Problem Set 2 - Testing, Proportion Problems

ECO 204 - Statistics for Business and Economics - II, Summer 2025
Faculty: Shaikh Tanvir Hossain, **TA:** Habiba Afroz
Due Date: 16th July, 10:00 PM, 2025

§

Instructions : Please form a group of three (max) and submit on Google Classroom by the due date. Submit the Excel / R file with the calculations and your answers in a single PDF file (handwritten solutions are fine). Please write all group members names and ID numbers on the first page of the PDF file.

Testing for Population Mean

[Note: Use the data file **Student_Lunch.xlsx**. Actually just to do any testing problem, you can use either the **critical value approach** or the **p-value approach** and always both methods will give you the same conclusion, but just to do some maths please do the testing following both methods here so that you have a better understanding of the concepts.]

1. Suppose someone named **Rocky Balboa** visited East West University and had a lunch in the cafeteria. He is worried about the quality of the food here and wants to do some hypothesis testing regarding the amount spent on lunch by the students at the Economics Dept. at EWU. To do this he collected a sample of $n = 64$ students from the Economics Dept. and recorded the amount spent on lunch by each student. The data is provided in the file **Student_Lunch.xlsx**. Assuming the amount spent on lunch is Normally distributed with a population standard deviation of $\sigma = 6$, Rocky wants to do the following hypothesis tests:
   (a) Suppose Rocky wants to test the hypothesis that **the average amount spent for lunch by the students at the Economics Dept. is 20 TK**. Clearly write both the hypotheses and check whether Rocky is correct at 5% significance level. Do the test using both the critical value approach and the p-value approach.
   (b) Now suppose Rocky wants to test the hypothesis that the average amount spent for lunch by the students at the Economics Dept. is at least 20 TK. Now test whether Rocky is correct at 1% significance level. Do the test using both the critical value approach and the p-value approach.
   (c) Finally, suppose Rocky wants to test the hypothesis that the average amount spent for lunch by the students at the Economics Dept. is at most 20 TK. Again test whether Rocky is correct at 10% significance level. Do the test using both the critical value approach and the p-value approach.

2. Continuing from question 1, do the same three tests (in (a), (b), and (c)) as in the previous question, but now assume that the population standard deviation is unknown, however, you can continue to assume that the amount spent on lunch is Normally distributed. Do the test using both the critical value approach and the p-value approach.

3. Finally, do the same three tests as in the previous two questions, but now assume that the amount spent on lunch follows some arbitrary probability distribution (i.e., the population distribution is not known). Again check whether Rocky is correct using the same significance levels as in the previous two questions. Do the test using both the critical value approach and the p-value approach.

Estimation and Testing for Population Proportion

4. Suppose now Rocky collected a dataset from the same students at the Economics Dept. at EWU, and asked them *whether they are satisfied with the quality of food served at the cafeteria*, the answers are recorded as *Yes = 1* or *No = 0*. The data is contained in the file `Student_Satisfaction.xlsx`. Now answer the following questions:
   (a) What is the population of this study?
   (b) If our goal is to find out whether the students in general, not just from the Economics dept. are satisfied with the quality of food served at the cafeteria, is this a good sample? Why or why not?
   (c) Suppose our target parameter is the **proportion of all students at the Economics Dept. who are satisfied with the quality of food served at the cafeteria**, let's call it p, in theory how can we calculate this number?
   (d) Of course it will be difficult to calculate the population proportion p, this is why Rocky collected a random sample of 64 students from the Economics Dept. Now calculate the sample proportion $\overline{p}$ of students who are satisfied with the quality of food served at the cafeteria.
   (e) Based on this sample what is the **point estimate of the population proportion of students who are satisfied** with the quality of food served at the cafeteria?

(f) If we think each data point (or each row) of the sample is a Bernoulli random variable, what is the mean and variance of the Bernoulli random variable? What is the estimated variance of the Bernoulli random variable in this case?

(g) If we assume i.i.d. assumption, this means all rows in the random sample follow the same distribution and they are independent of each other. Then what is the mean and variance of sample proportion $\overline{p}$ when it is a random variable (i.e., $\overline{p}$ is random when we think about repeated sampling)?

(h) What is the standard error of the sample proportion?

