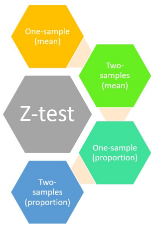


Z - Test Contd.

Please note that any topics that are not covered in today's lecture will be covered in the next lecture.

Content

1. Recap Framework for Hypothesis Testing
2. Two-sample Z-Test
3. Z-Test Proportions

Recap Framework for Hypothesis Testing

We start any Hypothesis Testing problem with 2 things:

- Assumption
- Data

We have a framework to compute a quantifiable metric that will help us decide if we should accept or reject our null hypothesis.

Let's summarise it into steps:-

1. Setup Null and Alternate Hypothesis
2. Choose the distribution (Gaussian, Binomial, etc), and hence the test statistic.
3. Select the Left vs Right vs Two-Tailed test, as per the hypothesis
4. Compute the P-Value
5. Compare the P-Value to the Significance Level (α) and Fail to reject/reject the Null Hypothesis accordingly.

Two-sample Z-Test

Imagine you're working for a renowned institution like ICMR, WHO, or FDA, and your task is to determine whether two different medicines, **M1** and **M2**, are equally effective in recovery time.

> This means that we want to test whether the population mean recovery times for medicines **M1** and **M2** are equal.

Let's set up our hypotheses:

- Null Hypothesis (H_0:): $\mu 1 = \mu 2$
- Alternative Hypothesis (H_a:): $\mu 1 \neq \mu 2$

where,

- μ1: Represents the population mean for group 1.
- μ2: Represents the population mean for group 2.

- For medicine M1, we have data from 100 patients, denoted as $x_{1,1}$, $x_{1,2}$, $x_{1,3}$, and so on, up to $x_{1,100}$.
- Similarly, for medicine M2, we have data from 90 patients, $x_{2,1}$, $x_{2,2}$, $x_{2,3}$, and so forth, up to $x_{2,90}$.
- We take sample means for both samples M1 and M2 which is represented as $\bar{x}_1$ (for M1) and $\bar{x}_2$for M2)

Since $n_1 = 100$ and $n_2 = 90$, we can compute our test statistic, Z as: Z = $\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

Where,

- Z: The z-score, a standard normal variable used to determine the probability of the observed difference between the two samples.
- $\bar{x}_1$: The mean of the first sample.
- $\bar{x}_2$: The mean of the second sample.
- σ_1: The standard deviation of the first population.
- σ_2: The standard deviation of the second population.
- n_1: The size of the first sample.
- n_2: The size of the second sample.

Now, we have two parts to test the test statistic

Part 1: Interpreting the Test Statistic: Is This Sensible?

Let's analyze the first part of our test statistic.

- If our null hypothesis is true, meaning that there's no difference in the effectiveness of the two medicines, our test statistic will be close to 0.
- Conversely, if the alternative hypothesis is correct, our test statistic will be significantly different from 0.

    - It will have either a large positive value or a large negative value

- Population standard deviations (σ) represent true values but are often unknown.
    - Since population standard deviations are unknown, sample standard deviations (S1 and S2) can be used as the reliable estimations (S1 ≈ σ1, S2 ≈ σ2).
- The reliability of these estimations improves with larger sample sizes (n1 and n2), ideally exceeding 30.

Using sample standard deviations as approximations for population standard deviations is a common and practical approach in real-world hypothesis testing scenarios.

Part 2: Distribution of Test Statistic Under the Null Hypothesis:

- It follows a **normal distribution** with a mean of 0 and a standard deviation of 1.
- This distribution is represented as `z(0,1)` or simply a standard normal distribution.

STEP 1:

> **What should be the null and alternate hypothesis?**

- Null Hypothesis (H_0:): The mean recovery times for medicines M1 and M2 are the same (i.e.,$\mu 1 = \mu 2$)
- Alternative Hypothesis (H_a:): The mean recovery times for medicines M1 and M2 are the not same (i.e.,$\mu 1 \neq \mu 2$)

STEP 2:

We have given **M1** for **100** patients and **M2** for **90** patients.

