

Food Statistics of College Students

MA4240 - Applied Statistics

Contributors: Chitipotu Kushwanth K.D.V.S Aditya
Tera Keshavardhan Reddy Manpurwar Ganesh Sai Satwik

May 2, 2024

Presentation Overview

- Survey Questions

1 Food statistics

- Type of diet
- Taste preferences
- Cuisine preferences
- Dessert preferences

2 Cost statistics

- Introduction
- Level of Adventurousness
- Dine out

3 Hypothesis Testing

- Hypothesis - 1
- Hypothesis - 2
- Hypothesis - 3
- Hypothesis - 4

Survey Questions

Survey Questions

- 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?

Food statistics

Type of diet

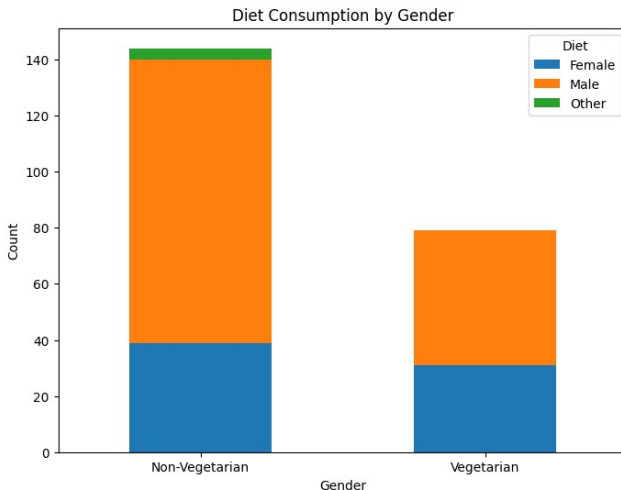


Figure: Diet Consumption by Gender

Type of diet

This subsection speaks about the proportions of our surveyors 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.

$$\text{No. of Vegetarians, } n_V = 79 \quad (1)$$

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

So,

$$\hat{\pi}_1 = \frac{n_V}{n_V + n_{NV}} = 0.35, \quad \hat{\pi}_2 = \frac{n_{NV}}{n_V + n_{NV}} = 0.65 \quad (3)$$

Confidence interval of ratio of vegetarians

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

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

By substituting, we get,

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

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.

Confidence interval of ratio of Non-Veg

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

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

Substituting the values, we get

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

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.

Taste preferences

This subsection 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...

$$\text{No. of that prefer sweet taste, } n_1 = 93 \quad (8)$$

$$\text{No. of that prefer salty taste, } n_2 = 54 \quad (9)$$

$$\text{No. of that prefer sour taste, } n_3 = 44 \quad (10)$$

$$\text{No. of that prefer spicy taste, } n_4 = 176 \quad (11)$$

$$\text{No. of that prefer umami taste, } n_5 = 21 \quad (12)$$

$$\text{No. of people, } n_T = 223 \quad (13)$$

So,

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

Taste preferences

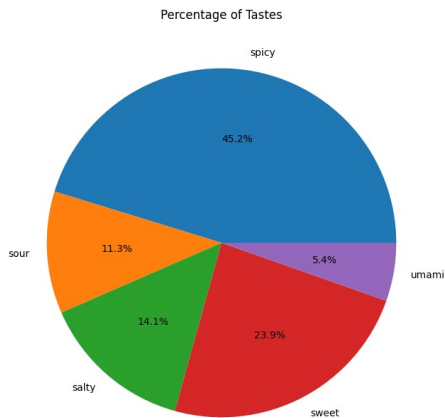


Figure: Distribution of Tastes

Confidence Interval

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

$$\hat{p}_i \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_i(1 - \hat{p}_i)}{n}} \quad \text{for } i = 1, 2, 3, 4, 5$$

The above formula can be used to calculate CI for proportions of people who prefer a particular taste. Applying, It results in the following CIs

Confidence Interval

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

Cuisine preferences

This subsection 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...

$$\text{No. of that prefer Japanese cuisine, } n_1 = 63 \quad (15)$$

$$\text{No. of that prefer Italian cuisine, } n_2 = 115 \quad (16)$$

$$\text{No. of that prefer Chinese cuisine, } n_3 = 135 \quad (17)$$

$$\text{No. of that prefer Mexican cuisine, } n_4 = 60 \quad (18)$$

$$\text{No. of people, } n_T = 223 \quad (19)$$

So,

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

Cuisine preferences

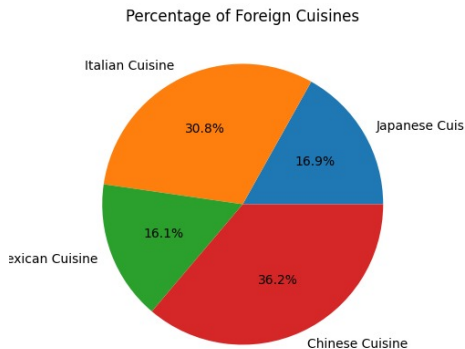


Figure: Percentage Distribution of Cuisine

Confidence Interval

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

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

Using the above formula, we can get CI for their Proportions related to each Cuisine Applying it gives the following outputs:

Confidence Interval

- 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.

