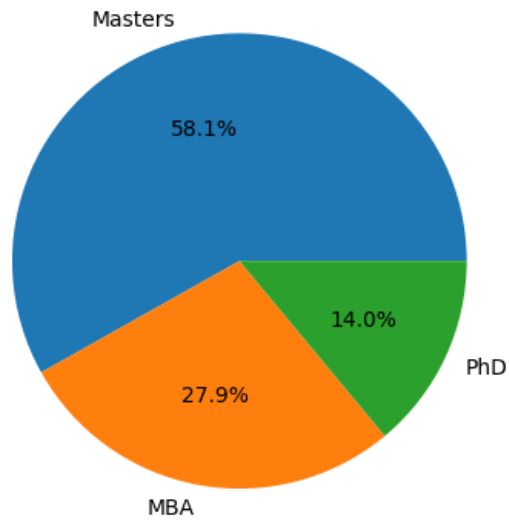
Different from Bachelors



Same as Bachelors



Preference of degree



Number of students persuing higher studies

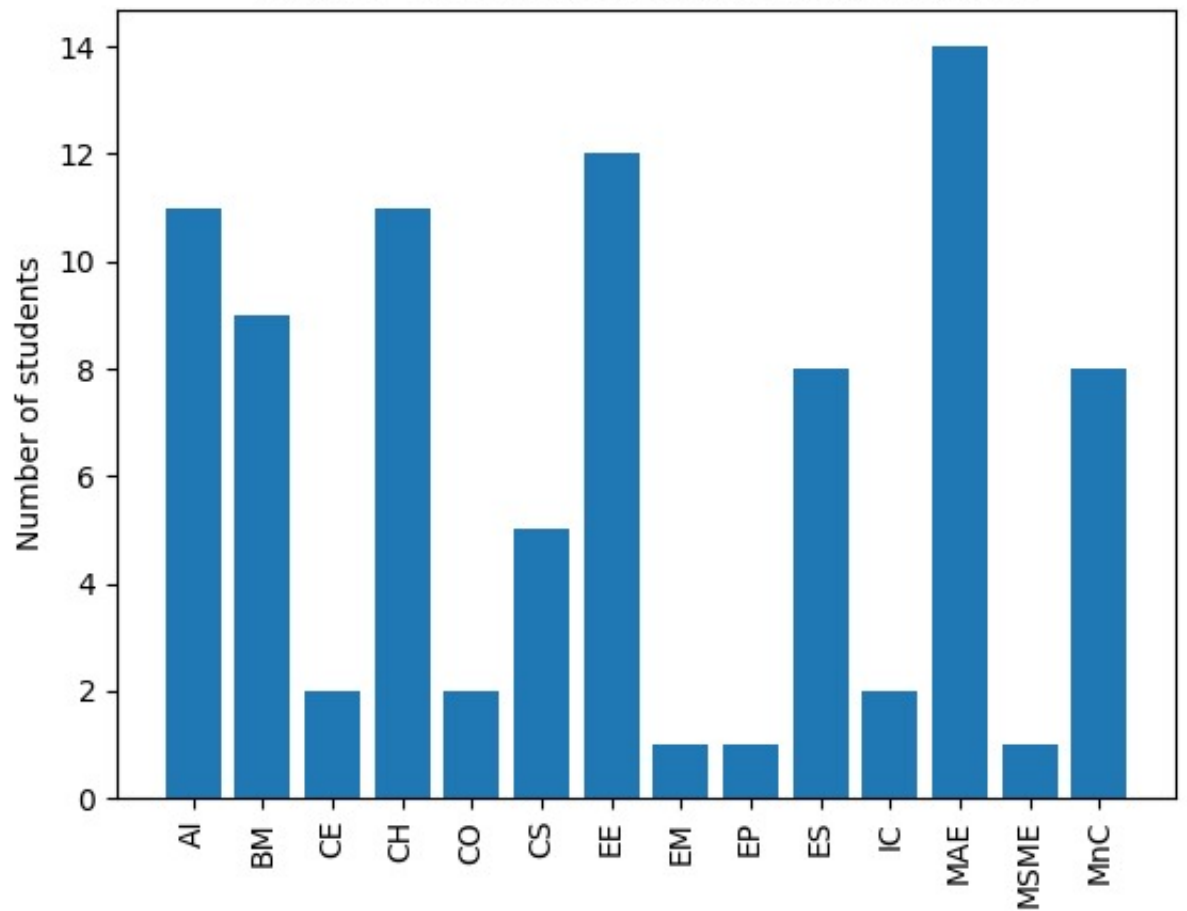
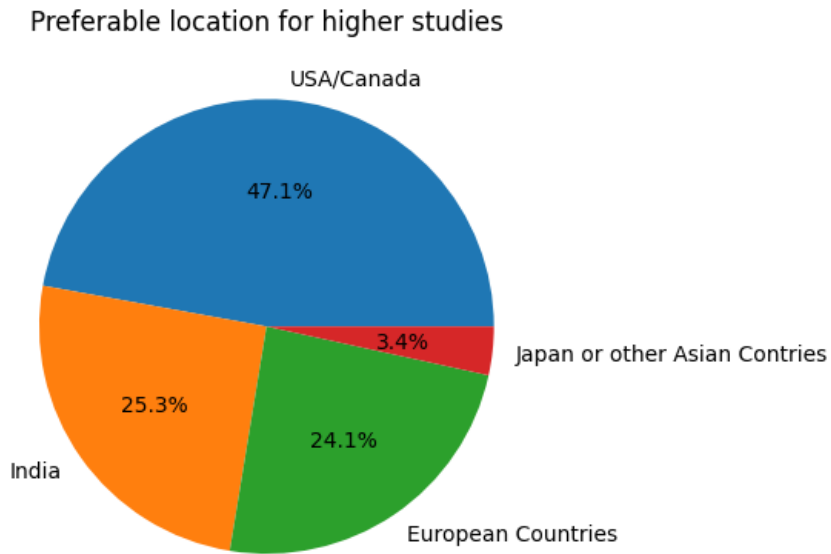
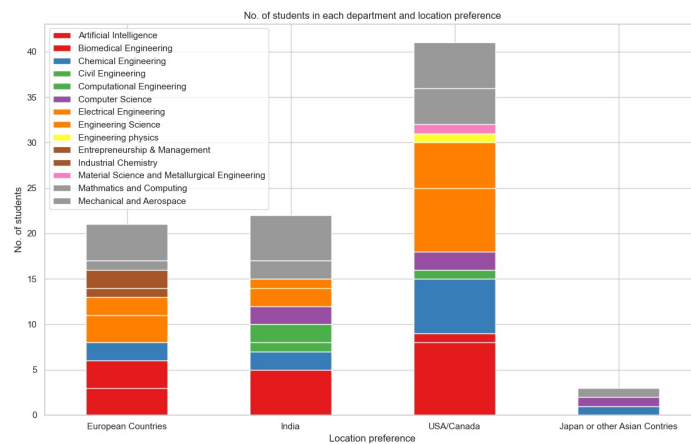