> **What is the distribution?**

- Normal distribution.

STEP 3:

> **Is the team looking for an effect towards the left side or right side or two-tailed?**

- Two-tailed test

Generating random recovery times for both medicines M1 & M2

```
import random
import numpy as np

# Set a random seed for reproducibility
random.seed(123)  # You can use any integer as the seed

# Create an empty list to store the recovery times
M1_data = []

# Generate 100 random recovery times
for _ in range(100):
    recovery_time = random.uniform(5.0, 20.0)  # Generating values between 10 and
    M1_data.append(np.round(recovery_time,0))

# Print the generated data
print("M1_data:",M1_data)

# Create an empty list to store the recovery times
M2_data = []

# Generate 90 random recovery times
for _ in range(90):
    recovery_time = random.uniform(5.0, 30.0)  # Generating values between 5 and
    M2_data.append(np.round(recovery_time,0))

# Print the generated data
print("M2_data:",M2_data)
```

```
M1_data: [6.0, 6.0, 11.0, 7.0, 19.0, 6.0, 13.0, 10.0, 18.0, 7.0, 10.0, 10.0, 9
M2_data: [24.0, 13.0, 14.0, 21.0, 29.0, 27.0, 15.0, 23.0, 22.0, 14.0, 11.0, 7
```

STEP 4:

We perform Two sample Z-test and calculate the P-Value

Link for the official Documentation

```
# import a library to perform a Z-test
from statsmodels.stats import weightstats as stests
from scipy import stats


z_score, pval = stests.ztest(x1 = M1_data, x2 = M2_data, value = 0, alternative =

# print the test statistic and corresponding p-value
print("Z-score: ", z_score)
print("p-value: ", pval)
```

```
Z-score:  -7.68917478890992
p-value:  1.4808703984296164e-14
```

STEP 5:

We defined $\alpha = 0.01$ for confidence level 99%

```
alpha = 0.01

if pval < alpha:
  print("Reject the null hypothesis, (i.e, The recovery time of two medicines are
else:
  print("Fail to reject the null hypothesis  (i.e, The recovery time of two medic
```

```
Reject the null hypothesis, (i.e, The recovery time of two medicines are diff
```

Since $1.4808703984296164e - 14 < \alpha$

This means that we can **Reject the null hypothesis**

So, we can report that the recovery times of two medicines M1 & M2 are different.

Let's look into another example:

```
A car manufacturer conducted a study to compare the fuel efficiency of two different engi
```

They collected data from two groups: Group X and Group Y.

- In Group X, a random sample of 50 cars with Engine X had an average fuel efficiency of 30 miles per gallon (mpg) with a standard deviation of 3 mpg.
- In Group Y, a random sample of 60 cars with Engine Y had an average fuel efficiency of 32 mpg with a standard deviation of 2.5 mpg.

The significance level (α) is set at 0.05.

Can it be concluded that one engine type is more fuel-efficient than the other?