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...

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

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

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

$$\text{No. of people, } n_T = 223 \quad (24)$$

So,

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

Dessert preferences

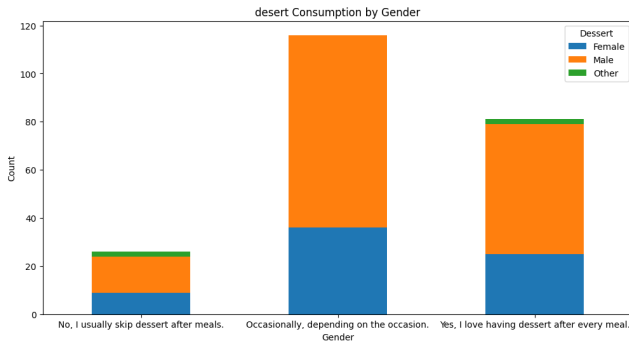


Figure: Dessert preference by gender

Dessert preferences

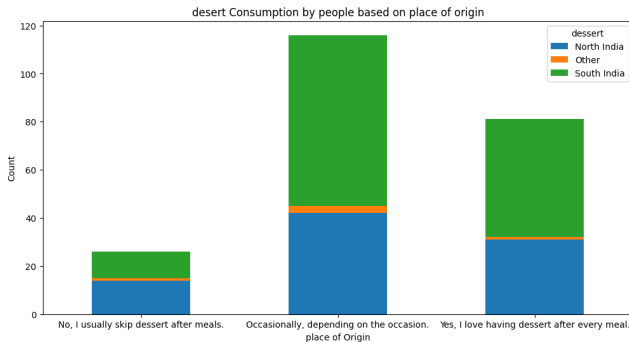


Figure: Dessert preference by place of origin

Confidence Interval

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

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

Using the above formula, we can obtain CI for their frequency of dessert consumption. Applying the formula, we get the following conclusions:

Confidence Interval

- 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.

Cost statistics

Introduction

Introduction

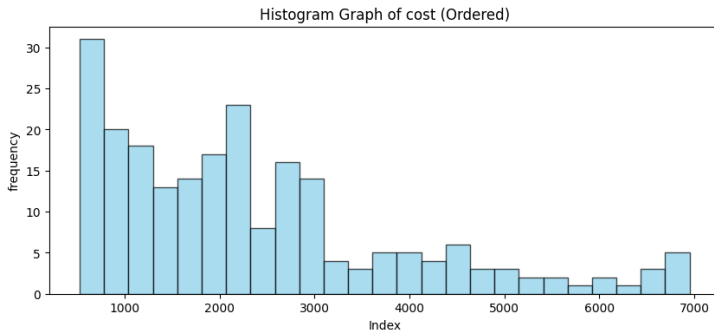


Figure: Histogram for Amount spent

Overview of Spending statistics

The central tendencies are as follows:

- **Count** n:223.000000.
- **Mean** \bar{x} : 2327.668161.
- **Std S**: 1493.292685
- **Min**: 513.000000.
- **25%**: 1086.000000.
- **50%**: 2080.000000.
- **75%**: 2881.500000.
- **Max**: 6900.000000.

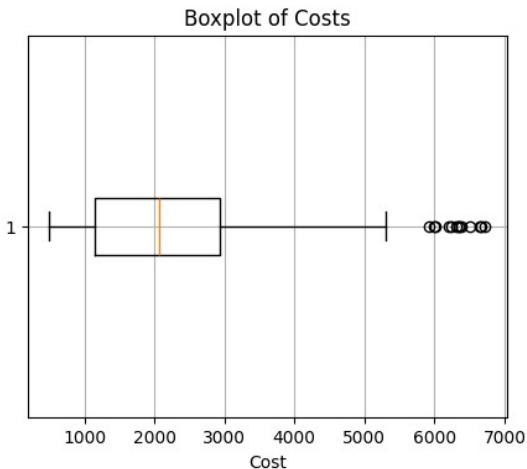


Figure: Box plots for costs

Confidence interval

If X_1, X_2, \dots, 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:

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

Substitution in the above formula,

$$= 2327.67 \pm 1.97 \frac{1493.29}{\sqrt{223}} \quad (26)$$

$$= 2327.67 \pm 196.996$$
$$= (2130.601, 2524.736) \quad (27)$$

Thus, we can say that the mean of money spent by the population of Campus lies in the interval (2130.601, 2524.736)

Level of Adventurousness

This part of the project tries to study how Adventurous students are and tries to find out the relation of this to money spent by them.

