Mess Food Usage by IITH Student Community

MA4240:Applied Statistics - Group Project

Blessy Anvitha Saanvi Amrutha Anirudh Srinivasan Arnav Asati
Lokesh Surana Bhanu Prasad Samar Singhai Shivanand

May 2, 2023

Outline

Introduction

Through this project we tried to understand **how the students of IIT Hyderabad are using the mess food**. We tried observing if the student community is properly using the mess food or there's a lot wastage. Most of the possible variables that effect the usage like reasons for skipping mess, degree the student pursuing, financial status, etc., are being asked for in the survey conducted. For drawing the results we opted observational study, We tried including all the possible groups and hence assuming that the data is a random population sample.

Variables of Interest

1. How many times a week you skip eating mess food?
2. Early sleeping or Late-night sleeping?
3. Which meal do you usually skip in a day?
4. Are you a vegetarian or non-vegetarian?
5. Food Preference?
6. What are the reasons for you to skip?
7. What are you pursuing?
8. What department are you in?
9. Gender and Age?
10. Family Financial Status?
11. What's your preferable alternative?

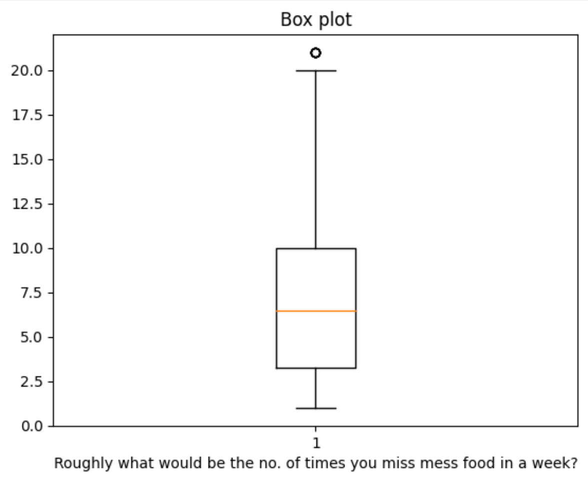
Analyzing the Uni-variate Numerical dataset

Table 1: How many times a week you skip eating mess food?

count	300
Mean	7.296
Median	6.5
Mode	10
Std	4.73
Min	0
25%	3.5
50%	6.5
75%	10
Max	21

Box Plot

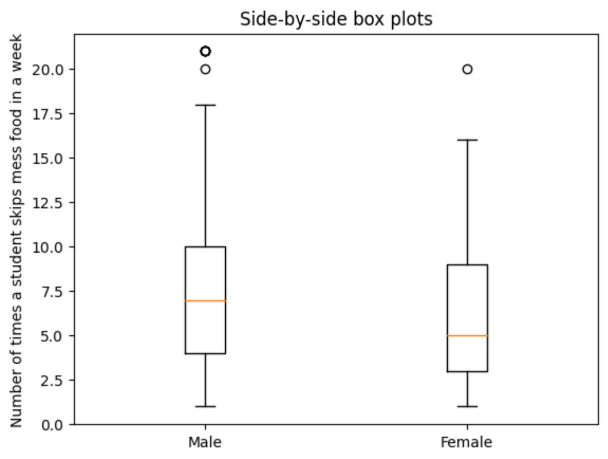
Figure 1: plot of uni-variate numerical dataset