When no data array is provided, the formula-based approach is necessary for solving the problem.

```
import numpy as np
from scipy import stats


# Null hypothesis: The population means of fuel efficiency for Engine X and Engin
# Alternative hypothesis: One engine type is more fuel-efficient than the other.

# Define the function to calculate the test statistic and corresponding p-value
def TwoSampZTest(samp_mean_1, samp_mean_2, samp_std_1, samp_std_2, n1, n2):
  # Calculate the test statistic
  denominator = np.sqrt((samp_std_1**2 / n1) + (samp_std_2**2 / n2))
  z_score = (samp_mean_1 - samp_mean_2) / denominator
  return z_score



# Given data
sample_mean_X = 30 # Average fuel efficiency for Group X (Engine X)
sample_mean_Y = 32 # Average fuel efficiency for Group Y (Engine Y)
sample_std_X = 3 # Standard deviation for Group X
sample_std_Y = 2.5 # Standard deviation for Group Y
significance_level = 0.05
sample_size_X = 50 # Sample size for Group X
sample_size_Y = 60 # Sample size for Group Y



# Calculate the z-score using the function
z_score = TwoSampZTest(sample_mean_X, sample_mean_Y, sample_std_X, sample_std_Y,

# Calculate the two-tailed p-value
p_value = 2 * (1 - stats.norm.cdf(abs(z_score)))

# A two-tailed z-test considers deviations from the null hypothesis in both posit
# Using abs(z_score) ignores the sign of the deviation, focusing solely on its ma

# Compare the p-value to the significance level
if p_value < significance_level:
  conclusion = "Reject the null hypothesis. Engine Y is more fuel-efficient."
else:
  conclusion = "Fail to reject the null hypothesis. No significant difference in

print(f'z-score: {z_score:.4f}')
print(f'p-value: {p_value:.4f}')
print('Conclusion:', conclusion)
```

```
z-score: -3.7518
p-value: 0.0002
Conclusion: Reject the null hypothesis. Engine Y is more fuel-efficient.
```

Conditions for using the Two Sample Z-Test:

- The population must have known finite means (μ) and standard deviations (σ).
- We need to know σ1 and σ2, or we should be able to estimate S1 and S2 (when n1 and n2 are not too small, preferably greater than 30).

- The data in each population must be continuous, not discrete.
- The data in each population should be approximately normally distributed.
- N1 and N2 don't have to be the same; (i.e., they can be the same or different.)

Z-Proportions Test

Let's say you are a Data Scientist working for a Web Application

- The website wants to add a new feature to make "more" customers buy their product to increase their Proportion of Sales
  - Proportion of Sales = No.of customers buying the product / No.of customers visiting the web page

What do you think can be the impact of adding the new feature?

Case 1: There is No impact

Case 2: There is impact

a) It may increase the proportion of sales

- By making **user experience better**

b) It may decrease the proportion of sales

- People may get confused
- May NOT like the new interface

> How do we check our assumptions and validate them?

We need to evaluate change in proportions of sales before and after introducing the new feature

- To decide whether should we even go with the new feature or not
- We need a way to test the effectiveness of changes we make in the system before we roll out the changes

This is where the Test of Proportions comes into the picture

- We need to test for consequences before we can release a change
- Test whether our assumptions/beliefs about those changes are even True or NOT
- This becomes our motivation to discuss the Test of Proportions

One Sample Z-Proportion Test

Imagine you are a product manager in a company, and you want to determine the satisfaction rate of customers with a new product.

- A proportion is a way to express a part of a whole.
- It's often used to measure the percentage of a specific outcome within a larger population.
- In our case, it's the proportion of satisfied customers.

Null Hypothesis (H_0): The proportion of satisfied customers with the new product is equal to a specified value, e.g., 0.70 (70%).

Alternative Hypothesis (H_a): The proportion of satisfied customers with the new product are not equal to the specified value.

To use this test, we need to ensure the following conditions are met:

- The sample size is sufficiently large, typically greater than 30.
- The data can be assumed to follow a normal distribution.

The Test Statistic:

Test statistic for the one-sample proportion test is as follows: $Z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$

Breaking it down:

- $\hat{p}$: is the sample proportion of satisfied customers(i.e., observed value of proportion)
  - $\hat{p} = \frac{x}{n}$
- p: is the specified proportion under the null hypothesis. (i.e., Hypothesized or Population proportion)
- n: is the sample size.

Interpreting the Test Statistic:

The test statistic tells us if there's a significant difference between the conversion rates.