Definition

Adventurous in the context of food is their willingness to explore diverse cuisines, flavours, and ingredients, often embracing unconventional or exotic dishes with enthusiasm and curiosity.

Categories

Four categories of people based on their level of adventurousness are **low**, **mid**, **high**, and **very high**.

Statistics

Statistics related to each level are as follows:

Level of Adventurousness (Contd.)

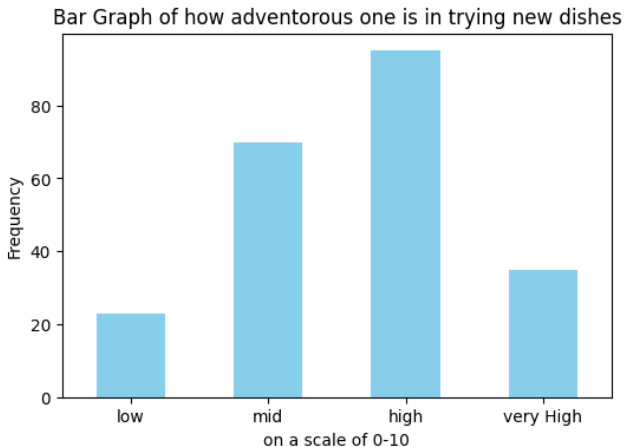


Figure: Adventurousness level grouped into low,mid,high,very high

Level of Adventurousness (Contd.)

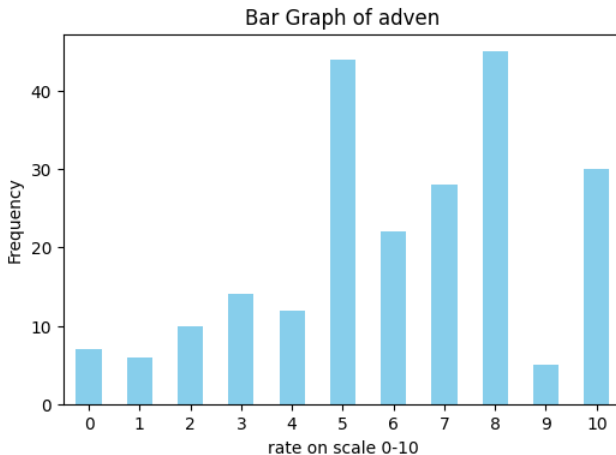


Figure: Adventurousness level on scale 1-10

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

$$\bar{x} \pm t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}} = 2898.78 \pm 2.07 \frac{1967.06}{\sqrt{23}} = (2048.161, 3749.403) \quad (28)$$

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

$$\bar{x} \pm t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}} = 2283.92 \pm 1.99 \frac{1430.83}{\sqrt{70}} = (1942.757, 2625.100) \quad (29)$$

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

$$\bar{x} \pm t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}} = 2271.83 \pm 1.985 \frac{1435.55}{\sqrt{95}} = (1979.394, 2564.269) \quad (30)$$

Very High

- **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

$$\bar{x} \pm t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}} = 2191.40 \pm 2.032 \frac{1397.0}{\sqrt{35}} = (1722.545, 2660.255) \quad (31)$$

Confidence intervals

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%**.
- 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%**.
- 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%**.
- 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%**.

This part studies the number of days our campus students 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.

Dine out (Contn.)

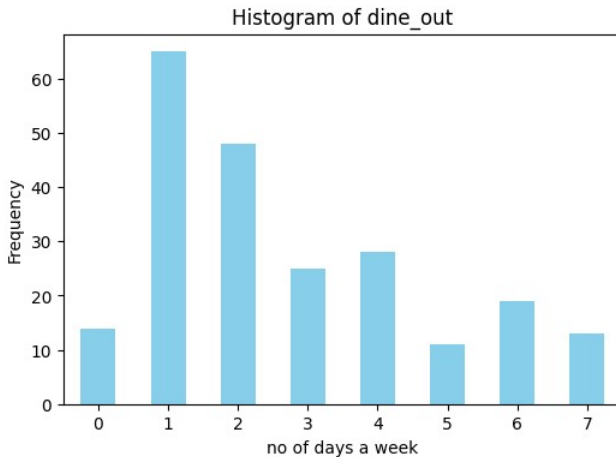


Figure: Histogram of Dine Out

Dine out (Contn.)

