Food Stastistics of College Students

MA4240 - Applied Statistics

Contents

1 Introduction **3**

2 Preprocessing of data **4**

3 Data Analysis **4**
3.1 Observing Amount spent by each person 4
3.2 Observing the Level of Adventurousness and its Proportion of Data 5
3.3 observing dineout frequencies per week. 8

4 Confidence Intervals **8**
4.1 Type of Diet 8
4.1.1 Confidence Interval for Ratio of vegetarians 9
4.1.2 Confidence Interval for Ratio of vegetarians 9
4.2 Taste Preferences 10
4.2.1 Confidence Intervals 10
4.3 Cuisine preferences 11
4.3.1 Confidence Intervals 11
4.4 Dessert preferences 12
4.4.1 Confidence Intervals 12
4.5 Cost statistics : Confidence Intervals for the amount spent by different adventurous level people 12
4.5.1 Confidence Intervals for the amount spent by low adventurous level people 13
4.5.2 Confidence Intervals for amount spent by medium adventurous level people 14
4.5.3 Confidence Intervals for amount spent by high adventurous level people 14
4.5.4 Confidence Intervals for amount spent by very high adventurous level people 15
4.6 Confidence Interval for the number of days the students dine out per week 16

5 Hypothesis Testing: **16**
5.1 Hypothesis -1 16
5.2 Hypothesis -2 18
5.3 Hypothesis -3 19
5.4 Hypothesis -4 21

6 Contributors **22**

1 Introduction

Overview The project tries to estimate the number of people based on many subparts that could be obtained from questions like the proportion of vegetarians and nonvegetarians in campus, Cuisine preference ratios, etc.
It also tries to estimate money spent by people on campus and some factors that might affect the money spent like their ability to experiment with new foods, number of dine-out days, etc.
On an overlook of this, we try to estimate population ratios for each type in the confidence intervals part while trying to test some hypotheses that seem general and we try to estimate where the hypotheses are true or not .

Survey Questions These are the survey questions that were used to collect the data for project

1. Gender
2. Type
3. Place of Origin
4. Which type of taste do you generally prefer in your food?
5. Which Indian cuisine do you like the most?
6. Which cuisine do you like to eat among foreign cuisines?
7. How adventurous are you when it comes to trying new foods? (On a scale of 0 to 10)
8. How often do you dine out in a typical week?
9. On average, how much do you spend on food per month, considering dining out and other related expenses?
10. Do you prefer having dessert after a meal, and if so, what type of dessert do you typically enjoy?

2 Preprocessing of data

The following are the steps taken for the preprocessing of data:

- We started by removing white spaces and some symbols like the rupee symbol from the cost.
- we one hot encoded multiple options like the taste and foreign cuisine preferences for further calculations. eg:[Spicy, Umami] - [1,0,0,0,1]
- From collected as northeast Indians are fewer we classified them into north India

3 Data Analysis

3.1 Observing Amount spent by each person

In our data, we collected about the amount and various preferences. We talk about the amount spent by them and the population preferences

Here are the statistics about the amount spent:

Table 1: Descriptive Statistics for amount spent

Statistic	**Value**
Count	223
Mean	2350.39
Standard Deviation	1561.59
Minimum	529
25th Percentile	1123
Median	2058
75th Percentile	2909.5
Maximum	6951

It is left skewed gaussian.

