Intricacies of Study Environments at IIT Hyderabad

MA4240 - Applied Statistics

Prakhar Patni K N Vardhan Varunaditya Singhal
Tanmay Goyal Tanay Yadav Sujal

May 8, 2022

Outline

1. Introduction
2. Data Visualization
3. Data Analysis and Conclusions

Introduction

This project is based on the behavioral patterns of students studying at IITH. The vision was to see the changes that happened in the academic structure of our college after the strike of the pandemic, as well as how the students are dealing and coping with it. No. of study hours, preference of place, and material of studies were a few of the most common queries we asked. We use Statistics to deduce conclusions from the given data, assuming that the data is a random population sample.

Variables of interest

1. Where do you like to study?
2. Gender?
3. Do you prefer to study alone?
4. How many hours do you study in one go?
5. Do you prefer snacks while studying?
6. In which block you stay?
7. Do you prefer to study on bed or study table?
8. Which program are you enrolled in?
9. Which department are you in?
10. Do you prefer live lectures or recording?
11. Do you prefer to study with lecture recording or lecture slides?

Data Analysis Tools

We used NumPy, Pandas, and Matplotlib modules in Python for data analysis, data cleaning (and pre-processing), and data plotting respectively.

Pre-Processing of Data

The following steps were taken to pre-process the data:

1. We began by removing white-spaces and dropping columns not required for further analysis.
2. Since the names of the Hostel Blocks were to be an important variable for our comparisons, all entries not having the Hostel Block name were dropped.
3. Any existing `NaN`s were replaced with the modal values of the specific column, since we do not have any model to predict the unentered values.
4. The *Hostel Block* names were replaced with their letters and the *Department* names with their departmental codes.
5. Finally, to make the data more interpretable, we assumed that any person studying between n and $n+1$ hours would be studying $n+\frac{1}{2}$ hours on average.

Data Visualization

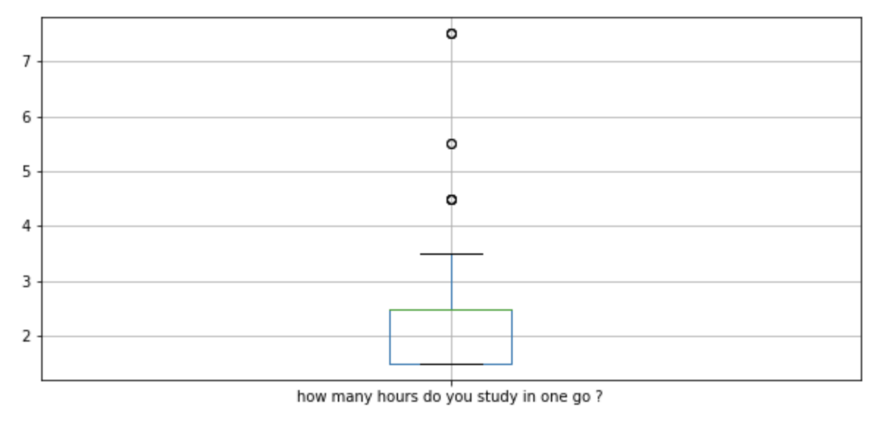
Analyzing the Uni-variate Numerical dataset

Table: How many hours do you study in one go? (in hours)

Count	118(non-null)
Mean	2.466102
Median	2.5
Mode	2.5
std	1.246754
min	1.5
25%	1.5
50%	2.5
75%	2.5
max	7.5
95% Confidence Interval	(2.238819810180139, 2.69338357965037)
99% Confidence Interval	(2.1656101375328065, 2.766593252297702)

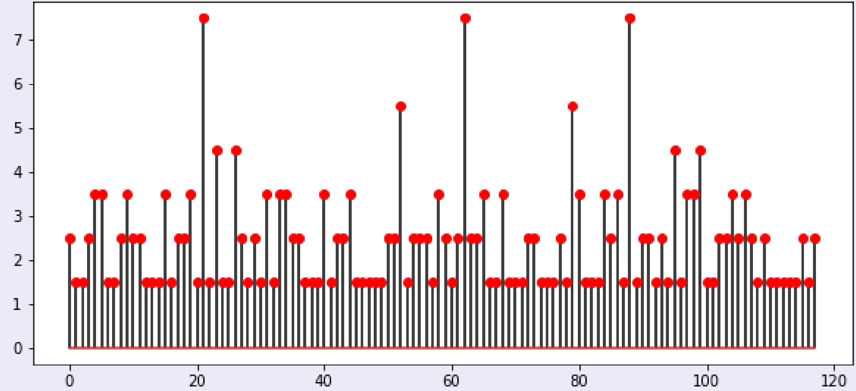
Visualizing the Uni-variate Numerical dataset