Data Visualization

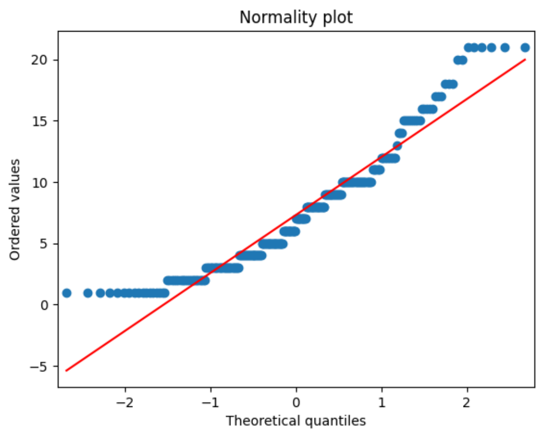
Male vs Female

Figure 2: Side-by-Side Boxplot



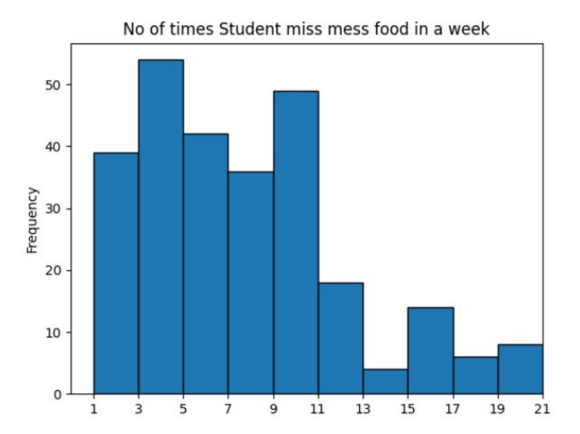
Normality plot

Figure 3: Analysing whether the univariate numerical dataset is normally distributed or not