- Under the **null hypothesis**, the test statistic is **close to 0** because we assume that the conversion rates are the same.
- However, under the **alternative hypothesis**, the test statistic can be either a **large positive** or a **large negative value**, indicating a difference between the web pages.
- The null and alternative hypothesis is given as:
  - $H_0 : \hat{p} = p$ **or** $\hat{p} \geq p$ **or** $\hat{p} \leq p$
  - $H_1 : \hat{p} \neq p$ **or** $\hat{p} < p$ **or** $\hat{p} > p$

Distribution of the Test Statistic:

- Our test statistic follows the standard normal distribution, denoted as $Z \sim (0, 1)$, under the null hypothesis.

- This means it's shaped like a bell curve with a mean of 0 and a standard deviation of 1.

The framework is same that we learned in the previous class about comparing means using the Z-test applies here.

> Let's see with an example.

Question:

- You are a product manager for a company that has recently launched a new product.
- Customer satisfaction is a critical metric, and you want to determine if the proportion of satisfied customers with the new product meets your target satisfaction level of 70%.
- You collected a random sample of 150 customer reviews, and 115 of them expressed satisfaction with the product.

STEP 1:

> **What should be the null and alternate hypothesis?**

- Null Hypothesis (H_0): The proportion of satisfied customers with the new product is equal to the target satisfaction level of 70%. (i.e.,$p = 0.70$).
- Alternative Hypothesis (H_a): The proportion of satisfied customers with the new product is not equal to the target satisfaction level of 70%.($p \neq 0.70$).

STEP 2:

> **What is the distribution it follows?**

- Normal Distribution.

STEP 3:

> **Is the team looking for an effect towards the left side or right side or two-tailed?**

- Two-tailed test

STEP 4:

We perform One sample Z- Proportion test and calculate the P-Value

```
import numpy as np
import scipy.stats as stats

# Define the data
# Number of satisfied customers
satisfied_customers = 115
# Total number of customers in the sample
total_customers = 150

# Specify the target satisfaction level (null hypothesis)
target_satisfaction = 0.70  # 70%

# Define the hypotheses
# Null Hypothesis (H0): The proportion of satisfied customers is equal to the tar
# Alternative Hypothesis (Ha): The proportion of satisfied customers is not equal
p = target_satisfaction

# Calculate the test statistic (Z)
# p_hat = sample proportion
p_hat = satisfied_customers / total_customers
n = total_customers
Z = (p_hat - p) / np.sqrt((p * (1 - p)) / n)

# Interpret the test statistic
# Z follows a standard normal distribution. We will calculate the two-tailed p-va

#Calculate the p-value
p_value = 2 * (1 - stats.norm.cdf(np.abs(Z)))

# Print the results
print(f"Z-statistic = {Z}")
print(f"P-value = {p_value}")
```

```
Z-statistic = 1.7817416127494983
P-value = 0.07479137758694376
```

STEP 5:

We defined $\alpha = 0.05$ for confidence level 95%

```
alpha = 0.05  # Significance level
if p_value < alpha:
    decision = "Reject the null hypothesis"
else:
    decision = "Fail to reject the null hypothesis"

if decision == "Reject the null hypothesis":
    conclusion = "There is enough evidence to conclude that the proportion of sat
else:
    conclusion = "There is not enough evidence to conclude that the proportion of

# Print the results
print(f"Decision: {decision}")
print(f"Conclusion: {conclusion}")
```

```
Decision: Fail to reject the null hypothesis
Conclusion: There is not enough evidence to conclude that the proportion of sa
```

Since $0.07479137758694376 > \alpha$, We **Fail to reject the null hypothesis**

Conclusion: There is not enough evidence to conclude that the proportion of satisfied customers with the new product is statistically different from the target satisfaction level.

Two Sample Z-Proportion Test

Imagine you run an online retail website and are keen to optimise your web pages to boost sales.

To optimise your web pages to boost sales. you focus on a critical metric in e-commerce called the **conversion rate**.