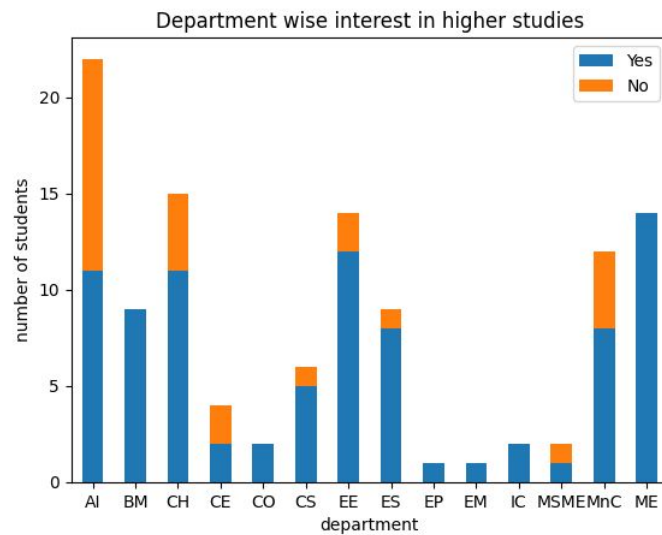
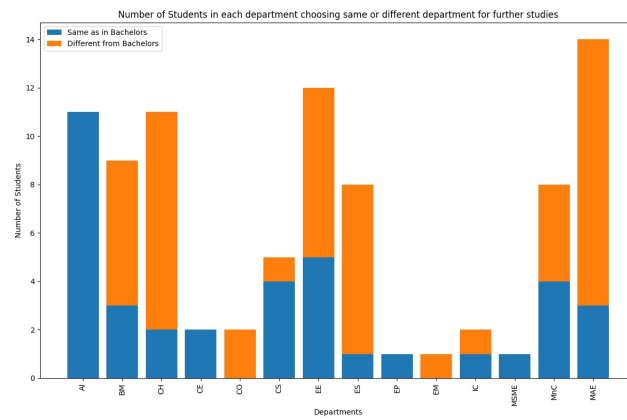
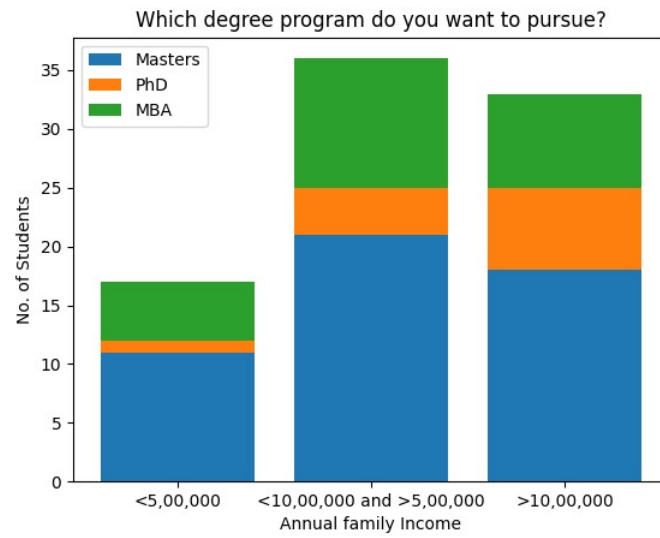


