MA4240: Applied Statistics
Project Report

Rutv Kocheta
MA21BTECH11014

Tanay Chandra
MA21BTECH11016

Vatsal Chaudhary
MA21BTECH11017

Ankit Saha
AI21BTECH11004

Pranav Balasubramanian
AI21BTECH11023

Aayush Prabhu
AI21BTECH11002

Pranav Seshu
MA21BTECH11008

Hari Vamshi
AI21BTECH11014

May 1, 2023

Abstract

The aim of this statistics project is to analyze and compare the population of IIT Hyderabad students based on their dietary habits, physical activity, expenditure, and sleep patterns. The study surveyed a sample of students. The data was analyzed using statistical methods. Overall, this study provides valuable insights into the relationships between these lifestyle factors.

1 Introduction

The project studies the relationship between a person's food habits and daily lifestyle in IIT-H. The primary aim of this study is to see how an individual's day-to-day life is affected by their food habits. The type of diet, hours of physical activity done in a week, sleep hours, and average expenditure were a few of the queries we asked in the survey. We use statistics to deduce conclusions from the given data, assuming that the data is a random population sample.

The survey consisted of the following questions:

1. Are you vegetarian or non-vegetarian?

2. How many hours a week do you engage in physical activities (e.g. gym, sports)

3. How many hours a day do you sleep?

4. How many times in a month do you eat outside the mess?

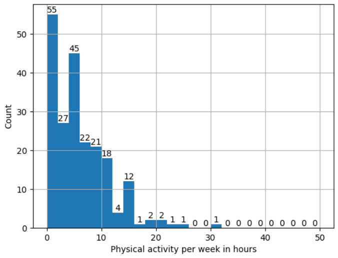
5. How much money do you spend on food (apart from mess food) per day? (includes food from vending machines, mess extras, canteen, restaurants, food courts, etc.)

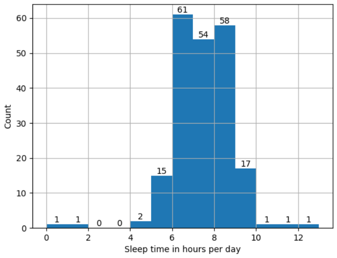
6. Which meal do you look forward to the most in a day?

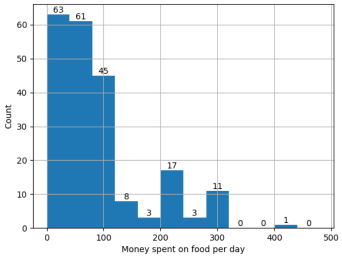
7. Do you take naps in the afternoon?

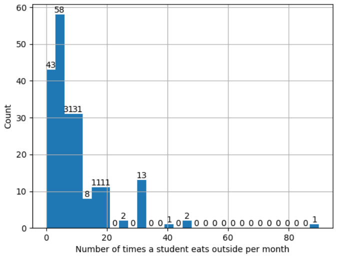
2 Data visualisation and pre-processing

2.1 Histograms

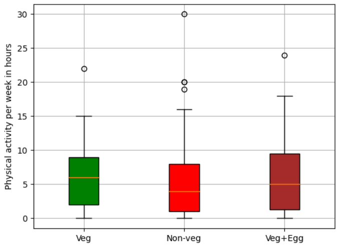


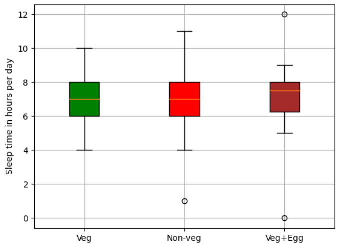


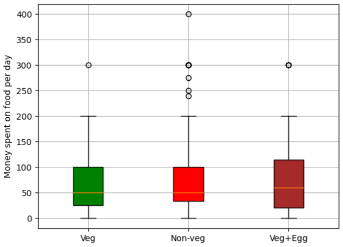


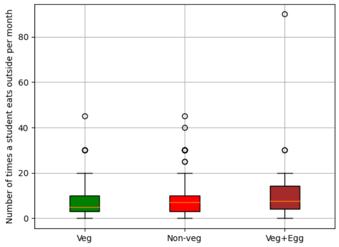


2.2 Box Plots









3 CLT Plots

The CLT plots are generated by taking a random sample of size n and finding the mean of the sample and a large number of such samples are plotted on the histogram. The large spikes in the graphs can be attributed to the high discreteness of the data.