(i) Construct a 90% confidence interval for the population proportion of students who are satisfied with the quality of food served at the cafeteria. Interpret the confidence interval.

(j) Now suppose Rocky wants to test the hypothesis that the proportion of students who are satisfied with the quality of food served at the cafeteria is at least 0.7. Test this hypothesis at 5% significance level.

Solution: Question 1

(a) The null and alternative hypotheses are:

$$H_0 : \mu = 20$$
$$H_a : \mu \neq 20$$

To test this hypothesis, we can use the sample mean and standard deviation from the data in `Student_Lunch.xlsx`. Assuming the population standard deviation is known, we can calculate the test statistic using the formula:

$$z_{calc} = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

Here we hve

- $\overline{x} = 21.52$
- $\mu_0 = 20$,
- $\sigma = 6$,
- $n = 64$.

Now we can calculate the test statistic[Check the Excel file for the Calculation]

$$z_{calc} = \frac{21.52 - 20}{6/\sqrt{64}} = 2.03$$

First we will do the calculations using the critical value approach.

Critical Value Approach:

In the critical value approach, the key task is to find the critical value for a two-tailed test at 5% significance level, which are (we give the Excel formula for the critical values):

$$z_{\alpha/2} = z_{0.025} = \texttt{NORM.INV}(0.025, 0, 1) = -1.96$$
$$z_{1-\alpha/2} = z_{0.975} = \texttt{NORM.INV}(0.975, 0, 1) = 1.96$$

Now we know that for the two tailed test, we reject the null hypothesis if $z_{calc} > z_{\alpha/2}$ (this happens when z_{calc} is positive) or $z_{calc} < z_{\alpha/2}$ (this happens when z_{calc} is negative). In this case $z_{calc} = 2.03 > z_{0.025} = 1.96$, so we reject the null hypothesis.

The picture for the two tailed test is as follows:

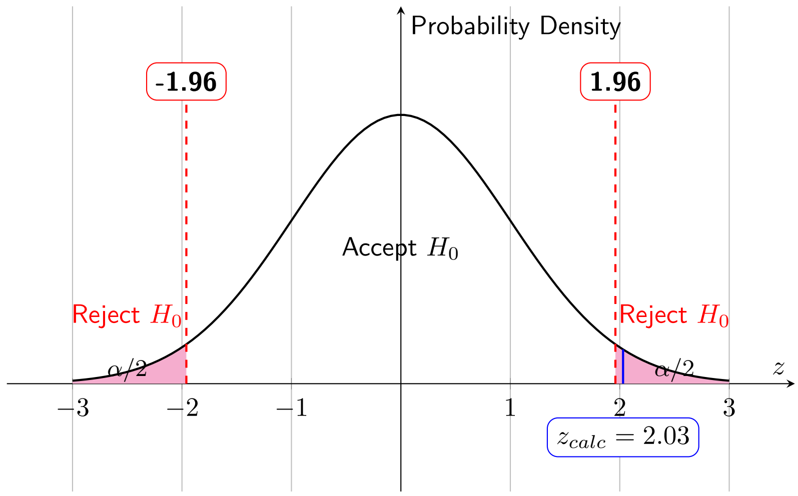

Figure 1: Two-tailed hypothesis test at 5% significance level. The blue line represents the test statistic $z_{calc} = 2.03$, and the red dashed lines are the critical values at $z_{0.025} = -1.96$ and $z_{0.975} = 1.96$. The shaded areas represent the rejection regions for the null hypothesis H_0.

p-value approach: For the p-value approach we can do following calculations

- If $z_{calc} > 0$, then p-value $= 2 \cdot \mathbb{P}(Z > z_{calc})$, which can be calculated in Excel using the formula:

$$p\text{-value} = 2 \cdot (1 - \texttt{NORM.DIST}(z_{calc}, 0, 1, 1))$$

- If $z_{calc} < 0$, then p-value $= 2 \cdot \mathbb{P}(Z < z_{calc})$, which can be calculated in Excel using the formula:

$$p\text{-value} = 2 \cdot \texttt{NORM.DIST}(z_{calc}, 0, 1, 1)$$

- If $z_{calc} = 0$, then p-value $= 1$.