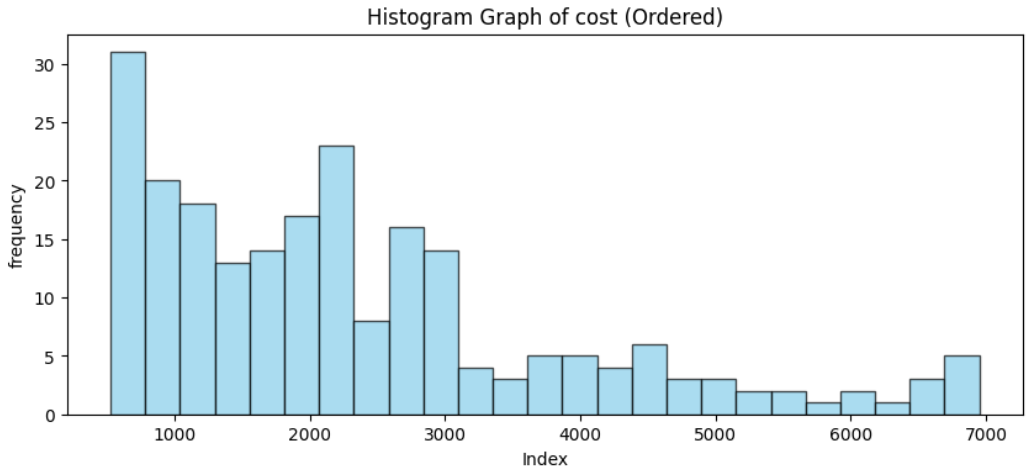

Figure 1: Histogram for amount spent

3.2 Observing the Level of Adventurousness and its Proportion of Data

By observing the graphs we can infer the number of people and their tendency to experiment with food .

Table 2: Descriptive Statistics for level of adventurous

Statistic	**Value**
Count	223
Mean	6.14
Standard Deviation	2.61
Minimum	0
25th Percentile	5
Median	6
75th Percentile	8
Maximum	10

This is later used to estimate money spent on food and their respective level of adventurousness