Figure: plot of uni-variate numerical dataset



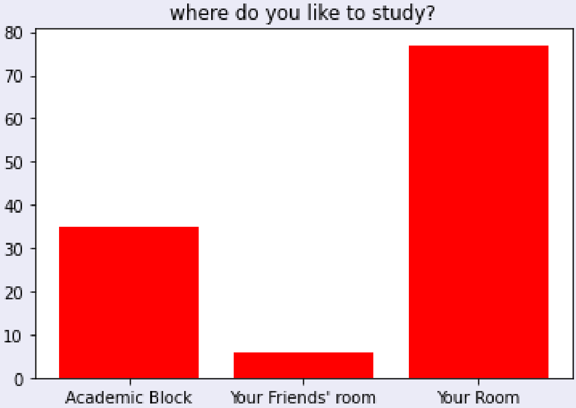
Visualizing the Uni-variate Numerical dataset

Figure: plot of uni-variate numerical dataset



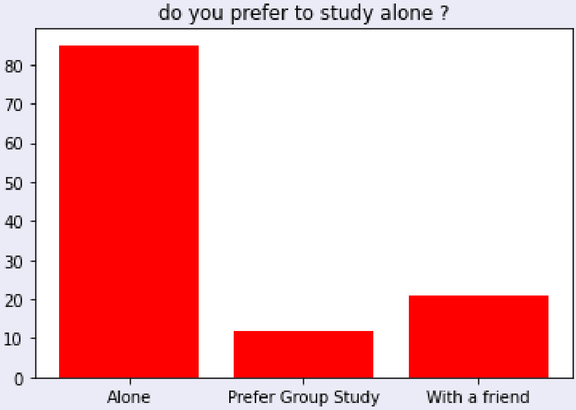
Data Visualization of Numerous Categorical Datasets

Figure: Counts to visualize the data of study place preference



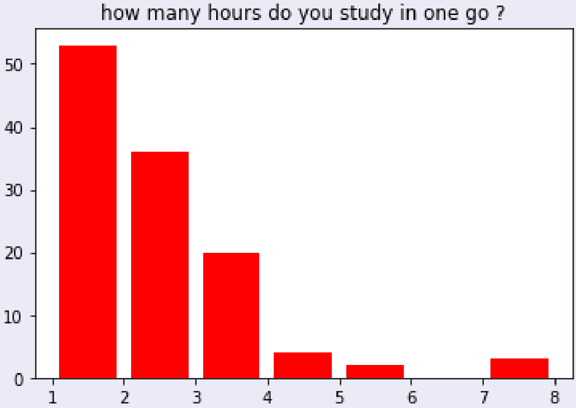
Data Visualization of Numerous Categorical Datasets

Figure: Counts to visualize the data of studying with friend/group or alone



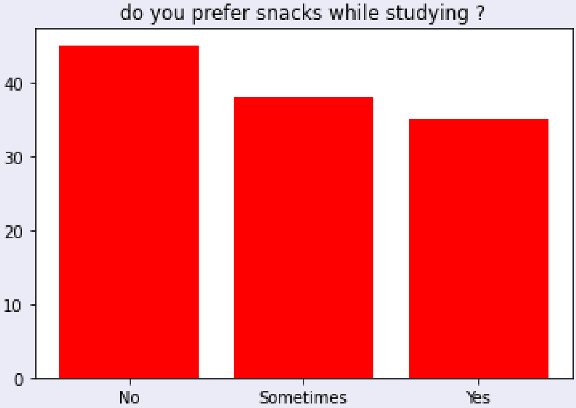
Data Visualization of Numerous Categorical Datasets