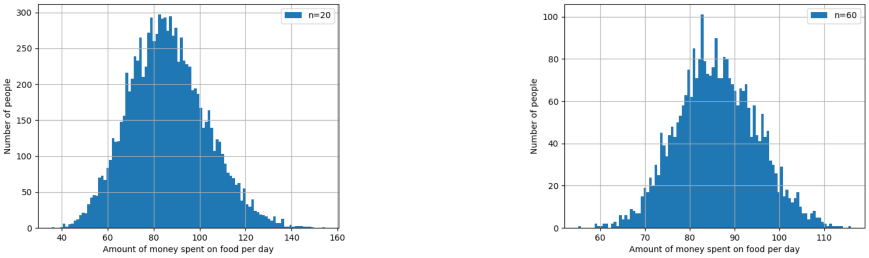


Figure 1: CLT plots of money spent with mean samples of size 20 and 60

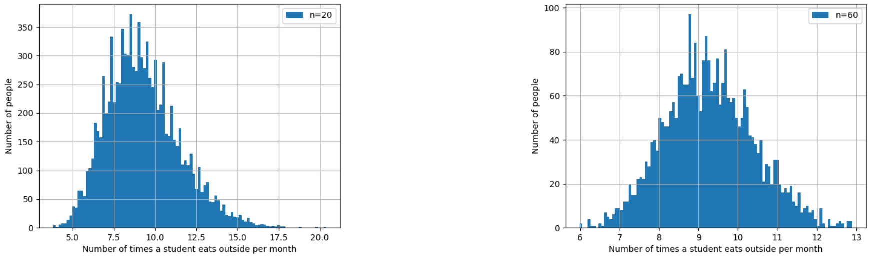


Figure 2: CLT plots of frequency of eating out per month with mean samples of size 20 and 60

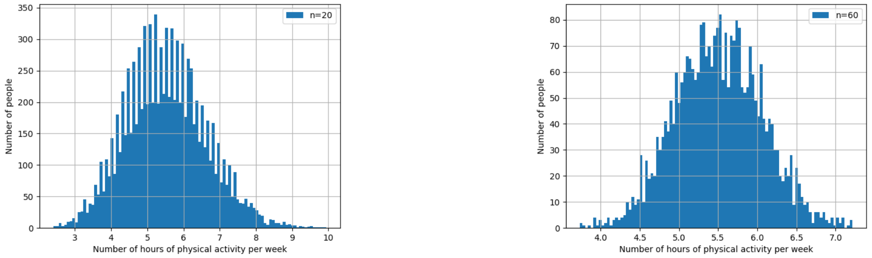


Figure 3: CLT plots of hours of physical activity per week with mean samples of size 20 and 60

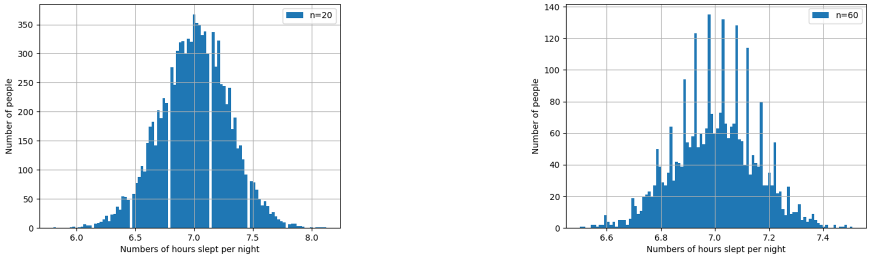


Figure 4: CLT plots of hours of sleeping time per day with mean samples of size 20 and 60

4 F-distribution plots

The F distribution plots are generated by taking random samples of size $n = 10$ for vegetarians and $m = 20$ for non-vegetarians, and taking a large number of such random samples, and calculating the ratio of the sample variance of the both samples. These ratios are then plotted and shown below.