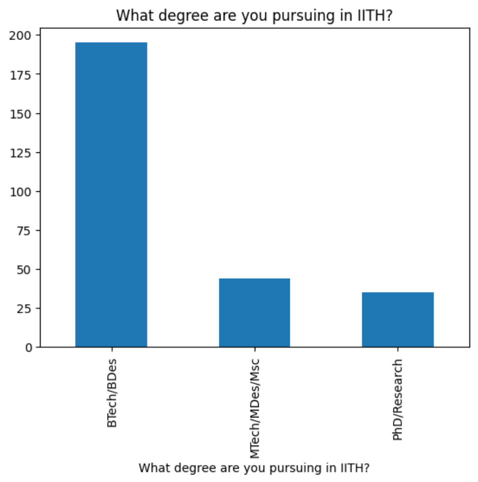
Frequency of skipping mess

Figure 4: Frequency of Number of times Meal skip



Participation based on degree

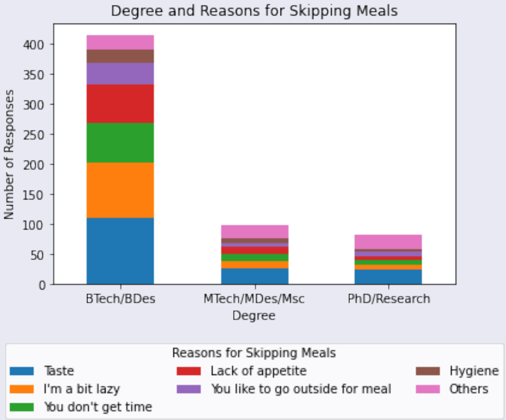
Figure 5: Degree vs number of responses



Data Analysis

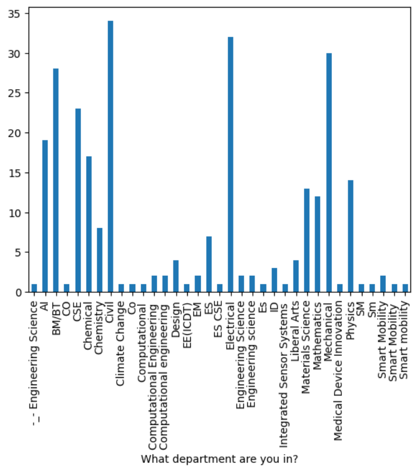
Reason of skipping based on Degree

Figure 6: Reason vs degree vs meals skipped



Participation based on branch

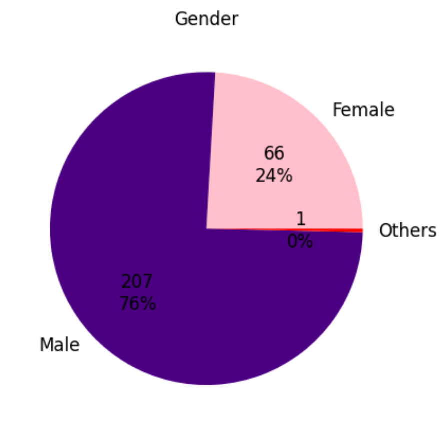
Figure 7: Branch vs number of responses



Exploratory Data Analysis

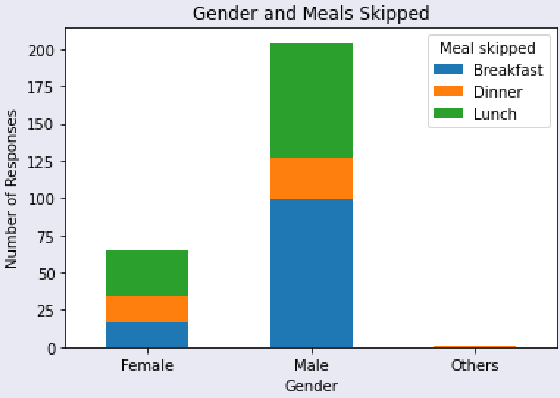
Involvement in terms of gender

Figure 8: Male responses vs female responses



Gender and skipping meals

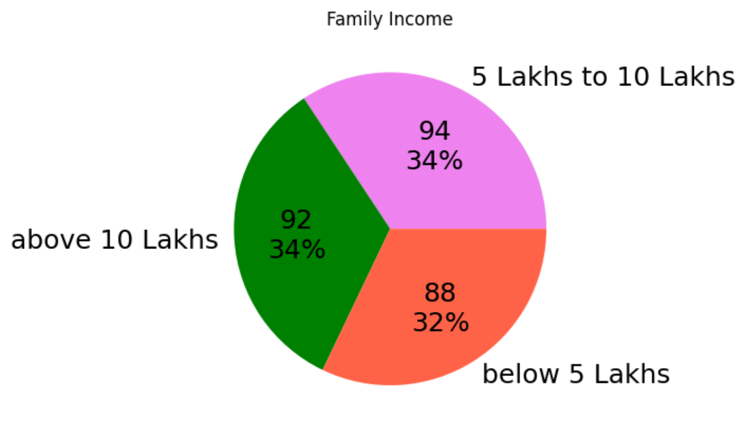
Figure 9: Gender vs Meals skipped