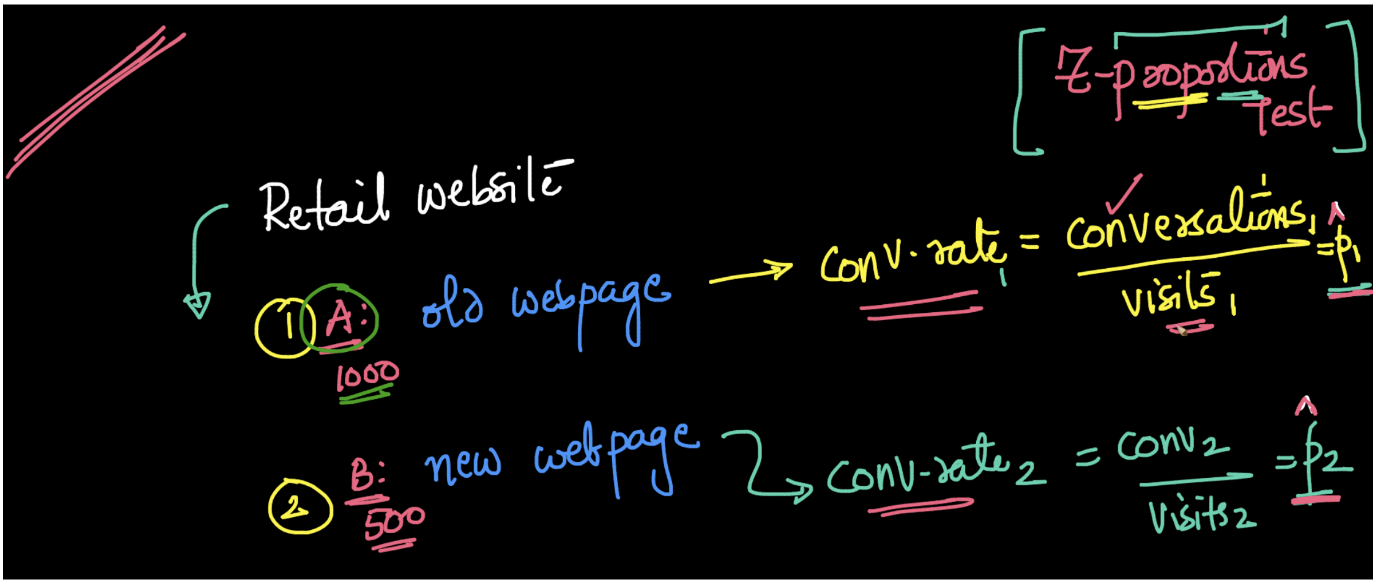
- Conversion rate is a simple yet vital metric that measures the proportion of visitors who make a purchase.
- It is the proportion of conversions (sales) to the total number of website visits.

We need to explore how to **compare conversion rates** between your old webpage and a new one.

Now, Imagine we have two web pages to compare: the old one and the new one.

To compare these two conversion rates, we turn to hypothesis testing. Here are our hypotheses:

- **Null Hypothesis (H_0)**: The conversion rates of the old and new web pages are the same ($p1 = p2$).
- **Alternative Hypothesis (H_a)**: The conversion rates of the old and new web pages are different ($p1 \neq p2$).



Conditions for the Z-Proportions Test:

Before we dive into the test, we need to ensure that our conditions are met:

- Sample sizes should be greater than 30.
- The data should follow a normal distribution.

The Test Statistic:

Our test statistic for the Z-proportions test is as follows:

- $Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$
- Breaking it down:
  - $\hat{p}_1$ and $\hat{p}_2$ are sample proportions for conversions on the old and new web pages.
    - $\hat{p}_1 = \frac{x_1}{n_1}$
    - $\hat{p}_2 = \frac{x_2}{n_2}$
  - $\hat{p}$: is the combined sample proportion for both web pages. calculated as the average of $\hat{p}_1$ and $\hat{p}_2$
    - $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$
  - n_1, n_2 sample sizes for groups 1 and 2, respectively.