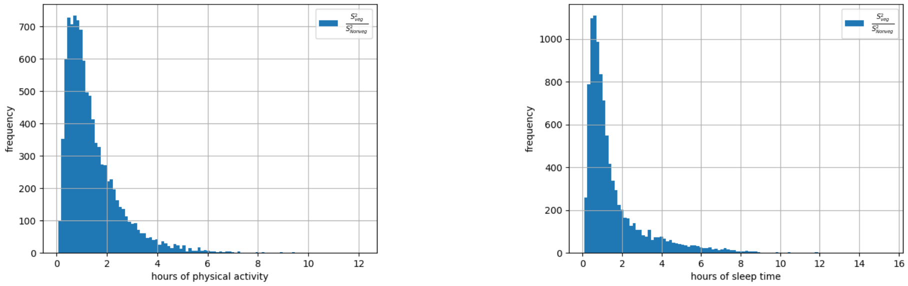


Figure 5: CLT plots of time slept per day with means samples of size 20 and 60

5 Normality plots

Normality plots are used to assess whether our data set is approximately following a normal distribution by checking if the quartile plot is approximately a straight line.

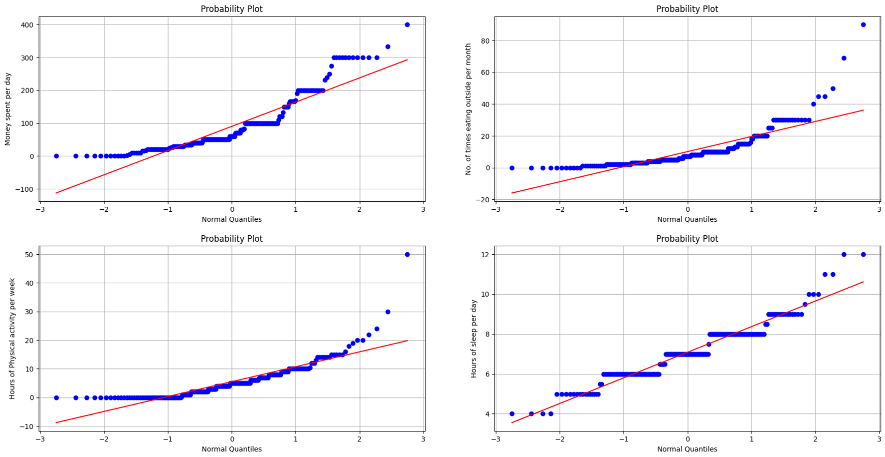


Figure 6: Normality Plots

From these plots we can infer that plot-2 and plot-3 (top right and bottom left) approximately follow normal distribution in the quartile range of -1 to 2

6 Hypothesis testing

Hypothesis 1

Non-vegetarian people spend more money on food on average compared to vegetarian people.

1.1 Formulation

We have a random sample of money spent $X_1, \ldots, X_n$ of size $n = 186$ from the non-vegetarian population and $Y_1, \ldots, Y_m$ of size $m = 47$ from the vegetarian population.
Our hypotheses are as follows:

$$H_0 : \mu_X \leq \mu_Y$$
$$H_a : \mu_X > \mu_Y$$

Since $n > 30$ and $m > 30$, we can safely use the *central limit theorem.* Assume that the non-vegetarian and vegetarian samples are independent, which is a reasonable assumption.

We have $\overline{x} = 97.54$, $\overline{y} = 75.73$, $s_X = 98.61$, $s_Y = 66.18$, $\mu_0 = 0$

$$\frac{s_X}{s_Y} = 1.49$$

Since $s_X/s_Y < 2$ and the random samples are independent, we can use the **right-tailed pooled t-test**. Consider a confidence level of 95%, i.e., $\alpha = 0.05$

1.2 Test statistic

The test statistic in this case is given by

$$t^* = \frac{(\overline{x} - \overline{y}) - \mu_0}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}} = 1.41$$

where the pooled variance s_p is given by

$$s_p = \sqrt{\frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2}} = 94.49$$

1.3 Rejection region approach

The critical t-value for $\alpha = 0.05$ and $n + m - 2 = 231$ degrees of freedom is