Figure 1: department wise students opting for higher studies



2.2 Cross plots





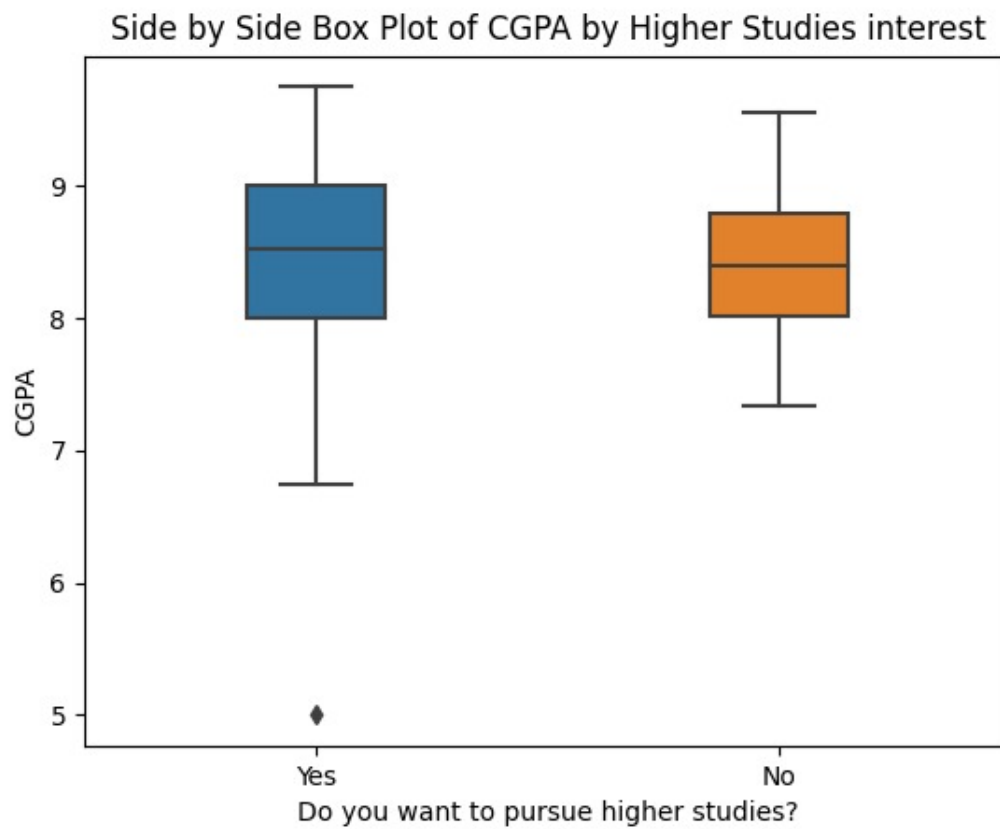
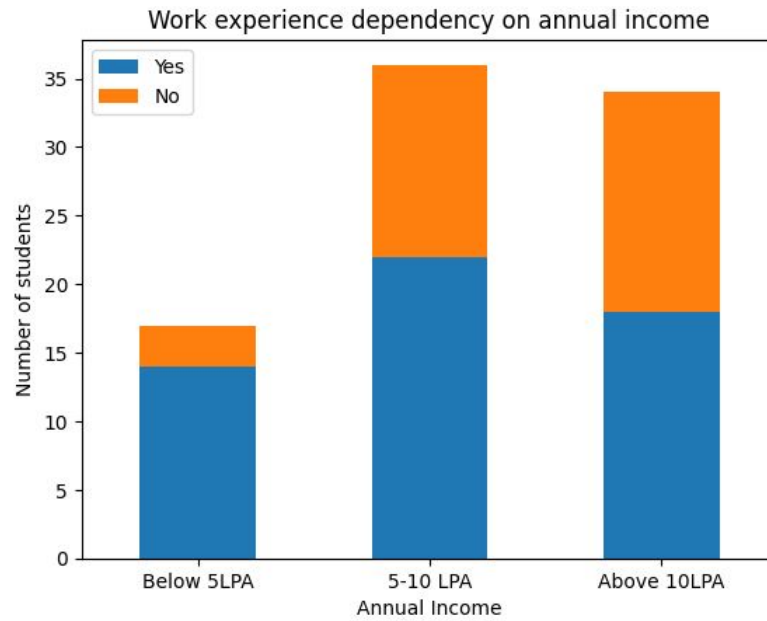


Table 1: Percentage Contingency table between Gender and interest in pursuing further studies

	Yes	No	Total (%)
Male	53.10	17.70	70.80
Female	23.89	5.31	29.20
Total (%)	76.99	23.01	100.00

Table 2: Percentage Contingency table between preferred program and preferred location

	MBA	Masters	PhD	Total (%)
European Countries	5.75	14.94	3.45	24.14
India	13.79	8.05	3.45	25.29
Japan or other Asian countries	1.15	1.15	1.15	3.45
USA/Canada	6.90	34.48	5.75	47.13
Total (%)	27.59	58.62	13.79	100.00

3 Analysing numerical data

How many years you will work before opting for Higher Education ?

Count	87(non-null)
Mean	1.333333
Median	1.000000
Mode	0.000000
Std	1.317738
Min	0.000000
25%	0.000000
50%	1.000000
75%	2.000000
Max	6.000000
95% Confidence Interval	(1.050951, 1.615715)
99% Confidence Interval	(0.961181, 1.705485)

4 Conclusions

With Around 113 students participating in the survey, 77% students are interested in pursuing higher studies and 23% students are not interested in pursuing higher studies.

Experience:

37.9 % students want to pursue higher studies with no experience, 62.1% students want to pursue higher studies with experience.

Degree preference:

58.1% want to pursue masters, 14.0% want to pursue PhD, 27.9% want to pursue MBA.

Location preference:

47.1% want to pursue in USA/Canada, 25.3% want to pursue in India, 24.1% want to pursue in European countries, 3.4% Japan or other Asian countries.

Field Interest:

72.56% students are interested in civil services, 27.44% students are not interested in civil services.