Further, we know $0 <= \hat{p} <= 1$

Interpreting the Test Statistic:

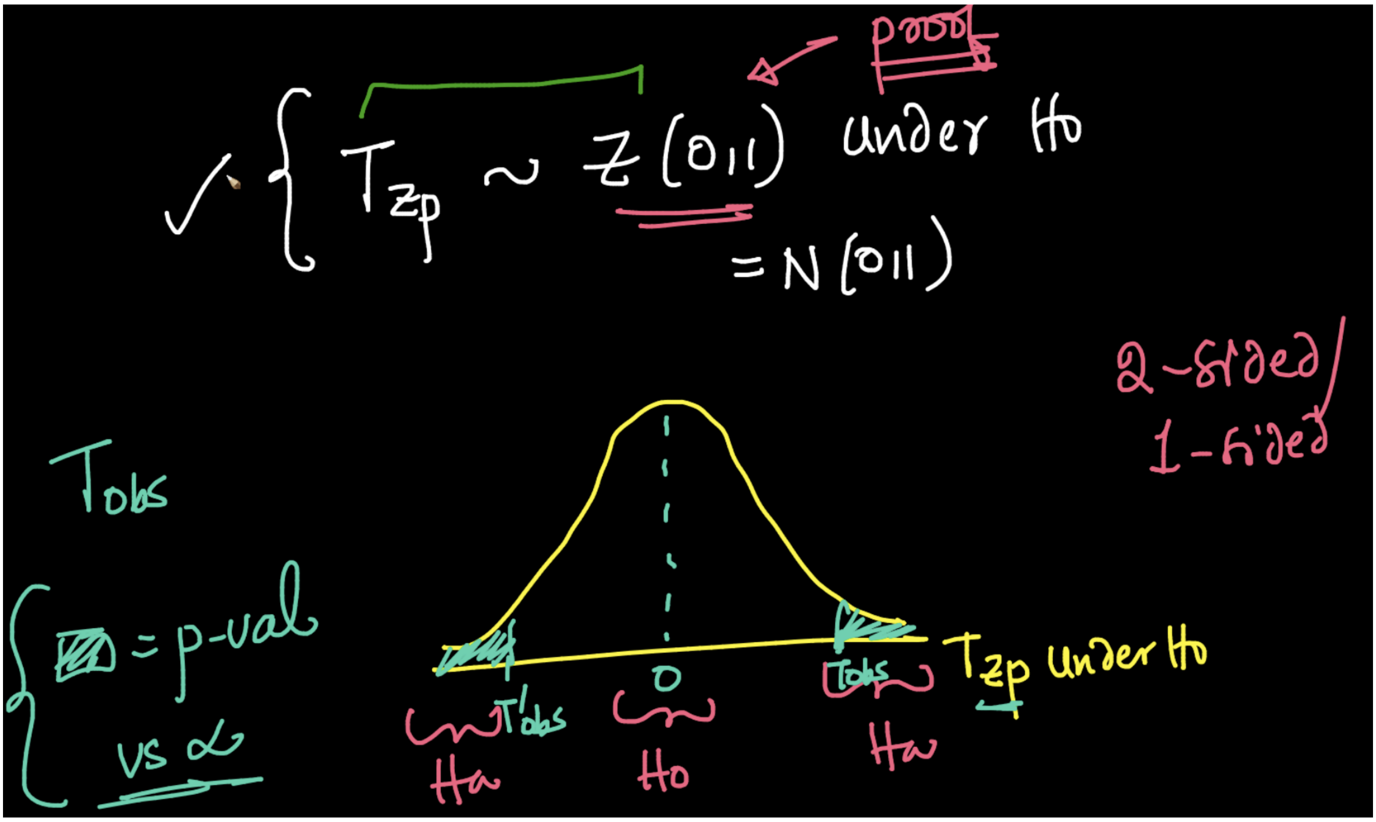
The test statistic tells us if there's a significant difference between the conversion rates.