$$t_{\alpha,\, n+m-2} = t_{0.05,\, 231} = 1.65$$

The rejection region for this right-tailed test is thus

$$\{t : t \geq 1.65\}$$

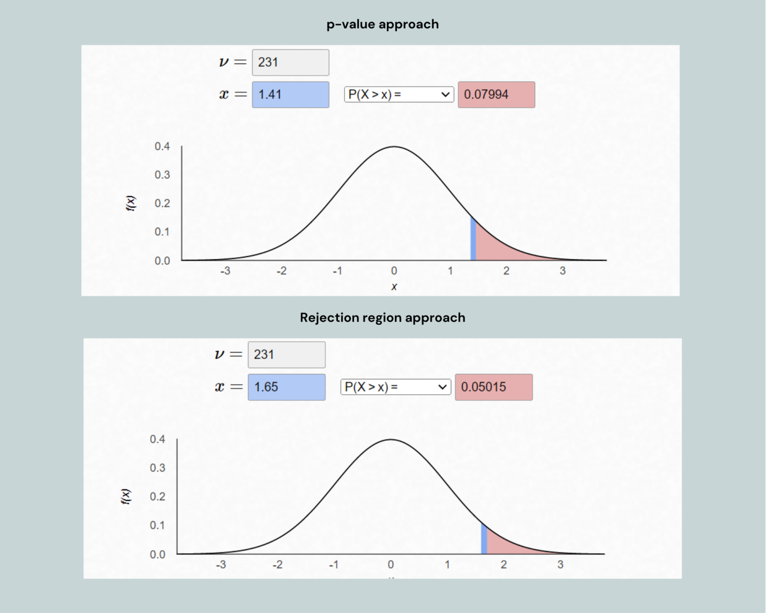
Since $1.41 < 1.65$, the test statistic lies outside the rejection region. Hence we fail to reject the null hypothesis.

1.4 p-value approach

The p-value corresponding to the above test statistic t^* is given by

$$p = \Pr\left(T_{231} \geq t^*\right) = \Pr\left(T_{231} \geq 1.41\right) = 0.079$$

Since $0.079 > 0.05$, the p-value is greater than the significance level α. Hence, we fail to reject the null hypothesis.



p-value approach

$\nu =$ 231

$x =$ 1.41 P(X > x) = 0.07994

Rejection region approach

$\nu =$ 231

$x =$ 1.65 P(X > x) = 0.05015

Figure 7: Rejection region and p-value of hypothesis 1

1.5 Conclusion

By both the above approaches, we fail to reject the null hypothesis. We thereby conclude that we do not have enough statistical evidence at 95% confidence to say that non-vegetarian people spend more money on food than vegetarian people.

Hypothesis 2

Physically active people have a more consistent number of sleep hours than those who are physically inactive, i.e., the variance in their sleeping hours is smaller than the variance in the sleeping hours of physically inactive people.

Note: People with more than 2 hours of physical activity per week are considered physically active.

2.1 Formulation

We have a random sample $X_1, \ldots, X_n$ of size $n = 150$ from the physically active population and $Y_1, \ldots, Y_m$ of size $m = 83$ from the physically inactive population. Our hypotheses are as follows:

$$H_0 : \ \sigma_X^2 \geq \ \sigma_Y^2$$
$$H_a : \ \sigma_X^2 < \ \sigma_Y^2$$

Since $n > 30$ and $m > 30$, we can use the **left-tailed F-test**. Consider a confidence level of 95%, i.e., $\alpha = 0.05$

2.2 Test statistic

We have $s_1^2 = 2.097, s_2^2 = 2.053$
The test statistic in this case is given by

$$F^* = \frac{s_1^2}{s_2^2} = 1.021$$

2.3 Rejection region approach

The critical F-value for $\alpha = 0.05$ and degrees of freedom $df_1 = n - 1 = 149$ and $df_2 = m - 1 = 82$ is