In this case our $z_{calc} = 2.03 > 0$, so we can use the first formula to calculate the p-value:

$$p\text{-value} = 2 \cdot (1 - \texttt{NORM.DIST}(2.03, 0, 1, 1)) = 0.043$$

and in our case we have $\alpha = 0.05$, so p-value $< \alpha$, hence we reject the null hypothesis. This means we have enough evidence to conclude that the average amount spent for lunch by the students at the Economics Dept. is not equal to 20 TK. So Rocky is not correct. Goodbye Rocky!

Shortcut: One issue with the two-tail test is we need to worry about the sign of z_{calc}, i.e., whether it is positive or negative. However, there is a nice shortcut here both for critical value approach and p value approach, these shortcurts are coming from symmetry of the Normal distribution. First we can calculate the absolute value of the test statistic, i.e., $|z_{calc}|$, and then we can use the absolute value to do the calculations.

- For critical value approach we will now check $|z_{calc}| > z_{1-\alpha/2}$ and if this is true, we reject the null hypothesis.
- For p-value approach: we can use the absolute value of the test statistic, with the formula:

$$p\text{-value} = 2 \cdot (1 - \texttt{NORM.DIST}(|z_{calc}|, 0, 1, 1))$$

  and if p-value $< \alpha$, we reject the null hypothesis.

(b) Now suppose Rocky wants to test the hypothesis that the average amount spent for lunch by the students at the Economics Dept. is at least 20 TK. Now test whether Rocky is correct at 1% significance level. Do the test using both the critical value approach and the p-value approach. In this case the null and alternative hypotheses are:

$$H_0 : \mu \geq 20$$
$$H_a : \mu < 20$$

Again we can use the same sample mean and standard deviation from the data in `Student_Lunch.xlsx`. Assuming the population standard deviation is known, we can calculate the test statistic using the formula:

$$z_{calc} = \frac{\overline{x} - \mu_0}{\sigma / \sqrt{n}}$$

Here entire calculation of the test statistic is same as in the previous part, so we have

$$z_{calc} = \frac{21.52 - 20}{6/\sqrt{64}} = 2.03$$

Critical Value Approach:

In the critical value approach, in this case, the key task is to find the critical value z_α, which can be calculated as follows,

$$z_\alpha = z_{0.01} = \texttt{NORM.INV}(0.01, 0, 1) = -2.33$$

Now we know that for the lower tail test, we reject the null hypothesis if $z_{calc} < z_\alpha$. However in our case, we have $z_{calc} = 2.03 > z_{0.01} = -2.33$, so we fail to reject the null hypothesis or simply we accept the null hypothesis. So Rocky is not correct.

The picture for the one tailed test is as follows:

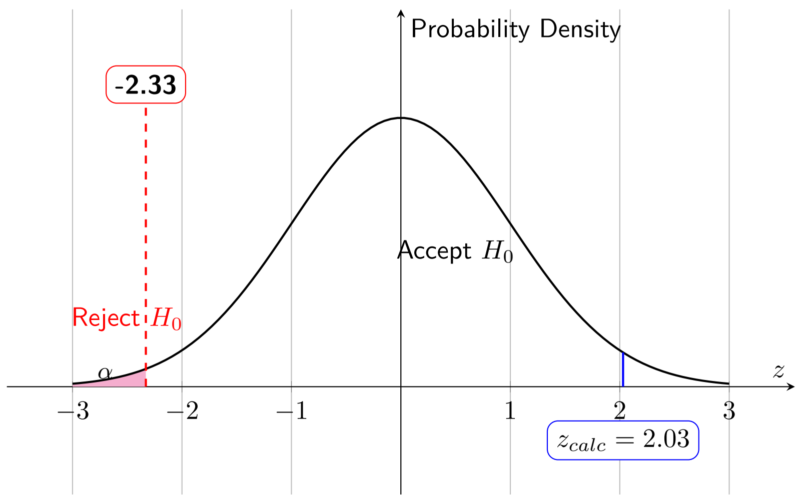

Figure 2: One-tailed hypothesis test at 1% significance level. The blue line represents the test statistic $z_{calc} = 2.03$, and the red dashed line is the critical value at $z_{0.01} = -2.33$. The shaded area represents the rejection region for the null hypothesis H_0.