- Under the **null hypothesis**, the test statistic is **close to 0** because we assume that the conversion rates are the same.
- However, under the **alternative hypothesis**, the test statistic can be either a **large positive** or a **large negative value**, indicating a difference between the web pages.

Distribution of the Test Statistic:

- Our test statistic follows the standard normal distribution, denoted as $Z \sim (0, 1)$, under the null hypothesis.
- This means it's shaped like a bell curve with a mean of 0 and a standard deviation of 1.

Given that our alternative hypothesis suggests that the conversion rates can differ in either direction, we perform a **two-tailed test**. This means that we look at both ends of the distribution.



The same framework we learned in the previous class about comparing means using the Z-test applies here.

In this case, we are comparing proportions using the Z-proportions test, with the condition that both sample sizes (n1 and n2) are large.

> Let's see with an example.

Question:

```
You are the manager of an e-commerce website, and you have recently implemented a new web

To evaluate the effectiveness of the new page, you collected data on the conversion rates
The conversion rate is defined as the proportion of visitors who make a purchase.

- For the old web page (Web Page A), you had **1000** visitors, resulting in **50** conve
- For the new web page (Web Page B), you had **500** visitors, resulting in **30** conver

Now, you want to determine if there is a statistically significant difference in the conv
```

STEP 1:

> **What should be the null and alternate hypothesis?**

- Null Hypothesis (H_0): The conversion rates of the old web page (Web Page A) and the new web page (Web Page B) are the same. ($p1 = p2$).
- Alternative Hypothesis (H_a): The conversion rates of the old web page (Web Page A) and the new web page (Web Page B) are different.($p1 \neq p2$).

STEP 2:

> **What is the distribution it follows?**

- Normal Distribution.

STEP 3:

> **Is the team looking for an effect towards the left side or right side or two-tailed?**

- Two-tailed test

STEP 4:

We perform two sample Z- Proportion test and calculate the P-Value

[Documentation]

```
import statsmodels.api as sm
import numpy as np

# Define the data
conversions = np.array([50, 30])  # Number of conversions for Web Page A and Web
visits = np.array([1000, 500])  # Number of visits for Web Page A and Web Page B

# Perform the Z-proportions test
z_stat, p_value = sm.stats.proportions_ztest(conversions, visits, alternative='tw

# Print the results
print(f"Z-statistic = {z_stat}")
print(f"P-value = {p_value}")
```

```
Z-statistic = -0.8125338562826986
P-value = 0.4164853677823287
```

Solving the same using the formula

- $Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

```
import numpy as np
import scipy.stats as stats

# Step 1: Define the data
# Data for the old web page (Web Page A)
visits_1 = 1000
conversions_1 = 50
```

STEP 5:

We defined $\alpha = 0.05$ for confidence level 95%

```
# Step 2: Define the hypotheses
# Make a decision and conclusion
if p_value < alpha:
    decision = "Reject the null hypothesis"
else:
    decision = "Fail to reject the null hypothesis"

if decision == "Reject the null hypothesis":
    conclusion = "There is enough evidence to conclude that the conversion rates
else:
    conclusion = "There is not enough evidence to conclude that the conversion ra

# Print the results
print(f"Decision: {decision}")
print(f"Conclusion: {conclusion}")
```

```
Decision: Fail to reject the null hypothesis
Conclusion: There is not enough evidence to conclude that the conversion rate:
```

Since $0.4164853677823288 > \alpha$

This means that we **Fail to reject the null hypothesis**

Conclusion: There is not enough evidence to conclude that the conversion rates of the new web page (Web Page B) are statistically different from the old web page (Web Page A).

The Z-proportions test is a powerful tool in e-commerce for comparing the conversion rates of different web pages.