$$F_{1-\alpha, df_1, df_2} = F_{0.95, 149, 82} = 0.731$$

The rejection region for this one-tailed test is thus

$$\{F : F \leq 0.731\}$$

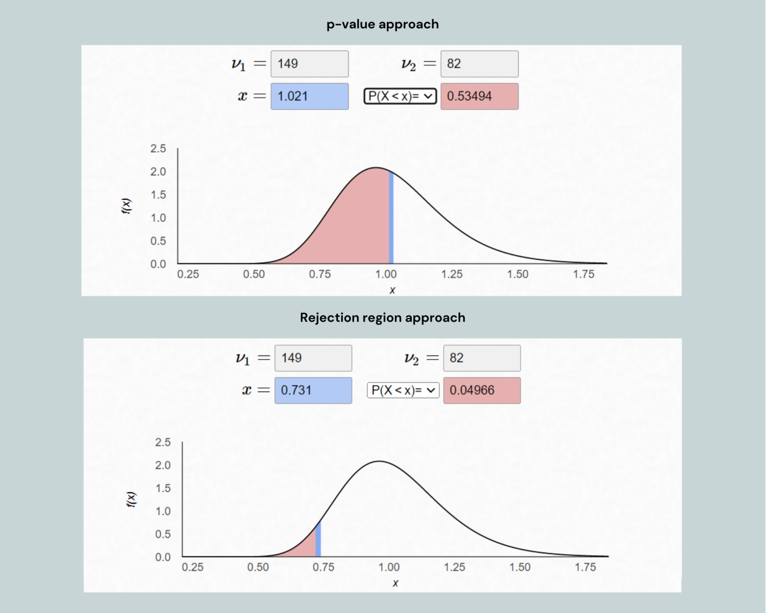
Since $1.021 > 0.731$, the test statistic lies outside the rejection region. Hence, we fail to reject H_0.

2.4 p-value approach

The p-value corresponding to the above test statistic F^* is given by

$$p = \Pr\left(F_{149,\,82} \leq F^*\right) = \Pr\left(F_{149,\,82} \leq 1.021\right) = 0.535$$

Since $0.535 > 0.05$, the p-value is greater than the significance level α. Hence, we fail to reject the null hypothesis.



p-value approach

$\nu_1 =$ 149 $\nu_2 =$ 82

$x =$ 1.021 P(X<x)= 0.53494

Rejection region approach

$\nu_1 =$ 149 $\nu_2 =$ 82

$x =$ 0.731 P(X<x)= 0.04966

Figure 8: Rejection region and p-value of the hypothesis 2

2.5 Conclusion

By both the above approaches, we fail to reject the null hypothesis. We thereby conclude that we do not have enough statistical evidence at 95% confidence to say that physically active people have a more consistent number of sleep hours than those who are physically inactive.

Hypothesis 3

The proportion of lunch-preferring people who take naps in the afternoon is greater than the proportion of people prefer meals other than lunch who take naps in the afternoon.

3.1 Formulation

We have a random sample $X_1, \dots, X_n$ of size $n = 45$ who prefer lunch and $Y_1, \dots, Y_m$ of size $m = 188$ who prefer breakfast or dinner.
Our hypotheses are as follows:

$$H_0 : \pi_X \leq \pi_Y$$
$$H_a : \pi_X > \pi_Y$$

We have $\hat{\pi}_X = 0.53$, $\hat{\pi}_Y = 0.51$
$n\hat{\pi}_X = 23.85$, $n(1 - \hat{\pi}_X) = 21.15$, $m\hat{\pi}_Y = 95.88$, $m(1 - \hat{\pi}_Y) = 92.12$. Since all values are greater than 5, we can use the **right-tailed test**. Consider a confidence level of 95%, i.e., $\alpha = 0.05$

3.2 Test statistic

The test statistic in this case is given by

$$z^* = \frac{\hat{\pi}_X - \hat{\pi}_Y}{\sqrt{\frac{\hat{\pi}_X(1-\hat{\pi}_X)}{n} + \frac{\hat{\pi}_Y(1-\hat{\pi}_Y)}{m}}} = 0.24$$

3.3 Rejection region approach

The critical z-value for $\alpha = 0.05$ is

$$z_{0.05} = 1.645$$

The rejection region for this right-tailed test is thus

$$\{z : z \geq 1.645\}$$

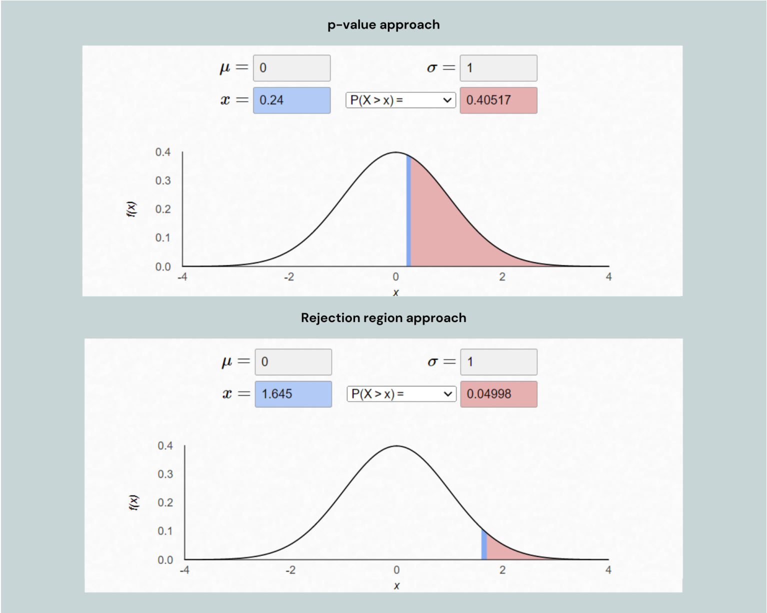
Since $0.24 < 1.645$, the test statistic lies outside the rejection region. Hence we fail to reject the null hypothesis.