Exploratory Data Analysis

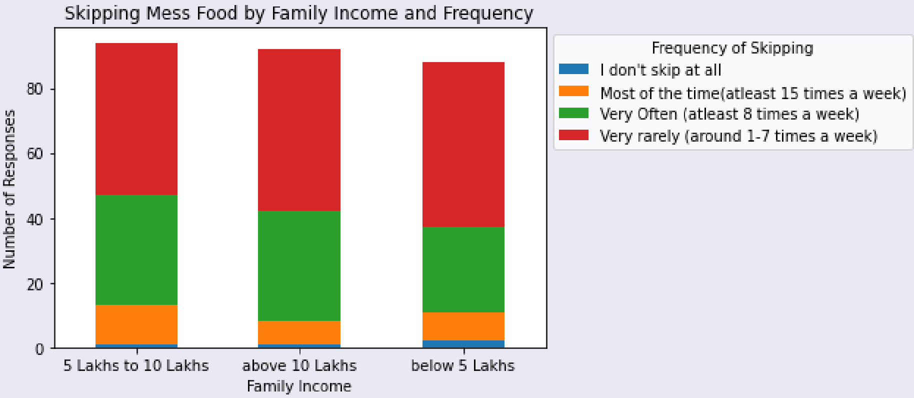
Economic classes

Figure 10: Share of different economic classes



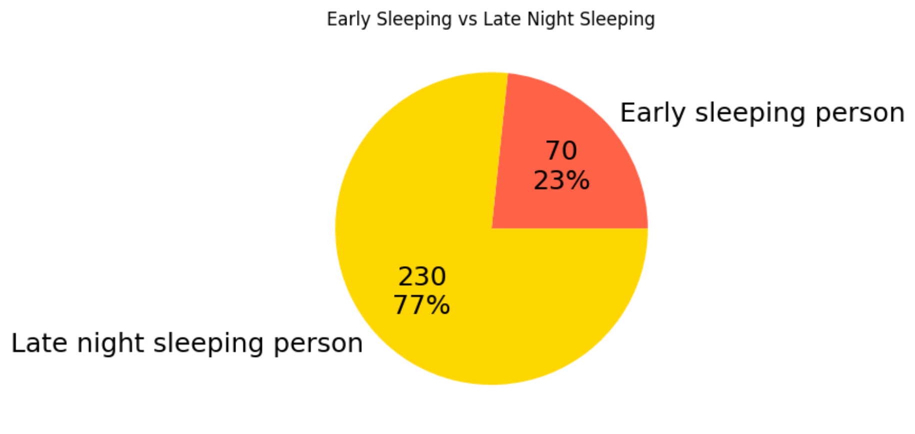
Relation of family income and skipping meals

Figure 11: Family income vs Meals skipped



Majority are late night sleepers!!

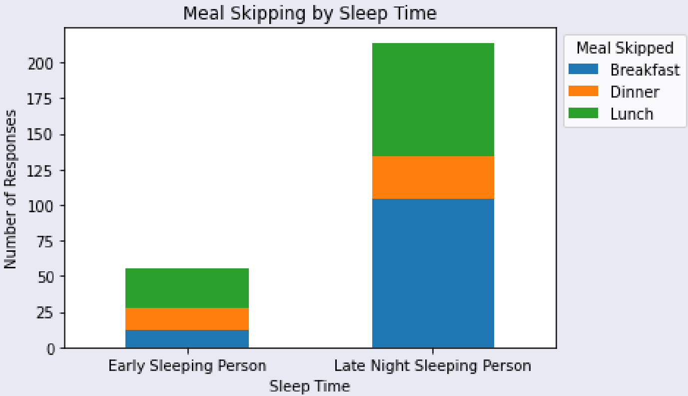
Figure 12: Early sleepers vs Late sleepers



Data Analysis

Sleep time and meal skipping

Figure 13: Sleep Time vs Meals skipped



Data analysis

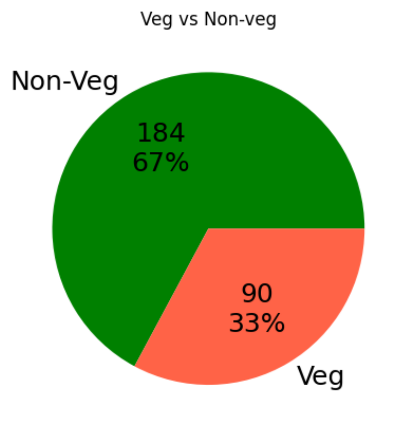
Contingency Table

Table 2: Contingency table

		Meal Skipped			
		Breakfast	Lunch	Dinner	Total
Sleep Time	Early sleeping person	12	28	16	56
	Late night sleeping person	104	80	30	214
	Total	116	108	46	270