***p*-Value Approach:**

For the p-value approach since this is a lower tail test, we can do following calculations

- Here p-value $= \mathbb{P}(Z < z_{calc})$, which can be calculated in Excel using the formula:

$$p\text{-value} = \texttt{NORM.DIST}(z_{calc}, 0, 1, 1)$$

In this case we can calculate the p-value as follows:

$$p\text{-value} = \texttt{NORM.DIST}(2.03, 0, 1, 1) = 0.979$$

and in our case we have $\alpha = 0.01$, so p-value $> \alpha$, hence we fail to reject the null hypothesis or simply we accept the null hypothesis.

This means using both approaches we do have enough evidence to conclude that the average amount spent for lunch by the students at the Economics Dept. at least 20 TK. So Rocky is correct here. Welcome Rocky!

(c) Finally, suppose Rocky wants to test the hypothesis that the average amount spent for lunch by the students at the Economics Dept. is at most 20 TK. Again test whether Rocky is correct at 10% significance level. In this case the null and alternative hypotheses are:

$$H_0 : \mu \leq 20$$
$$H_a : \mu > 20$$

Again entire calculation of the test statistic is same as in the previous parts, so we have

$$z_{calc} = \frac{21.52 - 20}{6/\sqrt{64}} = 2.03$$

Critical Value Approach:

In the critical value approach, in this case, the key task is to find the critical value $z_{1-\alpha}$. Here $\alpha = 0.10$, and we can calculate the critical value as follows,

$$z_{1-\alpha} = z_{0.10} = \texttt{NORM.INV}(0.10, 0, 1) = 1.28$$

Now we know that for the upper tail test, we reject the null hypothesis if $z_{calc} > z_{1-\alpha}$. And in our case, we do have have $z_{calc} = 2.03 > z_{0.10} = 1.28$, so we reject the null hypothesis.

The picture for the upper tail test is as follows:

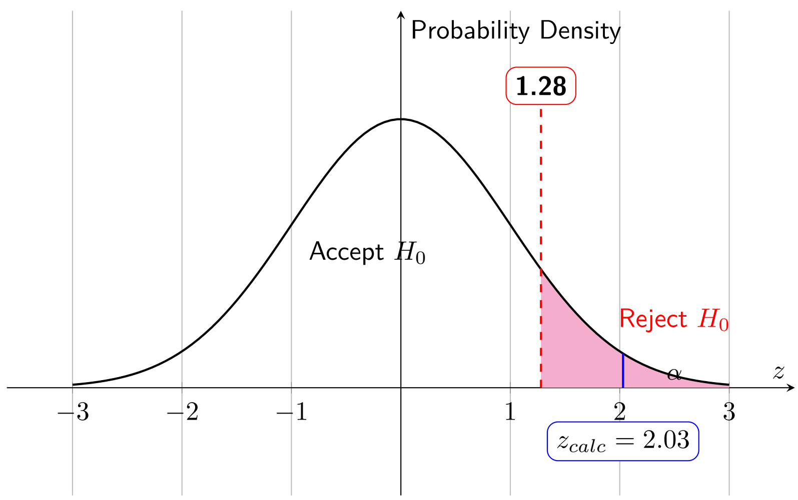

Figure 3: One-tailed hypothesis test at 10% significance level. The blue line represents the test statistic $z_{calc} = 2.03$, and the red dashed line is the critical value at $z_{0.10} = 1.28$. The shaded area represents the rejection region for the null hypothesis H_0.

p-**Value Approach:** For the p-value approach since this is an upper tail test, we can do following calculations

- Here p-value $= \mathbb{P}(Z > z_{calc})$, which can be calculated in Excel using the formula:

$$p\text{-value} = 1 - \texttt{NORM.DIST}(z_{calc}, 0, 1, 1)$$

In this case we can calculate the p-value as follows:

$$p\text{-value} = 1 - \texttt{NORM.DIST}(2.03, 0, 1, 1) = 0.021$$

and in our case we have $\alpha = 0.10$, so p-value $< \alpha$, hence we reject the null hypothesis.

This means using both approaches we do have enough evidence to conclude that the average amount spent for lunch by the students at the Economics Dept. is more than 20 TK. So Rocky is not correct here. Goodbye Rocky!

Solution: Question 2

We now assume that the population standard deviation is unknown, and do the same test using both the critical value approach and the p-value approach.