74.33% students are interested in software related jobs, 25.66% students are not interested in software related jobs.

5 Hypothesis Testing

5.1 Case 1: Comparing CGPA of students who are willing to pursue higher studies with students who don't want to pursue

We assume our null hypothesis to be that average CGPA of students willing to go for higher studies is greater than or equal to those who don't want to. Let $\alpha = 0.05$.

1. \bar{x}_1 = sample mean of CGPA of people willing to go for higher studies
2. \bar{x}_2 = sample mean of CGPA of people who don't want to go for higher studies
3. s_1^2 = sample standard deviation of CGPA of people willing for higher studies
4. s_2^2 = sample standard deviation of CGPA of people who don't want

For Hypothesis Testing we make the following statements:

$$H_0 = \mu_1 - \mu_2 \geq 0 \quad (1)$$

$$H_a = \mu_1 - \mu_2 < 0 \quad (2)$$

Information:

$$\bar{x}_1 = 8.4447 \quad (3)$$

$$\bar{x}_2 = 8.3919 \quad (4)$$

$$s_1^2 = 0.7694 \quad (5)$$

$$s_2^2 = 0.5672 \quad (6)$$

$$n_1 = 87 \quad (7)$$

$$n_2 = 26 \quad (8)$$

$$(9)$$

since $\frac{s_1^2}{s_2^2} < 4$, we can assume the population variances would be equal.

Degrees of freedom,

$$dof = n_1 + n_2 - 2 = 111$$

The pooled variance will be:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = 0.5317 \quad (10)$$

The test statistic is given by:

$$t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{s_p \sqrt{\frac{n_2 + n_1}{n_1 n_2}}} = 0.3314 \quad (11)$$

Using the rejection region approach, we reject H_0 if $t_{0.05,111} \geq -t$ where $t_{0.05,111} = -1.6587$. We have enough statistical evidence to reject null hypothesis since observed t is lesser than 1.6587

5.2 Case 2: Hypothesized testing if students want an average work experience of more than 1 year before going for further studies

Let us assume the null hypothesis as the average work experience of students willing to go for higher studies is less than 1 year. Let $\alpha = 0.05$.

The hypotheses are:

$$H_0 = \mu \leq \mu_0$$

$$H_a = \mu > \mu_0$$

where $\mu_0 = 1$

1. \bar{X} = Average years of work experience before going for further studies

2. S^2 = Sample standard deviation of number of years of work experience of students willing to go for higher studies
3. n = Number of students planning to go for further studies

Information:

$$\bar{X} = 1.33 \quad (12)$$

$$S^2 = 1.74 \quad (13)$$

$$n = 87 \quad (14)$$

$$df = n - 1 = 86 \quad (15)$$

The test statistic is given by:

$$t^* = \frac{\bar{X} - \mu_0}{S / \sqrt{n}} = \frac{1.33 - 1}{\frac{\sqrt{1.74}}{\sqrt{87}}} = 2.34 \quad (16)$$

Using the rejection region approach, we reject H_0 if $t^* \geq t_{0.05,86}$ where $t_{0.05,86} = 1.6628$. Hence, we have enough statistical evidence to reject null hypothesis H_0 .

5.3 Case 3: Comparing average work experience students take before going for higher studies in Management field vs in Research field

Let us assume the null hypothesis as avg work experience of students going for MBA is less than or equal to avg work experience of students going for MS/PhD
Let $\alpha = 0.05$

1. \bar{X}_1 = sample mean of avg work experience of students willing to go for MBA
2. \bar{X}_2 = sample mean of avg work experience of students willing to go for MS/PhD
3. \bar{s}_1 = sample variance of avg work experience of students willing to go for MBA
4. \bar{s}_2 = sample variance of avg work experience of students willing to go for MS/PhD