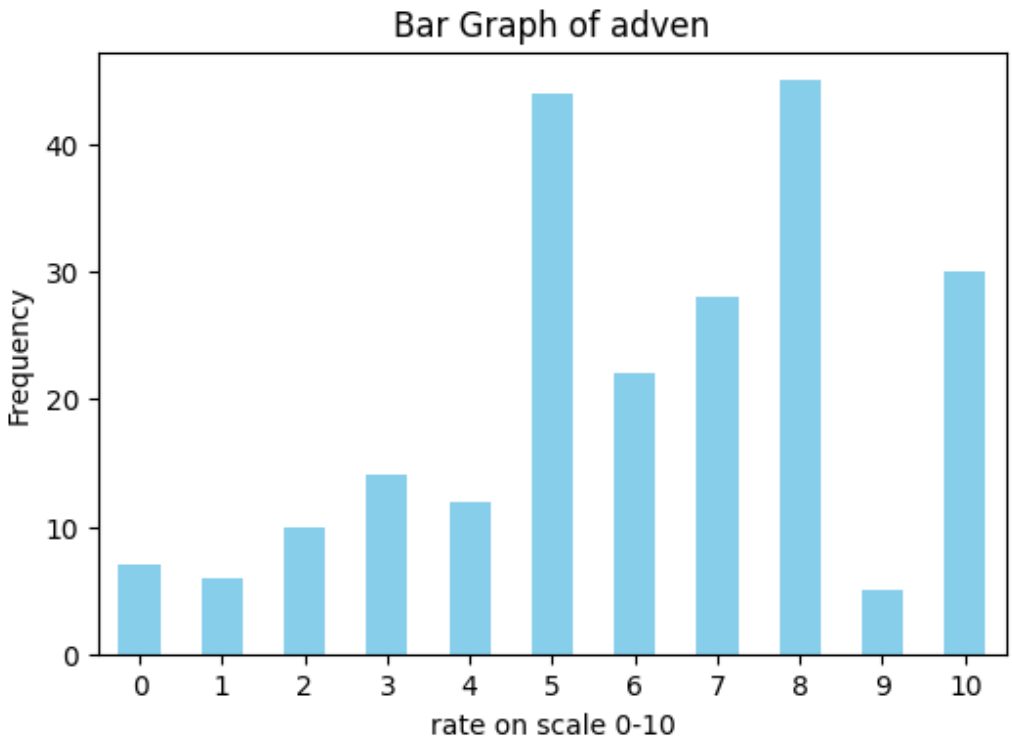

Figure 2: Bar Graph on Adventurous Level

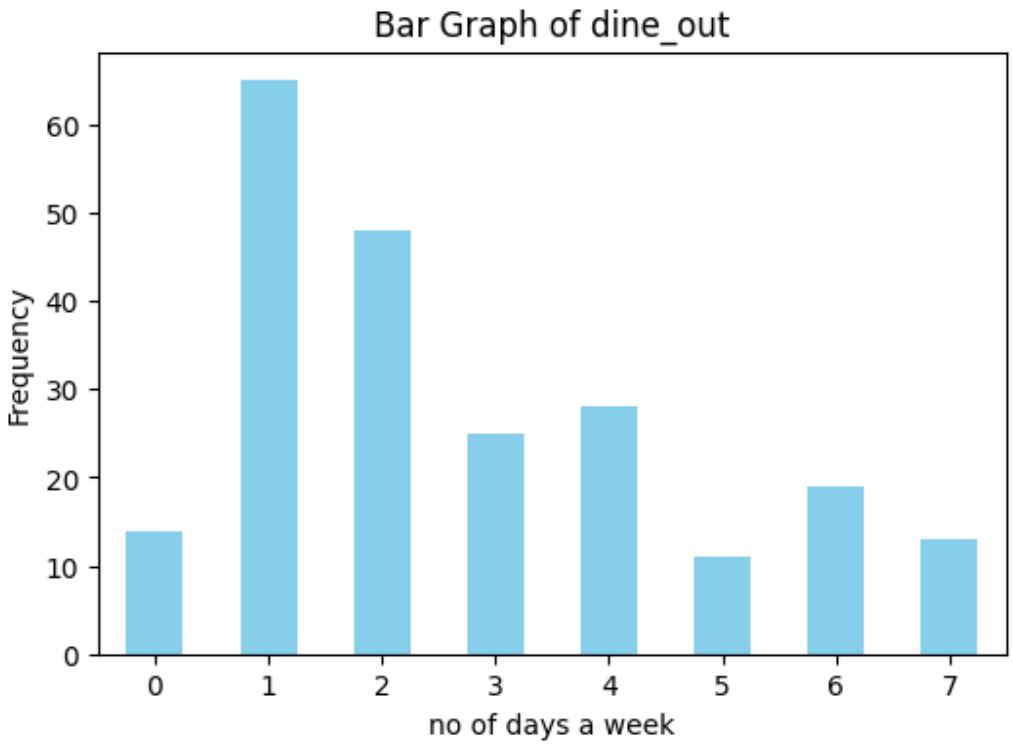

Figure 3: Bar Graph on Dine Out

3.3 observing dineout frequencies per week.

Statistics obsvered for this:

Table 3: Descriptive Statistics for dine-out frequency

Statistic	**Value**
Count	223
Mean	2.73
Standard Deviation	1.98
Minimum	0
25th Percentile	1
Median	2
75th Percentile	4
Maximum	7

This is later used to estimate average number of days average person of campus eats out.

4 Confidence Intervals

4.1 Type of Diet

This section speaks about the proportions of our survivors that prefer vegetarian and non-vegetarian.

The following are stats related to it.

Let π_1 denote the proportion of Vegetarians and π_2 denote the proportion of Non-Vegetarians.

$$\text{No. of Vegetarians, } n_{\text{V}} = 79 \tag{1}$$

$$\text{No. of Non-Vegetarians, } n_{\text{NV}} = 144 \tag{2}$$

$$\hat{\pi}_1 = \frac{n_{\text{V}}}{n_{\text{V}} + n_{\text{NV}}} = 0.35, \quad \hat{\pi}_2 = \frac{n_{\text{NV}}}{n_{\text{V}} + n_{\text{NV}}} = 0.65 \tag{3}$$

4.1.1 Confidence Interval for Ratio of vegetarians

The following formula can be used to calculate the confidence interval of the fraction of vegetarians

$$\hat{\pi}_1 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n}} \tag{4}$$

By substituting, we get,

$$0.35 \pm 1.96 \times \sqrt{\frac{0.35(1-0.35)}{223}} \tag{5}$$

i.e.,

$$0.35 \pm 0.062 = (0.291, 0.417)$$

Thus, we can be 95% confident that between the ratio of 0.291 and 0.417 of the population on the campus prefers vegetarianism.

4.1.2 Confidence Interval for Ratio of vegetarians

Similarly, by using the same formula, we can calculate the CI for the fraction of non-vegetarians as

$$\hat{\pi}_2 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n}} \tag{6}$$

Substituting the values, we get

$$0.65 \pm 1.96 \times \sqrt{\frac{0.65(1-0.65)}{223}} \tag{7}$$

i.e.,

$$0.65 \pm 0.062 = (0.582, 0.708)$$

Thus, we can be 95% confident that between the ratio of 0.582 and 0.708 the population on the campus prefer Non-vegetarian.