Figure: Counts to visualize the data of number of hours they study at one go



Data Visualization of Numerous Categorical Datasets

Figure: Counts to visualize the data of snacks preference while studying



Data Visualization of Numerous Categorical Datasets

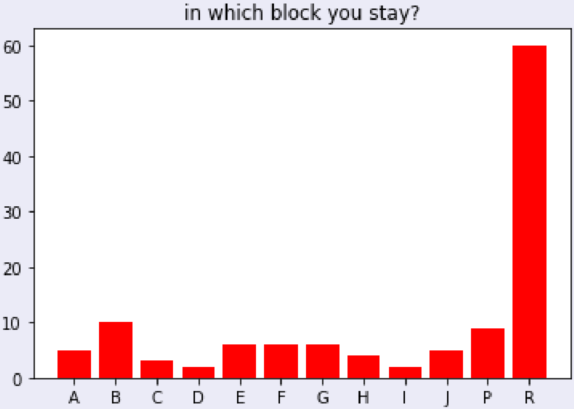
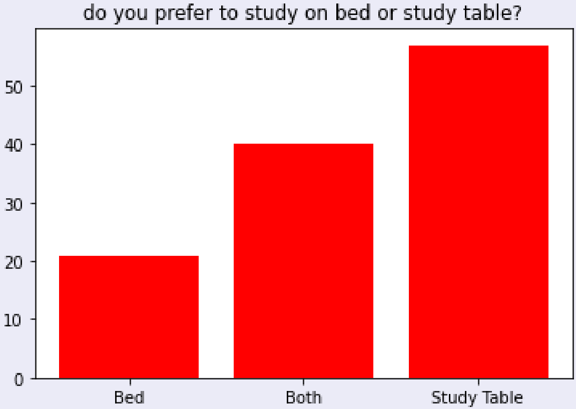


Figure: Counts to visualize the data of hostel block residence

Data Visualization of Numerous Categorical Datasets

Figure: Counts to visualize the data of preference of study on study table or bed



Data Visualization of Numerous Categorical Datasets

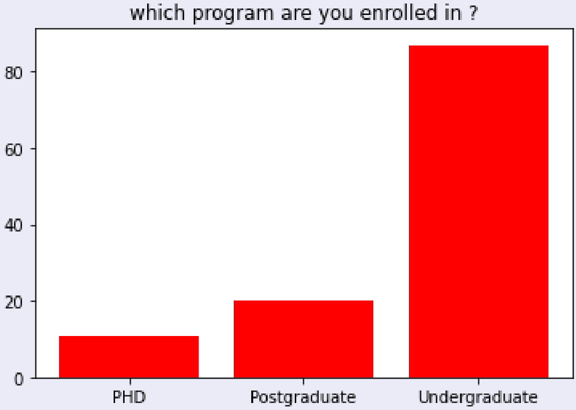
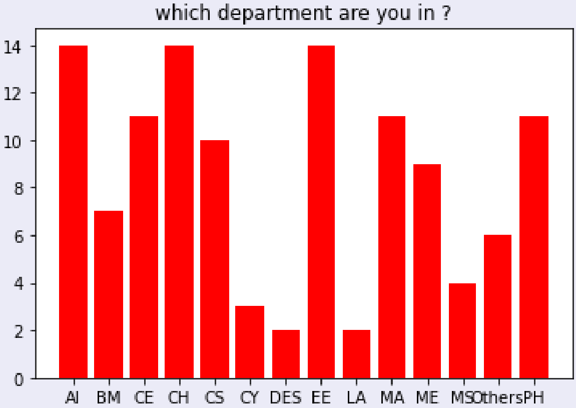


Figure: Counts to visualize the data of program enrolled in IITH

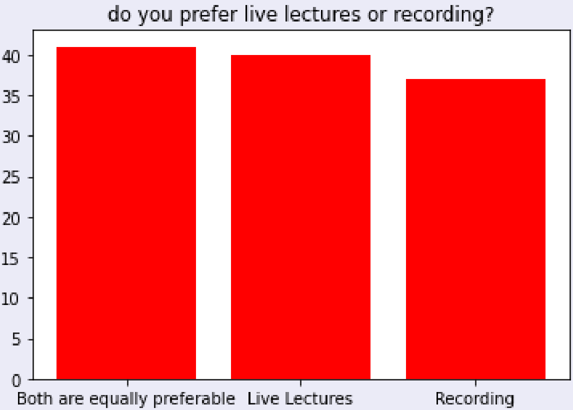
Data Visualization of Numerous Categorical Datasets