(a) The null and alternative hypotheses are:

$$H_0 : \mu = 20$$
$$H_a : \mu \neq 20$$

To test this hypothesis, we can use the sample mean and standard deviation from the data in `Student_Lunch.xlsx`. Assuming the population standard deviation is unknown, we can calculate the test statistic using the formula:

$$t_{calc} = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

Here we have

- $\overline{x} = 21.52$
- $\mu_0 = 20$,
- $s = 6.89$ (sample standard deviation), formula in Excel: `=STDEV.S(array)`),
- $n = 64$.

Now we can calculate the test statistic[Check the Excel file for the Calculation]

$$t_{calc} = \frac{21.52 - 20}{6.89/\sqrt{64}} = 1.76$$

Critical Value Approach:

In the critical value approach, the key task is to find the critical value for a two-tailed test at 5% significance level, which can be calculated as follows:

$$t_{\alpha/2} = t_{0.025,63} = \texttt{T.INV}(0.05, 63) = -1.99$$
$$t_{1-\alpha/2} = t_{0.975,63} = \texttt{T.INV}(0.95, 63) = 1.99$$

Now we know that for the two-tailed test, we reject the null hypothesis if $t_{calc} \geq t_{1-\alpha/2}$ (this happens when t_{calc} is positive) or $t_{calc} \leq t_{\alpha/2}$ (this happens when t_{calc} is negative). In this case $t_{calc} = 1.76 < 1.99 = t_{1-\alpha/2}$, so we fail to reject the null hypothesis or simply we accept the null hypothesis.

The picture for the two-tailed test is as follows:

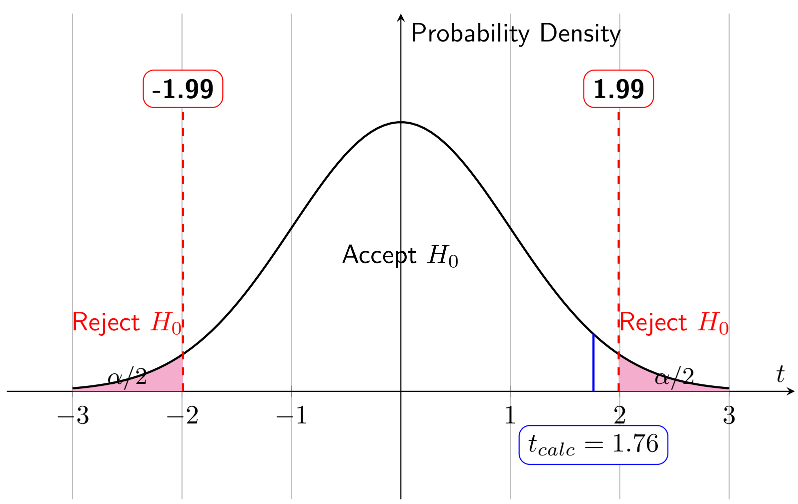

Figure 4: Two-tailed hypothesis test at 5% significance level. The blue line represents the test statistic $t_{calc} = 1.76$, and the red dashed lines are the critical values at $t_{0.025,63} = -1.99$ and $t_{0.975,63} = 1.99$. The shaded areas represent the rejection regions for the null hypothesis H_0.

p-**value approach:** For the p-value approach we can do following calculations

- If $t_{calc} > 0$, then p-value $= 2 \cdot \mathbb{P}(T > t_{calc})$, which can be calculated in Excel using the formula:

$$p\text{-value} = 2 \cdot (1 - \texttt{T.DIST}(t_{calc}, 63, 1))$$

- If $t_{calc} < 0$, then p-value $= 2 \cdot \mathbb{P}(T < t_{calc})$, which can be calculated in Excel using the formula:

$$p\text{-value} = 2 \cdot \texttt{T.DIST}(t_{calc}, 63, 1)$$

- If $t_{calc} = 0$, then p-value $= 1$.

In this case our $t_{calc} = 1.76 > 0$, so we can use the first formula to calculate the p-value:

$$p\text{-value} = 2 \cdot (1 - \texttt{T.DIST}(1.76, 63, 1)) = 0.083$$

and in our case we have $\alpha = 0.05$, so p-value $> \alpha$, hence we fail to reject the null hypothesis or simply we accept the null hypothesis. This means we do not have enough evidence to conclude that the average amount spent for lunch by the students at the Economics Dept. is not equal to 20 TK. So Rocky is correct here. Welcome Rocky!