4.2 Taste Preferences

This section speaks about the proportions of our survey related to their taste preferences

The following are stats related to it.

Let π_1 denote the proportion of people who prefer sweet and π_2 denote the proportion of people who prefer salt...

$$\text{No. of that prefer sweet taste, } n_1 = 93 \tag{8}$$
$$\text{No. of that prefer salty taste, } n_2 = 54 \tag{9}$$
$$\text{No. of that prefer sour taste, } n_3 = 44 \tag{10}$$
$$\text{No. of that prefer spicy taste, } n_4 = 176 \tag{11}$$
$$\text{No. of that prefer umami taste, } n_5 = 21 \tag{12}$$
$$\text{No. of people, } n_{\mathrm{T}} = 223 \tag{13}$$

So,

$$\hat{\pi}_1 = 0.42, \quad \hat{\pi}_2 = 0.25, \quad \hat{\pi}_3 = 0.2, \quad \hat{\pi}_4 = 0.79, \quad \hat{\pi}_5 = 0.09 \tag{14}$$

Now, For large random samples, a $100(1-\alpha)\%$ confidence interval for population proportion p_i is:

$$\hat{p}i \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{n}} \quad \text{for } i = 1, 2, 3, 4$$

4.2.1 Confidence Intervals

On applying the above formula we get the following confidence intervals

- Thus, we can be 95% confident that between the ratio of 0.352 and 0.482, the population on the campus prefers sweet flavour.
- Thus, we can be 95% confident that between the ratio of 0.19 and 0.303, the population on the campus prefers Salty flavor.
- Thus, we can be 95% confident that between the ratio of 0.145 and 0.303, the population on the campus prefers Sour flavor.
- Thus, we can be 95% confident that between the ratio of 0.736 and 0.843, the population on the campus prefers Spicy flavor.
- Thus, we can be 95% confident that between the ratio of 0.056 and 0.133, the population on the campus prefers Umami Flavor.

4.3 Cuisine preferences

This section speaks about surveyors and their foreign cuisine preferences.

The following are stats related to it.

Let π_1 denote the proportion of people who prefer sweet and π_2 denote the proportion of people who prefer salt...

$$\text{No. of that prefer Japanese cuisine, } n_1 = 63 \tag{15}$$
$$\text{No. of that prefer Italian cuisine, } n_2 = 115 \tag{16}$$
$$\text{No. of that prefer Chinese cuisine, } n_3 = 135 \tag{17}$$
$$\text{No. of that prefer Mexican cuisine, } n_4 = 60 \tag{18}$$
$$\text{No. of people, } n_{\mathrm{T}} = 223 \tag{19}$$

So,

$$\hat{\pi}_1 = 0.28, \quad \hat{\pi}_2 = 0.52, \quad \hat{\pi}_3 = 0.60, \quad \hat{\pi}_4 = 0.27 \tag{20}$$

Now, For large random samples, a $100(1-\alpha)\%$ confidence interval for population proportion p_i is:

$$\hat{p}i \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{n}} \quad \text{for } i = 1, 2, 3, 4$$

4.3.1 Confidence Intervals

On applying the above formula we get the following confidence intervals

- Thus, we can be 95% confident that between the ratio of 0.223 and 0.342, the population on the campus prefers Japanese cuisine.
- Thus, we can be 95% confident that between the ratio of 0.45 and 0.581, the population on the campus prefers Italian cuisine.
- Thus, we can be 95% confident that between the ratio of 0.541 and 0.67, the population on the campus prefers Chinese cuisine.
- Thus, we can be 95% confident that between the ratio of 0.211 and 0.327, the population on the campus prefers Mexican cuisine.

4.4 Dessert preferences

This subsection speaks about our surveyors and Their frequency of desert consumption

The following are stats related to it.