Figure: Counts to visualize the data of department of study



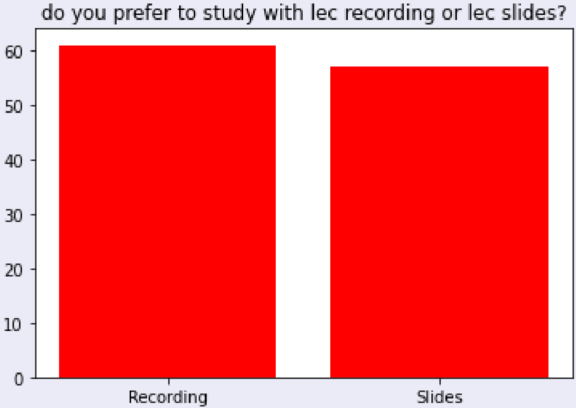
Data Visualization of Numerous Categorical Datasets

Figure: Counts to visualize the data live lectures preference



Data Visualization of Numerous Categorical Datasets

Figure: Counts to visualize the data of recordings/slides preference



Data Visualization with Segmented Bar plots

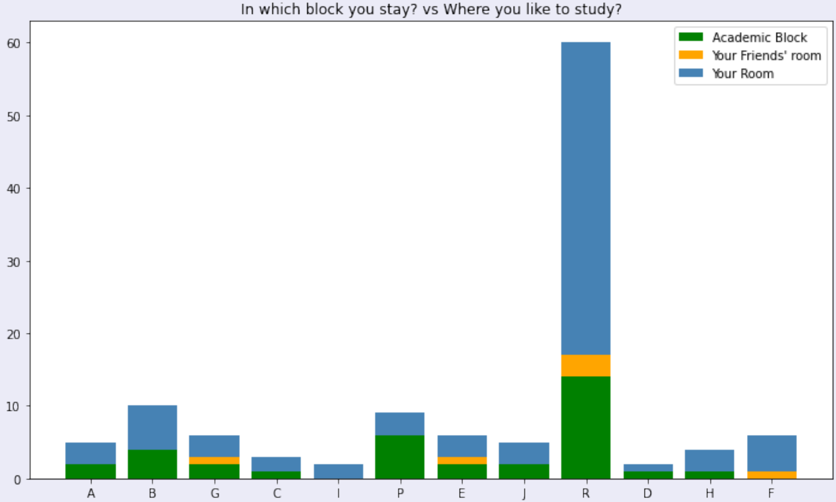


Figure: Categorical variables in a segmented bar plot

Data Visualization with Segmented Bar plots

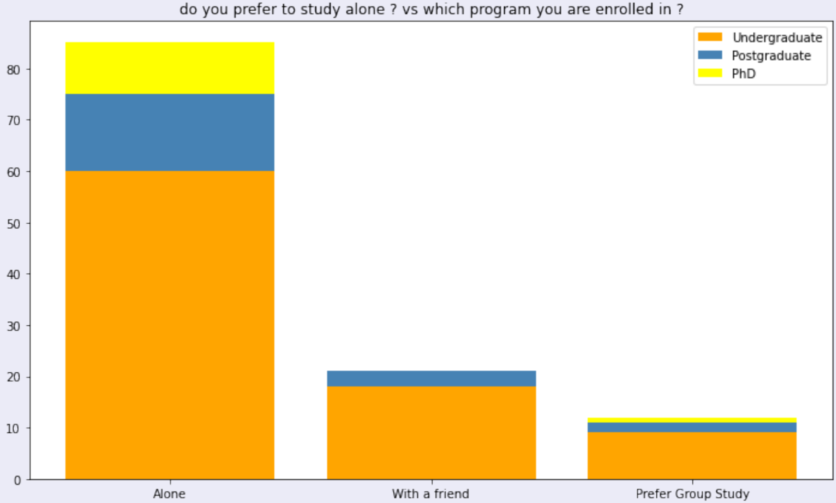
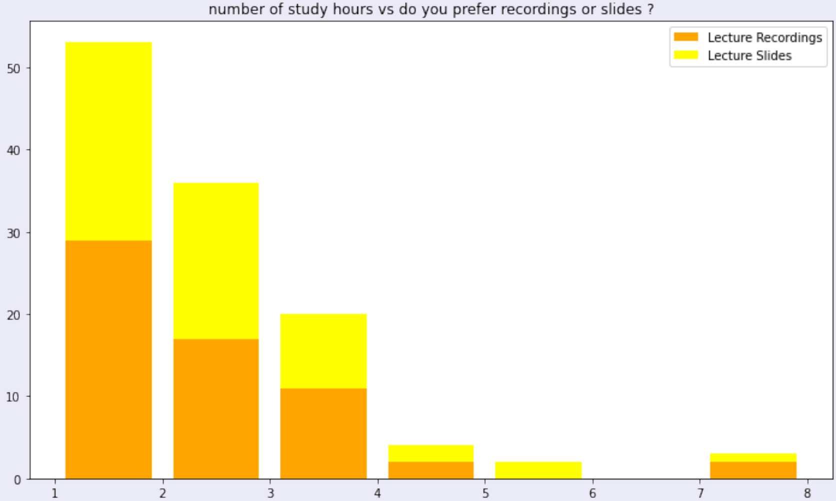