The hypotheses are:

$$H_0 = \mu_1 - \mu_2 \leq 0 \quad (17)$$

$$H_a = \mu_1 - \mu_2 > 0 \quad (18)$$

Information:

$$\bar{X}_1 = 1.54 \quad (19)$$

$$\bar{X}_2 = 1.25 \quad (20)$$

$$\bar{s}_1 = 1.30 \quad (21)$$

$$\bar{s}_1 = 1.90 \quad (22)$$

$$\bar{n}_1 = 24 \quad (23)$$

$$\bar{n}_2 = 63 \quad (24)$$

$$(25)$$

$$c = \frac{\frac{s_1^2}{n_1}}{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \frac{\frac{(1.30)^2}{24}}{\frac{(1.30)^2}{24} + \frac{(1.90)^2}{63}} = 0.57 \quad (26)$$

$$df = \frac{(n_1 - 1)(n_2 - 1)}{(1 - c)^2(n_1 - 1) + c^2(n_2 - 1)} \quad (27)$$

$$= \frac{(24 - 1)(63 - 1)}{(1 - 0.57)^2(24 - 1) + (0.57)^2(63 - 1)} = 58.45 \quad (28)$$

$$df \approx 54 \quad (29)$$

The test static is given by:

$$t' = \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (30)$$

where, $D_0 = 0$

$$t' = \frac{(1.54 - 1.25) - 0}{\sqrt{\frac{(1.30)^2}{24} + \frac{(1.90)^2}{63}}} = 0.81 \quad (31)$$

RR: $t' \geq t_{0.05}$, at $df = 54$

Here, $t_{0.05,54} = 1.67 > t'$

Since, t' does not fall in rejection region for $\alpha = 0.05$, we fail to reject H_0 . Here, we don't have enough evidence to say that avg work experience of MBA willing students is more than MS/PhD willing students.

5.4 Case 4: Hypothesized proportion testing if there is enough evidence that the proportions of people opting for masters, MBA, Phd are not all equal

Sample data :

Masters	MBA	PhD	Total
50	24	13	87

Let P_{Ms} , P_{MBA} , P_{PhD} denote proportions of students willing to pursue Masters, MBA, PhD for higher studies

$$H_0 : P_{Ms} = P_{MBA} = P_{PhD} = \frac{1}{3} \quad H_a : \text{atleast one } P \neq \frac{1}{3} \quad \alpha = 0.05$$

Also,

$$E = \frac{1}{3} \times 87 = 29 \quad (32)$$

and,

$$\chi^2 = \sum \frac{(O - E)^2}{E} \quad (33)$$

$$\begin{aligned} \chi^2 &= \frac{(50 - 29)^2}{29} + \frac{(24 - 29)^2}{29} + \frac{(13 - 29)^2}{29} \\ &= 15.2 + 0.862 + 8.827 \\ &= 24.889 \end{aligned} \quad (34)$$

At $df = 3 - 1 = 2$, p value = 0.0001

p value < $\alpha = 0.05$

Hence there is enough evidence that population proportions are not all equal.

6 Contribution

1. Kola Akshitha

- (a) Plots in Data Visualisation
- (b) Slides of Data visualisation
- (c) Latex report
- (d) Cleaning of data
- (e) Data Collection

2. Mouktika Cherukupalli

- (a) Plots in Data Visualisation
- (b) Slides of Data visualisation
- (c) Data Collection
- (d) Analysis of numerical data

3. Shreyas Wankhede
 - (a) Ideation of project
 - (b) Data Collection
 - (c) Data Cleaning
 - (d) Slides of Hypothesis testing
 - (e) Deciding cases for hypothesis testing and working in one of them
4. Pradeep Mundlik
 - (a) Ideation of project
 - (b) Data Collection
 - (c) Slides of Hypothesis testing
 - (d) Deciding cases for hypothesis testing and working in one of them
5. Avinash Malothu
 - (a) Plots and contingency tables in Data Visualisation
 - (b) Slides of Data visualisation
 - (c) Data Collection
6. Dhatri Reddy
 - (a) Ideation of project
 - (b) Data Collection
 - (c) Slides of Hypothesis testing
 - (d) Worked on one case of hypothesis testing
7. Sumeeth Kumar
 - (a) Worked on one case of Hypothesis testing
 - (b) Slides of hypothesis testing
 - (c) Data collection
8. Sathvika Marri
 - (a) Analysis and conclusions of categorical data
 - (b) Slides of Conclusions
 - (c) Data Collection