(b)

(c)

Solution: Question 3

We now assume that the population standard deviation is unknown, and do the same test using both the critical value approach and the p-value approach.

(a)

(b)

(c)

Solution: Question 4

(a) The population of this study is **all students from the Economics Dept. at East West University**.

(b) No, this is not a good sample, because the sample is only from the Economics Dept. at EWU, but we want to find out whether the students in general, not just from the Economics dept. are satisfied with the quality of food served at the cafeteria. So this sample is not representative of the entire population of students at EWU.

(c) In theory we can calculate the population proportion p as follows:

$$p = \frac{\text{Number of students who are satisfied with the quality of food served at the cafeteria}}{\text{Total number of students at the Economics Dept.}}$$

(d) The sample proportion $\overline{p}$ of students who are satisfied with the quality of food served at the cafeteria can be calculated as follows:

$$\overline{p} = \frac{\text{Number of students in the sample who are satisfied with the quality of food served at the cafeteria}}{\text{Total number of students in the sample}}$$

From the data in `Student_Satisfaction.xlsx`, we can calculate $\overline{p}$ as follows:

$$\overline{p} = \frac{40}{64} = 0.625$$

(e) The point estimate of the population proportion of students who are satisfied with the quality of food served at the cafeteria is $\overline{p} = 0.625$.

(f) If we think each of the row is a Bernoulli random variable which have same distribution and independent, then we are thinking a random sample

Table 1: Bernoulli Random Random Sample

X_i
X_1
$\vdots$
X_n

where $X_i \sim \text{Bern}(p)$ for $i = 1, 2, \ldots, n$, and they are independent.

In this case we know that the mean of the Bernoulli random variable is

$$\mathbb{E}[X_i] = p$$
$$\mathbb{V}(X_i) = p(1-p)$$

The estimated variance of the Bernoulli random variable in this case is:

$$\hat{\mathbb{V}}(X_i) = \overline{p}(1-\overline{p}) = 0.625(1-0.625) = 0.234375$$

(g) With the i.i.d. assumption, the mean and variance of the sample proportion $\overline{p}$ when it is a random variable are (you should be aboe to prove it, we already proved it in the class):

$$\mathbb{E}[\overline{p}] = p$$
$$\mathbb{V}(\overline{p}) = \frac{p(1-p)}{n}$$

Central Limit Theorem tells us that the sample proportion $\overline{p}$ is approximately normally distributed when n is large enough, i.e.,

$$\overline{p} \sim \mathcal{N}\Big(\mathbb{E}(\overline{p}), \mathbb{V}(\overline{p})\Big) \quad \text{approximately as } n \to \infty$$
$$\overline{p} \sim \mathcal{N}\Big(p, \frac{p(1-p)}{n}\Big) \quad \text{approximately as } n \to \infty$$

(h) The standard error of the sample proportion is given by the formula:

$$S.E.(\overline{p}) = \sqrt{\mathbb{V}(\overline{p})} = \sqrt{\frac{p(1-p)}{n}}$$

However this quantity is unknown, so we can use the estimated variance of the Bernoulli random variable to calculate the estimated standard error of the sample proportion:

$$\widehat{S.E.}(\overline{p}) = \sqrt{\hat{\mathbb{V}}(\overline{p})} = \sqrt{\frac{\overline{p}(1-\overline{p})}{n}} = \sqrt{\frac{0.625(1-0.625)}{64}} = 0.061$$