Let π_1 denote the proportion of people who prefer sweet and π_2 denote the proportion of people who prefer salt...

$$\text{No. of people that prefer dessert often , } n_1 = 81 \tag{21}$$

$$\text{No. of people that prefer dessert rarely , } n_2 = 116 \tag{22}$$

$$\text{No. of people that do not prefer dessert, } n_3 = 26 \tag{23}$$

$$\text{No. of people, } n_{\mathrm{T}} = 223 \tag{24}$$

So,

$$\hat{\pi}_1 = 0.36, \quad \hat{\pi}_2 = 0.52, \quad \hat{\pi}_3 = 0.12 \tag{25}$$

Now, For large random samples, a $100(1-\alpha)\%$ confidence interval for population proportion p_i is:

$$\hat{p}i \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{n}} \quad \text{for } i = 1, 2, 3, 4$$

4.4.1 Confidence Intervals

On applying the above formula we get the following confidence intervals

- We can be 95% confident that between the ratio of **0.3 and 0.426**, the population on the campus prefers Dessert **Often**.
- We can be 95% confident that between the ratio of **0.455 and 0.586**, the population on the campus prefers Dessert **rarely**.
- Thus, we can be 95% confident that between the ratio of **0.074 and 0.159**, the population on the campus **does not prefer** having Dessert.

4.5 Cost statistics : Confidence Intervals for the amount spent by different adventurous level people

In this section, we examine expenditure patterns across different levels of adventurousness among individuals. Confidence intervals were calculated to estimate mean expenditures for each category

If $X_1, X_2, \ldots, X_n$ are normally distributed with unknown mean μ and variance σ^2, then a $(1-\alpha) \times 100\%$ confidence interval for the population mean μ is:

$$\bar{x} \pm t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}}$$

4.5.1 Confidence Intervals for the amount spent by low adventurous level people

The central tendencies are as follows:

- **Count** n:23.000000.
- **Mean** $\bar{x}$: 2898.782609.
- **Std** S: 1967.060519
- **Min**: 564.000000.
- **25%**: 1152.000000.
- **50%**: 2267.000000.
- **75%**: 4731.000000.
- **Max**: 6627.000000.

Using the following formula we get CI with confidence 95% as

$$\bar{x} \pm t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}} = 2898.78 \pm 2.07 \frac{1967.06}{\sqrt{23}} = (2048.161, 3749.403) \tag{26}$$

From the respective calculations, we can see that:

The mean of money spent by people that fall into the category of **low adventurous level** will lie in the interval **(2048.161, 3749.403)** with a confidence of **95%**.

4.5.2 Confidence Intervals for amount spent by medium adventurous level people

The central tendencies are as follows:

- **Count** n:70.000000.
- **Mean** $\bar{x}$: 2283.928571.
- **Std** S: 1430.839931
- **Min**: 587.000000.
- **25%**: 1141.500000.
- **50%**: 2056.000000.
- **75%**: 2692.000000.
- **Max**: 6890.000000.

Using the following formula we get CI with confidence 95% as

$$\bar{x} \pm t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}} = 2283.92 \pm 1.99\frac{1430.83}{\sqrt{70}} = (1942.757, 2625.100) \tag{27}$$

From the respective calculations, we can see that:

The mean of money spent by people that fall into the category of **mid-adventurous level** will lie in the interval **(1942.757, 2625.100)** with a confidence of **95%**.

4.5.3 Confidence Intervals for amount spent by high adventurous level people

The central tendencies are as follows:

- **Count** n:95.000000.
- **Mean** $\bar{x}$: 2271.831579.
- **Std** S: 1435.554846
- **Min**: 513.000000.

- **25%**: 1086.000000.
- **50%**: 1935.000000.
- **75%**: 2963.000000.
- **Max**: 6900.000000.

Using the following formula, we get CI with confidence 95% as

$$\bar{x} \pm t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}} = 2271.83 \pm 1.985\frac{1435.55}{\sqrt{95}} = (1979.394, 2564.269) \quad (28)$$

From the respective calculations, we can see that:

The mean of money spent by people that fall into the category of **high adventurous level** will lie in the interval **(1979.394, 2564.269)** with a confidence of **95%**.

4.5.4 Confidence Intervals for amount spent by very high adventurous level people

The central tendencies are as follows:

- **Count** n:35.000000
- **Mean** $\bar{x}$: 2191.400000.
- **Std** S: 1397.002931
- **Min**: 549.000000.
- **25%**: 934.500000.
- **50%**: 2161.000000.
- **75%**: 2638.000000.
- **Max**: 5263.000000.