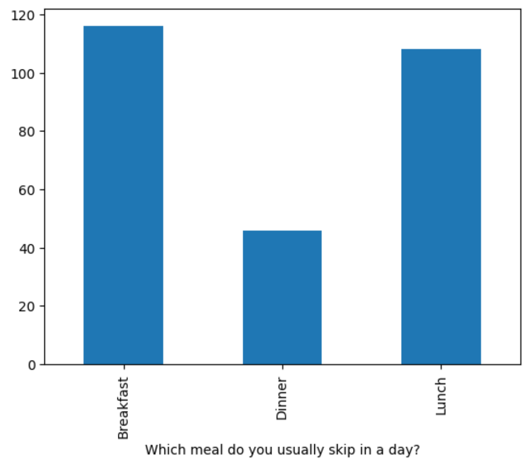
Non-veg vs veg

Figure 14: Share of non-veg and veg



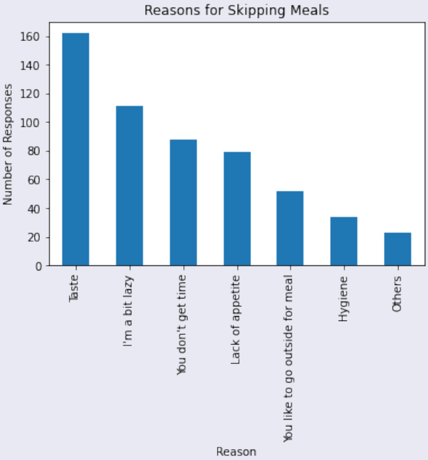
Which meal is being skipped the most??

Figure 15: Meal skip vs type of meal



Reasons for skipping

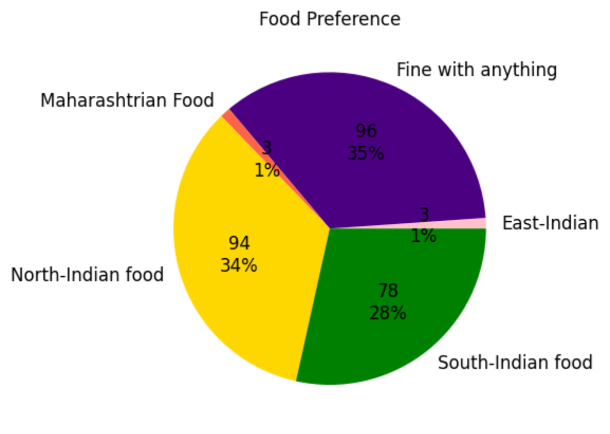
Figure 16: Reasons vs Number of responses



Data Analysis

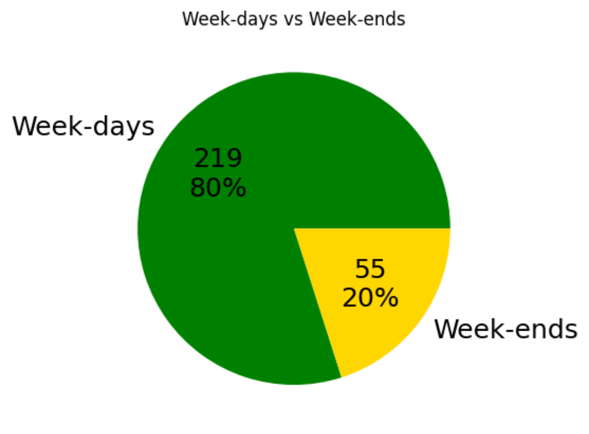
Which food is preferred the most?