Figure: Categorical variables in a segmented bar plot

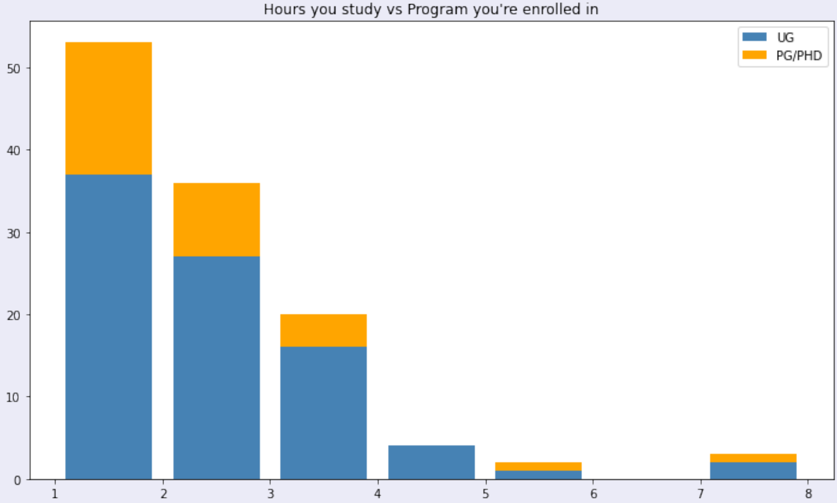
Data Visualization with Segmented Bar plots

Figure: Categorical variables in a segmented bar plot



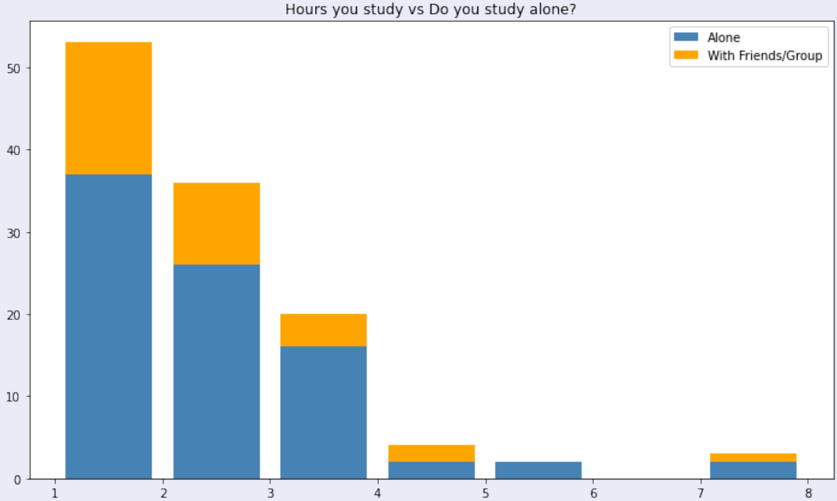
Data Visualization with Segmented Bar plots