Using the following formula, we get CI with confidence 95% as

$$\bar{x} \pm t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}} = 2191.40 \pm 2.032\frac{1397.0}{\sqrt{35}} = (1722.545, 2660.255) \quad (29)$$

From the respective calculations, we can see that:

The mean of money spent by people that fall into the category of **very high adventurous level** will lie in the interval **(1722.545, 2660.255)** with a confidence of **95%**.

4.6 Confidence Interval for the number of days the students dine out per week

This part studies the number of days the students of our campus dine out per week.

- **Count** n:223.000000
- **Mean** $\bar{x}$: 2.726457.
- **Std** S: 1.977707
- **Min**: 0.000000.
- **25%**: 1.000000.
- **50%**: 2.000000.
- **75%**: 4.000000.
- **Max**: 7.000000.

Using the following formula, we get CI with confidence 95% as

$$\bar{x} \pm t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}} = 2.73 \pm 1.970\frac{1.977}{\sqrt{223}} = (2.465, 2.987) \tag{30}$$

Thus, the average number of days of dining out of our campus students falls in the interval (2.465,2.987) with a confidence of 95%

5 Hypothesis Testing:

5.1 Hypothesis -1

Hypothesis: The proportion of vegetarians who prefer dessert after meals is higher than the proportion of non-vegetarians who prefer dessert after meals

Let π_1 denote the proportion of Vegetarians preferring dessert after meals and π_2 denote the proportion of Non-Vegetarians preferring dessert after meals.

$$\text{No. of Vegetarians in sampled data, } n_{\text{V}} = 49 \tag{31}$$
$$\text{No. of Non-Vegetarians in sampled data, } n_{\text{NV}} = 100 \tag{32}$$
$$\text{No. of Vegetarians preferring dessert, } n_{\text{Vp}} = 23 \tag{33}$$
$$\text{No. of Non-Vegetarians preferring dessert, } n_{\text{NVp}} = 30 \tag{34}$$

So,

$$\hat{\pi}_1 = \frac{n_{\text{Vp}}}{n_{\text{V}}} = 0.47, \quad \hat{\pi}_2 = \frac{n_{\text{NVp}}}{n_{\text{NV}}} = 0.30 \tag{35}$$

Now,

$$H_0 : \pi_1 - \pi_2 \leq 0 \tag{36}$$
$$H_a : \pi_1 - \pi_2 > 0 \tag{37}$$

Now, we check the conditions:

$$n_1\hat{\pi}_1 \geq 5, \quad n_1(1-\hat{\pi}_1) \geq 5 \tag{38}$$
$$n_2\hat{\pi}_2 \geq 5, \quad n_2(1-\hat{\pi}_2) \geq 5 \tag{39}$$

So, the test statistic is

$$Z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}} = 2.00 \tag{40}$$

and

$$z_{0.05} = 1.645 \tag{41}$$

Rejection Region approach: We will reject H_0 if the test statistic $Z > Z_\alpha$. With a significance level $\alpha = 0.05$, $z_{0.05} < Z$. Since $Z > z_{0.05}$, which means it's lying in the rejection region, so we will reject H_0.

Inference: We can infer that, statistically with a significance level of 0.05, The proportion of vegetarians who prefer dessert after meals is higher than the proportion of non-vegetarians who prefer dessert after meals

5.2 Hypothesis -2

Hypothesis: The mean expenditure on food per month by females is greater than the mean expenditure on food per month by males

From the data sampled from whole data, we have mean amount spent by male (x_m) and female (x_f) are as follows,

$$x_m = 2380.65$$

$$x_f = 2593.78$$

Since we have $x_f > x_m$, we will do Hypothesis Testing with,

$$H_0 : \mu_f - \mu_m \leq 0 \tag{42}$$
$$H_a : \mu_f - \mu_m > 0 \tag{43}$$

From the data we got,

$$S_f = 1619.76$$

$$S_m = 1702.65$$

$$\Rightarrow \frac{1}{2} < \frac{S_f}{S_m} \approx 0.95 < 2.0$$

So we assume population variances to be the same and then pooled variance will be equal to ($n_f = 60$, $n_m = 130$),

$$S_p = \sqrt{\frac{49 \times 1619.76 \times 1619.76 + 99 \times 1722.02 \times 1722.02}{148}} \approx= 1647.668 \tag{44}$$

The test statistic will be,

$$t = \frac{(x_f - x_m) - 0}{S_p \sqrt{\frac{1}{n_f} + \frac{1}{n_m}}} \tag{45}$$
$$= \frac{2593.78 - 2380.65}{1647.688 \times 0.173} \tag{46}$$
$$= \frac{213}{285.05} = 0.747 \tag{47}$$

Rejection Region Approach: We will reject H_0 if the test statistic $t > t_{\alpha, n_f + n_m - 2}$ with $\alpha = 0.05$ (Significance Level), $t_{0.05,188} = 1.653$. The

observed test statistic (0.747) is less than 1.655, hence it isn't in the rejection region.