3.4 p-value approach

The p-value corresponding to the above test statistic z^* is given by

$$p = \Pr\left(z \geq z^*\right) = \Pr\left(z \geq 0.24\right) = 0.405$$

Since $0.405 > 0.05$, the p-value is greater than the significance level α. Hence, we fail to reject the null hypothesis.



p-value approach

$\mu =$ 0 $\sigma =$ 1

$x =$ 0.24 P(X > x) = 0.40517

Rejection region approach

$\mu =$ 0 $\sigma =$ 1

$x =$ 1.645 P(X > x) = 0.04998

Figure 9: Rejection region and p-value of the hypothesis 3

3.5 Conclusion

By both the above approaches, we fail to reject the null hypothesis. We thereby conclude that we do not have enough statistical evidence at 95% confidence to say that the proportion of lunch-preferring people who take naps in the afternoon is greater than the proportion of breakfast/dinner-preferring people who take naps in the afternoon.

7 Confidence intervals

We find confidence intervals for the mean expenditure of vegetarian and non-vegetarian people.

7.1 Mean expenditure of non-vegetarian people

The confidence interval for the mean expenditure of non-vegetarian people is given by

$$\left(\overline{x} - t_{\alpha/2,\, n-1}\frac{S_X}{\sqrt{n}},\ \overline{x} + t_{\alpha/2,\, n-1}\frac{S_X}{\sqrt{n}}\right)$$

We set a 95% confidence level so that $\alpha = 0.05$. We have $\overline{x} = 97.54$, $s_X = 98.61$, $n = 186$ and $t_{0.025,\, 185} = 1.97287$. Thus, the confidence interval is

$$(83.275,\ 111.804)$$

7.2 Mean expenditure of non-vegetarian people

The confidence interval for the mean expenditure of vegetarian people is given by

$$\left(\overline{y} - t_{\alpha/2,\, m-1}\frac{S_Y}{\sqrt{m}},\ \overline{y} + t_{\alpha/2,\, m-1}\frac{S_Y}{\sqrt{m}}\right)$$

We set a 95% confidence level so that $\alpha = 0.05$. We have $\overline{y} = 75.73$, $s_Y = 66.18$, $m = 47$ and $t_{0.025,\, 46} = 2.0129$. Thus, the confidence interval is

$$(56.298,\ 95.161)$$

8 Individual contributions

1. Rutv
   - Ideation for the project
   - Collecting Data
   - F-distribution plots and their code
   - Confidence interval estimation
2. Tanay
   - Ideation for the project
   - Collecting Data
   - Hypothesis Testing
   - LaTeX Report
3. Vatsal

    - Ideation for the project
    - Collecting Data
    - Hypothesis Testing
    - LaTeX Report

4. Ankit
    - Ideation for the project
    - Collecting Data
    - Hypothesis Testing
    - Confidence interval estimation
    - LaTeX Report

5. Pranav B
    - Ideation for the project
    - Collecting Data
    - Histograms and box plots
    - CLT plots
    - p-value and rejection region plots
    - LaTeX Report

6. Aayush
    - Ideation for the project
    - Collecting data
    - Normality Plots and their code
    - LaTeX Report

7. Pranav S
    - Ideation for the project
    - Collecting Data
    - Hypothesis Testing
    - LaTeX Report

8. Vamshi

- Ideation for the project
- Collecting Data
- Hypothesis Testing
- LaTeX Report

9 Software used

1. LaTeX for typesetting
2. Numpy for numerical computations
3. Pandas for processing data
4. Matplotlib for plotting various plots
5. Google Forms for collecting data
6. `https://homepage.divms.uiowa.edu/~mbognar/applets/` for calculating critical values and p-values for different distributions