Figure: Categorical variables in a segmented bar plot



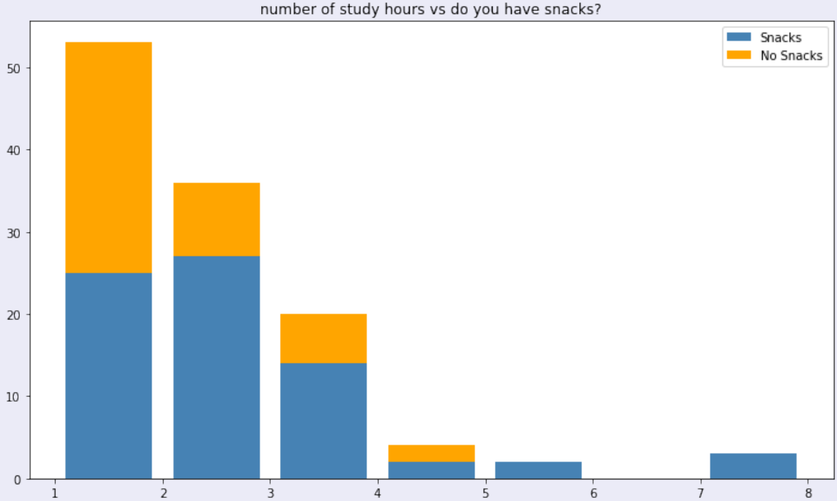
Data Visualization with Segmented Bar plots

Figure: Categorical variables in a segmented bar plot



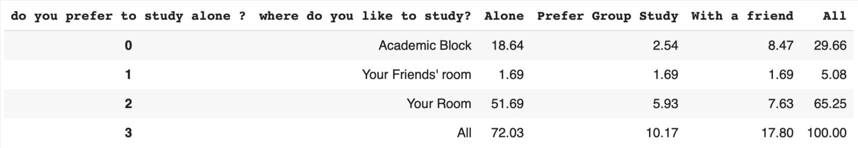
Data Visualization with Segmented Bar plots

Figure: Categorical variables in a segmented bar plot



Data Visualization with Contingency Table

Figure: Percentage Contingency table



do you prefer to study alone ?	where do you like to study?	Alone	Prefer Group Study	With a friend	All
0	Academic Block	18.64	2.54	8.47	29.66
1	Your Friends' room	1.69	1.69	1.69	5.08
2	Your Room	51.69	5.93	7.63	65.25
3	All	72.03	10.17	17.80	100.00

Inference

1. 51.69% people prefer to study alone in room.
2. While 18.64% people prefer to study alone in Academic Blocks.
3. And a very few percentage of people, i.e., 1.69% prefer to study in their friend's room.

Data Analysis and Conclusions

Uni-variate Numerical Dataset

With around 118 students participating in the survey, we get to see that the average time (in hours) the students study at one go is around 2.46 hours, i.e., $\mu = 2.466$ hrs with a standard deviation of 1.24, i.e., $\sigma = 1.246$ hrs, and median, mode being 2.5 hours.
And the confidence intervals for 95% and 99% are,

95% C.I - (2.239, 2.693)
99% C.I - (2.166, 2.767)

Hypothesis Testings

Case 1: Comparing the study hours for Undergraduates and Postgraduates

For Hypothesis Testing, we make the following statements -

$H_0 : \mu_1 - \mu_2 \geq 0$ and $H_a : \mu_1 - \mu_2 < 0$. Now,

$$\begin{aligned} \bar{x_1} &= 2.5 \text{ hours} \quad \bar{x_2} = 2.43 \text{ hours} \\ S_1^2 &= 0.5 \quad S_2^2 = 0.629 \\ n_1 &= 81 \quad n_2 = 37 \end{aligned}$$

Since $\dfrac{S_1^2}{S_2^2} = 1.258 < 4$, we can assume the population variances would be equal.