Figure 17: Preference of Regional food



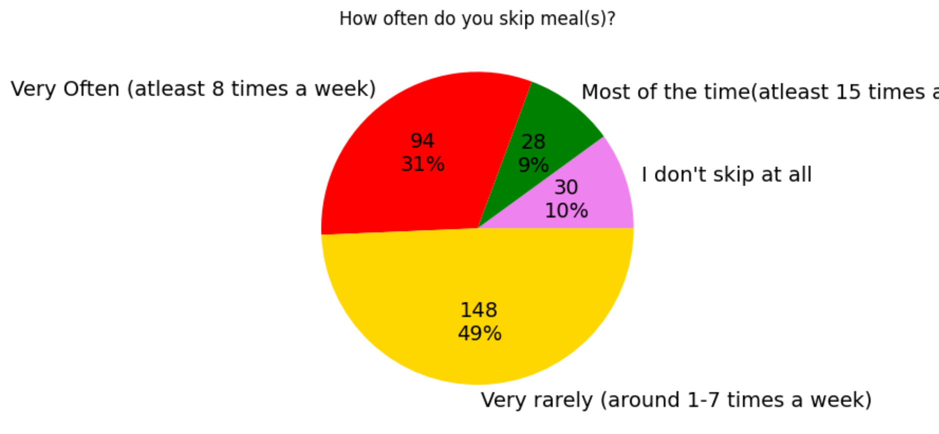
Impact of working days

Figure 18: Meal skip based on part of the week



Number of times mess is skipped

Figure 19: Range frequency mess is skipped



Confidence Interval Estimation

Case-1: Confidence Interval of μ, where μ is the mean number of meals skipped by a student at IITH in a week.

Consider a random sample of size $n = 50$. Sample mean and sample standard deviation are $\bar{x}$ and S respectively, where

$$\bar{x} = 6.98, S = 4.18$$

Also, for a 95% CI, $\alpha = 0.05, t_{\alpha/2,n-1} = t_{0.025,49} = 2.0096$
Now, the resulting 95% CI is:

$$\bar{x} - t_{\alpha/2,n-1}\left(\frac{S}{\sqrt{n}}\right) \leq \mu \leq \bar{x} + t_{\alpha/2,n-1}\left(\frac{S}{\sqrt{n}}\right)$$
$$5.792 \leq \mu \leq 8.168$$

Thus, based on the sample data, CI of mean number of meals skipped by a student in a week is $[5.792, 8.168]$

Confidence Interval Estimation

Case-2: CI of σ,where σ is the standard deviation of the population.

Consider a random sample of size $n = 50$.

Here sample variance, $S^2 = 17.49$

Take, significance level $\alpha = 0.05$

$$a = \chi^2_{1-\alpha/2,n-1} = \chi^2_{0.975,49} = 67.505$$
$$b = \chi^2_{\alpha/2,n-1} = \chi^2_{0.025,49} = 30.096$$

Now, the resulting 95% CI is:

$$\frac{(n-1)S^2}{b} \leq \sigma^2 \leq \frac{(n-1)S^2}{a}$$
$$12.695 \leq \sigma^2 \leq 28.476$$

Thus,the obtained CI of σ^2 is $[12.695, 28.476]$

This leads to 95% CI for σ :$[3.563, 5.336]$

Hypothesis Testing

Case-1: Verifying if the mean count of students who skip mess during week-days is greater than the mean count of the students who skip mess during week-ends with level of significance 0.05.
Here,

$$n_1 = 216 \quad n_2 = 54$$

As both the sample sizes are greater than 30, The condition that population distributions are normal with equal variances is satisfied.
Let's now declare the Null and Alternate Hypothesis

$$H_0 : \mu_1 - \mu_2 \leq 0 \text{ vs } H_a : \mu_1 - \mu_2 > 0$$
$$\bar{x_1} = 7.643 \quad \bar{x_2} = 5.907$$
$${S_1}^2 = 23.396 \quad {S_2}^2 = 15.935$$

Here, $\frac{{S_1}^2}{{S_2}^2} = 1.468$ which is less than 4. So, we can assume that both the variances are equal.

Hypothesis Testing

Let's now calculate the degrees of freedom and pooled variance:

$$df = n_1 + n_2 - 2 = 268$$
$${S_p}^2 = \frac{(n_1 - 1){S_1}^2 + (n_2 - 1){S_2}^2}{df} = 21.92$$

Test Statistic: :

$$t = \frac{(\bar{x_1} - \bar{x_2}) - D_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n2}}} = 0.5205$$

Rejection Region Approach : Reject H_o if $t \geq 1.6506$ where $t_{0.05,268} = 1.6506$ (here considered $\alpha = 0.05$).
Result: Because the observed value of t = 0.5205 is less than 1.6506 and hence is not in the rejection region, there is insufficient evidence to conclude that mean count of students who skip mess during week-days is greater than the mean count of the students who skip mess during week-ends.