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

$$\bar{x} \pm t_{\frac{\alpha}{2}, n-1} \frac{S}{\sqrt{n}} = 2.73 \pm 1.970 \frac{1.977}{\sqrt{223}} = (2.465, 2.987) \quad (32)$$

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%

CIs for difference of two population proportions

For large random samples, an approximate $100(1 - \alpha)\%$ confidence interval for the difference in two population proportions $p_1 - p_2$ is:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}$$

Substituting in the formula, we get

$$(0.35 - 0.40) \pm 1.96 \sqrt{\frac{0.50 \times 0.50}{400} + \frac{0.40 \times 0.60}{250}} \quad (33)$$

which simplifies to

$$0.10 \pm 0.078 = (0.022, 0.178) \quad (34)$$

Hypothesis Testing

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.

$$\text{No. of Vegetarians in sampled data, } n_V = 49 \quad (35)$$

$$\text{No. of Non-Vegetarians in sampled data, } n_{NV} = 100 \quad (36)$$

$$\text{No. of Vegetarians preferring dessert, } n_{Vp} = 23 \quad (37)$$

$$\text{No. of Non-Vegetarians preferring dessert, } n_{NVp} = 30 \quad (38)$$

So,

$$\hat{\pi}_1 = \frac{n_{Vp}}{n_V} = 0.47, \quad \hat{\pi}_2 = \frac{n_{NVp}}{n_{NV}} = 0.30 \quad (39)$$

Hypothesis - 1

Now,

$$H_0 : \pi_1 - \pi_2 \leq 0 \quad (40)$$

$$H_a : \pi_1 - \pi_2 > 0 \quad (41)$$

Now, we check the conditions:

$$n_1 \hat{\pi}_1 \geq 5, \quad n_1(1 - \hat{\pi}_1) \geq 5 \quad (42)$$

$$n_2 \hat{\pi}_2 \geq 5, \quad n_2(1 - \hat{\pi}_2) \geq 5 \quad (43)$$

So, the test statistic is

$$Z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}} = 2.00 \quad (44)$$

and

$$z_{0.05} = 1.645 \quad (45)$$

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

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 the whole data, we have the mean amount spent by males (x_m) and females (x_f) as follows,

$$x_m = 2380.65$$

$$x_f = 2593.78$$

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

Hypothesis - 2

$$H_0 : \mu_f - \mu_m \leq 0 \quad (46)$$

$$H_a : \mu_f - \mu_m > 0 \quad (47)$$

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 ($n_f = 50$, $n_m = 100$),

$$S_p = \sqrt{\frac{99 \times 1702.65 \times 1702.65 + 49 \times 1619.76 \times 1619.76}{148}} \approx 1647.668 \quad (48)$$

Hypothesis - 2

The test statistic will be,

$$t = \frac{(x_f - x_m) - 0}{S_p \sqrt{\frac{1}{n_f} + \frac{1}{n_m}}} \quad (49)$$

$$= \frac{2667.83 - 2259.61}{1647.668 \times 0.173} \quad (50)$$

$$= 0.747 \quad (51)$$

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), $t_{0.05, 188} = 1.655$. 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

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.

$$\text{No. of South Indians in sampled data, } n_S = 100 \quad (52)$$

$$\text{No. of North Indians in sampled data, } n_N = 50 \quad (53)$$

$$\text{No. of South Indians preferring South cuisine, } n_{Sp} = 83 \quad (54)$$

$$\text{No. of North Indians preferring North cuisine, } n_{Np} = 24 \quad (55)$$

So,

$$\hat{\pi}_1 = \frac{n_{Sp}}{n_S} = 0.83, \quad \hat{\pi}_2 = \frac{n_{Np}}{n_N} = 0.48 \quad (56)$$

Hypothesis - 3

Now,

$$H_0 : \pi_1 - \pi_2 \leq 0 \quad (57)$$

$$H_a : \pi_1 - \pi_2 > 0 \quad (58)$$

Now, we check the conditions:

$$n_1 \hat{\pi}_1 \geq 5, \quad n_1(1 - \hat{\pi}_1) \geq 5 \quad (59)$$

$$n_2 \hat{\pi}_2 \geq 5, \quad n_2(1 - \hat{\pi}_2) \geq 5 \quad (60)$$

So, the test statistic is

$$Z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}} = 4.38 \quad (61)$$

and

$$z_{0.05} = 1.645 \quad (62)$$

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

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,

No. of People preferring Japanese Cuisine in sampled data, $n_J = 28$ (63)

Total people in sampled data, $n_n = 100$ (64)

(65)

The estimated proportion is,

$$\hat{\pi} = \frac{n_J}{n_n} = 0.28 \quad (66)$$

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

$$H_0 : \pi \leq 0.5 \quad \text{vs.} \quad H_a : \pi > 0.5 \quad (67)$$

Hypothesis - 4

Test statistic:

$$Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} = \frac{0.28 - 0.500}{\sqrt{\frac{0.28 \times 0.72}{100}}} = -4.89 \quad (68)$$

p-value approach:

$$p = P(z > Z) = P(z > -4.89) = 0.9998 \quad (69)$$

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.

The End