Hypothesis Testings

Case 1: Continued

The degrees of freedom, $df = n_1 + n_2 - 2 = 116$, and the pooled variance will be:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = 1.255 \tag{1}$$

The test statistic t is then given by:

$$t = 0.584 \tag{2}$$

Using the rejection region approach, we reject H_0 if $t \leq -t_{0.05,116}$, where $t_{0.05,116} = -1.658$.
Because the observed value of $t = 0.584$ is less than 1.658, we have enough statistical evidence to reject the null hypothesis, and thus, we can say, the postgraduates study more in one go than the undergraduates on average.

Hypothesis Testings

Case 2: Comparing the study hours of people who study alone and who study in groups

For Hypothesis Testing, we make the following statements -

$H_0 : \mu_1 - \mu_2 \geq 0$ and $H_a : \mu_1 - \mu_2 < 0$

$$\begin{aligned} \bar{x_1} &= 2.488 \text{ hours} & \bar{x_2} &= 2.409 \text{ hours} \\ S_1^2 &= 0.535 & S_2^2 &= 0.647 \\ n_1 &= 85 & n_2 &= 33 \end{aligned}$$

Since $\dfrac{S_1^2}{S_2^2} = 1.209 < 4$, we can assume the population variances would be equal.

Hypothesis Testings

Case 2: Continued

The degrees of freedom, $df = n_1 + n_2 - 2 = 116$, and the pooled variance will be:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = 0.5658 \tag{3}$$

The test statistic t is then given by:

$$t = 0.512 \tag{4}$$

Using the rejection region approach, we reject H_0 if $t \leq -t_{0.05,116}$, where $t_{0.05,116} = -1.658$.
Because the observed value of $t = 0.512$ is less than 1.658, we have enough statistical evidence to reject the null hypothesis, and thus, we can say, those who study in groups study more on average in one go than those who study alone.

Hypothesis Testings

Case 3: Comparing the study hours of people who study while eating snacks and who study without eating snacks

For Hypothesis Testing, we make the following statements -

$H_0 : \mu_1 - \mu_2 \geq 0$ and $H_a : \mu_1 - \mu_2 < 0$

$$\bar{x_1} = 2.671 \text{ hours} \quad \bar{x_2} = 2.379 \text{ hours}$$
$$S_1^2 = 0.91 \quad S_2^2 = 0.4$$
$$n_1 = 35 \quad n_2 = 83$$

Since $\dfrac{S_1^2}{S_2^2} = 2.275 < 4$, we can assume the population variances would be equal.

Hypothesis Testings

Case 3: Continued

The degrees of freedom, $df = n_1 + n_2 - 2 = 116$, and the pooled variance will be:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = 0.55 \tag{5}$$

The test statistic t is then given by:

$$t = 1.9685 \tag{6}$$

Using the rejection region approach, we reject H_0 if $t \leq -t_{0.05,116}$, where $t_{0.05,116} = -1.658$.

Because the observed value of $t = 1.9685$ is greater than 1.658, we fail to reject the null hypothesis, and thus, do not have enough evidence to say that those who do not eat snacks while studying study for longer at one go than those who do not.

Hypothesis Testings

Case 4: Comparing the study hours of people who study from Lecture Slides and those who study from Lecture Recordings

For Hypothesis Testing, we make the following statements -

$H_0 : \mu_1 - \mu_2 \geq 0$ and $H_a : \mu_1 - \mu_2 < 0$

$$\bar{x_1} = 2.5 \text{ hours} \quad \bar{x_2} = 2.43 \text{ hours}$$
$$S_1^2 = 0.5 \quad S_2^2 = 0.629$$
$$n_1 = 57 \quad n_2 = 61$$

Since $\dfrac{S_1^2}{S_2^2} = 1.258 < 4$, we can assume the population variances would be equal.

Hypothesis Testings

Case 4: Continued

The degrees of freedom, $df = n_1 + n_2 - 2 = 116$, and the pooled variance will be:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = 0.5764 \tag{7}$$

The test statistic t is then given by:

$$t = 0.5 \tag{8}$$

Using the rejection region approach, we reject H_0 if $t \leq -t_{0.05,116}$, where $t_{0.05,116} = -1.658$.

Because the observed value of $t = 0.5$ is lesser than 1.658,we have enough statistical evidence to reject the null hypothesis, and thus, we can say, those who study from recordings study more on average in one go than those who study from slides.

THANK YOU

MA4240 - Applied Statistics

May 8, 2022