Hypothesis Testing

Case-2: Verifying if the mean count of Undergraduate students who skip mess is different from 6 with 5% as level of significance.

Here, sample size $n = 127 > 30$, So we can assume that the population distribution is normal.

Now, we set up the research hypotheses :

$$H_0 : \mu = 6 \quad \text{vs} \quad H_a : \mu \neq 6$$
$$\bar{x} = 7.285$$
$$S = 4.572$$

Test Statistic :

$$t^* = \frac{\bar{x} - \mu_0}{S/\sqrt{n}} = \frac{7.285 - 6}{4.572/\sqrt{127}} = 3.1673$$
$$df = 127 - 1 = 126 \text{ and } \alpha = 0.05$$

Hypothesis Testing

p-value Approach: Since H_a is two-tailed,

$$p - value = 2 \times P(t > |t^*|) = 2 \times P(t > |3.1673|)$$

Since we do not find the exact value of 3.1673 in t-table at 126 d.f., we try to find a range. It can be seen from the t-table that the value falls between 3.1562 and 3.1892, and corresponding to them the right tail probabilities are 0.001 and 0.0009 respectively.
$\therefore$ The p-value would be between $2 \times (0.0009) = 0.0018$ and $2 \times (0.001) = 0.002$.
Result: Since p-value is less than α, there is significant evidence that the mean count of Undergraduate students who skip mess is different from 7.

Hypothesis Testing

Case-3: Verifying if the variability in count of students of age 19-20 years who skip mess is less than 7 with 0.05 as level of significance.
Here,

$$H_0 : \sigma^2 \geq 7 \text{ vs } H_a : \sigma^2 < 7$$
$$n = 124 > 30$$
$$S^2 = 16.497$$

Test Statistic:

$$\chi^2 = \frac{(n-1)S^2}{{\sigma_0}^2} = \frac{123 \times 16.497}{49} = 41.4108$$

Rejection region Approach: Reject H_0 if the value of TS is less than 149.8846, for df = n-1 = 123 and $1 - \alpha = 0.95$.
Result: Since the computed value 41.4108 is less than the critical value of 149.8846, there is sufficient evidence to reject H_0 i.e., there is significant evidence that the variability in count of students of age 19-20 years who skip mess is less than 7.

Hypothesis Testing

Case-4: Verifying if the percentage of AI and CSE students who skip mess is greater than 0.1 with level of significance 0.05.

Here,

$$H_0 : \pi \leq 0.1 \text{ vs } H_a : \pi > 0.1$$
$$n = 41 > 30$$

Test Statistic :

$$Z = \frac{\hat{\pi} - \pi_0}{\sigma_{\hat{\pi}}}$$

From the Survery data,

$\hat{\pi} = \frac{41}{270} = 0.1518$ and $\sigma_{\hat{\pi}} = \sqrt{\frac{0.1518(1-0.1518)}{270}} = 0.0218$

Also,

$$n(\pi_0) = 270(0.1518) = 40.986 > 5 \text{ and}$$
$$n(1 - \pi_0) = 270(1 - 0.155) = 229.014 > 5.$$

Hypothesis Testing

Thus, the sample considered is valid and we obtain :

$$Z = \frac{\hat{\pi} - \pi_0}{\sigma_{\hat{\pi}}} = \frac{0.1518 - 0.1}{0.0218} = 2.3761$$

Rejection Region Approach: Reject H_0 if $Z > 1.645$ for $\alpha = 0.05$
Result: Since the observed value of Z exceeds the critical value of 1.645, we conclude there is significant evidence that the percentage of AI and CSE department students who skip mess exceeds the percentage of 10%.

THANK YOU

May 2, 2023