(i) We give a summary of confidence interval before we give this answer, recall whenever we assume Normality the confdence interval follows a similar pattern. First note, that

$$\overline{X} \pm S.E.(\overline{X}) \cdot z_{1-\alpha/2} \quad = \quad \overline{X} \pm \underbrace{\sigma/\sqrt{n}}_{\sqrt{\mathbb{V}(\overline{X})}} \cdot z_{1-\alpha/2}$$

where we used

$$S.E.(\overline{X}) = \sqrt{\mathbb{V}(\overline{X})} = \sigma/\sqrt{n}$$

Or if we don't know σ we don't know $\mathbb{V}(\overline{X})$ we can use estimated standard error, i.e., then we can write the confidence interval as follows:

$$\overline{X} \pm \widehat{S.E.}(\overline{X}) \cdot t_{n-1,1-\alpha/2} \quad = \quad \overline{X} \pm \underbrace{s/\sqrt{n}}_{\widehat{\mathbb{V}}(\overline{X})} \cdot t_{n-1,1-\alpha/2}$$

where we used

$$\widehat{S.E.}(\overline{X}) = \sqrt{\widehat{\mathbb{V}}(\overline{X})} = s/\sqrt{n}$$

and when we apply CLT, we have

$$\overline{X} \pm \underbrace{s/\sqrt{n}}_{\widehat{\mathbb{V}}(\overline{X})} \cdot z_{1-\alpha/2}$$

For proportion we will use the CLT result, so our confidence interval for the population proportion of students who are satisfied with the quality of food served at the cafeteria is given by:

$$\overline{p} \pm \underbrace{\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}}_{\widehat{S.E.}(\overline{p}) = \sqrt{\widehat{\mathbb{V}}(\overline{p})}} \cdot z_{1-\alpha/2}$$

where we used the estimated standard error of the sample proportion

$$\widehat{S.E.}(\overline{p}) = \sqrt{\widehat{\mathbb{V}}(\overline{p})} = \sqrt{\frac{\overline{p}(1-\overline{p})}{n}}$$

So the formula for the confidence interval is

$$\overline{p} \pm \sqrt{\frac{\overline{p}(1-\overline{p})}{n}} \cdot z_{1-\alpha/2}$$

where we can construct the interval with

$$\begin{aligned} &0.625 \pm 0.061 \cdot 1.64 \\ &0.625 \pm 0.10 \end{aligned}$$

With Interval we can write

$$(0.525,\ 0.725)$$

Interpretation: This is a fixed confidence interval estimate of the unknown population proportion p, Here is p is the proportion of students who are satisfied with the lunch served at the cafeteria. The unknown p is either in this interval or not. However if we take similar samples 100 times then we can say that out of 100, roughly 90 times p will fall inside the interval and 10 times it will not.

(j) Now we have a testing problem, we want to test the hypothesis that the proportion of students who are satisfied with the quality of food served at the cafeteria is at least 0.7. The null and alternative hypotheses are:

$$H_0 : p \geq 0.7$$
$$H_a : p < 0.7$$

This will be a large sample test for the population proportion, and we can use the sample proportion $\overline{p}$ to calculate the test statistic. The test statistic is given by:

$$z_{calc} = \frac{\overline{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

where $p_0 = 0.7$ is the hypothesized population proportion under the null hypothesis (Recall we do testing with sampling distribution of the sample proportion under the Null). So in our case we have

$$z_{calc} = \frac{0.625 - 0.7}{\sqrt{\frac{0.7(1-0.7)}{64}}} = -1.30$$

This is a lower tail test, so we can use the critical value approach or the p-value approach to test the hypothesis.

Critical Value Approach:

We need to calculate z_α, where $\alpha = 0.05$ is the significance level. We can calculate the critical value as follows:

$$z_\alpha = z_{0.05} = -1.645$$

Now we know that for the lower tail test, we reject the null hypothesis if $z_{calc} < z_\alpha$. In this case, we have $z_{calc} = -1.30 > -1.645 = z_{0.05}$, so we fail to reject the null hypothesis or simply we accept the null hypothesis. This means we do not have enough evidence to conclude that the proportion of students who are satisfied with the quality of food served at the cafeteria is less than 0.7.

P Value Approach:

We can do the same test using the p-value approach. The p-value for this test is given by:

$$\begin{aligned} p\text{-value} &= \texttt{NORM.DIST}(z_{calc}, 0, 1, 1) \\ &= \texttt{NORM.DIST}(-1.30, 0, 1, 1) \\ &= 0.0968 = 9.6\% \end{aligned}$$

So since p-value $> \alpha = 0.05$, we fail to reject the null hypothesis or simply we accept the null hypothesis. Same conclusion as before. So Rocky is correct here. Welcome Rocky!

"The Only Relevant Test of the Validity of a Hypothesis is Comparison of Prediction with Experience."

- Milton Friedman