Inference: We can infer that, statistically with a significance level of 0.05, The mean expenditure on food per month by females may not be greater than the mean expenditure on food per month by males

5.3 Hypothesis -3

Hypothesis The proportion of South Indians who prefer their regional cuisine is higher than the proportion of North Indians who prefer their regional cuisine

Let π_1 denote the proportion of South Indian people preferring South cuisine and π_2 denote the proportion of North Indian people preferring North cuisine.

$$\text{No. of South Indians in sampled data, } n_{\text{S}} = 100 \tag{48}$$
$$\text{No. of North Indians in sampled data, } n_{\text{N}} = 50 \tag{49}$$
$$\text{No. of South Indians preferring South cuisine, } n_{\text{Sp}} = 83 \tag{50}$$
$$\text{No. of North Indians preferring North cuisine, } n_{\text{Np}} = 24 \tag{51}$$

So,

$$\hat{\pi}_1 = \frac{n_{\text{Sp}}}{n_{\text{S}}} = 0.83, \quad \hat{\pi}_2 = \frac{n_{\text{Np}}}{n_{\text{N}}} = 0.48 \tag{52}$$

Now,

$$H_0 : \pi_1 - \pi_2 \leq 0 \tag{53}$$
$$H_a : \pi_1 - \pi_2 > 0 \tag{54}$$

Now, we check the conditions:

$$n_1\hat{\pi}_1 \geq 5, \quad n_1(1 - \hat{\pi}_1) \geq 5 \tag{55}$$
$$n_2\hat{\pi}_2 \geq 5, \quad n_2(1 - \hat{\pi}_2) \geq 5 \tag{56}$$

So, the test statistic is

$$Z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}} = 4.38 \tag{57}$$

and

$$z_{0.05} = 1.645 \tag{58}$$

Rejection Region approach: We will reject H_0 if the test statistic $Z > Z_\alpha$. With a significance level $\alpha = 0.05$, $z_{0.05} < Z$. Since $Z > z_{0.05}$, which means it's lying in the rejection region, so we will reject H_0.

Inference: We can infer that, statistically with a significance level of 0.05, The proportion of South Indians who prefer their regional cuisine is higher than the proportion of North Indians who prefer their regional cuisine

5.4 Hypothesis -4

Hypothesis The number of people who prefer Japanese cuisine is greater than the number of people who do not prefer it

From the data collected,

$$\text{No. of People preferring Japanese Cuisine in sampled data, } n_{\text{J}} = 28 \tag{59}$$

$$\text{Total people in sampled data, } n_{\text{n}} = 100 \tag{60}$$

$$\tag{61}$$

The estimated proportion is,

$$\hat{\pi} = \frac{n_{\text{J}}}{n_{\text{n}}} = 0.28 \tag{62}$$

Let π denote the proportion of people preferring Japanese Cuisine. Here,

$$H_0 : \pi \leq 0.5 \quad \text{vs.} \quad H_a : \pi > 0.5 \tag{63}$$

Test statistic:

$$Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.28 - 0.500}{\sqrt{\frac{0.28 \times 0.72}{100}}} = -4.89 \tag{64}$$

p-value approach:

$$p = P(z > Z) = P(z > -4.89) = 0.9998 \tag{65}$$

With significance level $\alpha = 0.05$, $\alpha < p$
so we cannot reject H_0.

Inference: There is no enough evidence to conclude that people generally prefer to have Japanese Cuisine.

6 Contributors

- Manpurwar Ganesh - AI22BTECH11017
- K D V S Aditya - AI22BTECH11013
- Ch Kushwanth - AI22BTECH11006
- T Keshavardhan - AI22BTECH11029
- S Sai Satwik - AI22BTECH11025