SNS COLLEGE OF TECHNOLOGY

An Autonomous Institution
Coimbatore-35

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & DATA SCIENCE

23ADT202 – FUNDAMENTALS OF DATA SCIENCE AND ANALYTICS

II YEAR IV SEM

UNIT III –Z Test

Z-test

EMPATHY:

- A Z-test is a type of hypothesis test that compares the sample's average to the population's average
- It calculates the Z-score and tells us how much the sample average is different from the population average by looking at how much the data normally varies.
- It is particularly useful when the sample size is large >30. It is also known as Z-Statistics and its formula is:

$$\text{Z-Score} = \frac{\bar{x} - \mu}{\sigma}$$

where:

- $\bar{x}$: mean of the sample.
- μ : mean of the population.
- σ : Standard deviation of the population.

Let's understand with the help of example The average family annual income in India is 200k with a standard deviation of 5k and the average family annual income in Delhi is 300k. Then Z-Score for Delhi will be.

$$\begin{aligned}\text{Z-Score} &= \frac{\bar{x} - \mu}{\sigma} \\ &= \frac{300 - 200}{5} \\ &= 20\end{aligned}$$

This indicates that the average family's annual income in Delhi is 20 standard deviations above the mean of the population (India).

DEFINE:

Steps to perform Z-test

1. First we identify the null and alternate hypotheses.
2. Then we determine the level of significance (α).
3. Next we find the critical value of Z in the z-test.
4. Then we calculate the z-test statistics using the formula :

$$Z = \frac{(\bar{x} - \mu)}{(\sigma/\sqrt{n})}$$

Where:

- $\bar{x}$: mean of the sample.
- μ : mean of the population.
- σ : Standard deviation of the population.
- n : sample size.

5. Now we compare with the hypothesis and decide whether to reject or not reject the null hypothesis.

Example Problem: One-Sample Z-Test (Step-by-Step)

Problem Statement

A data analytics team claims that the **average daily website traffic** is **10,000 users**. To verify this, a sample of **64 days** is taken. The results show:

- Sample mean ($\bar{X}$)= **10,400 users**
- Population standard deviation (σ) = **1,600 users**
- Significance level (α) = **0.05**

Test whether the claim is valid using a **Z-test**.

IDEATE:

Step 1: State the Hypotheses

Since we are checking whether the mean is **different**, this is a **two-tailed test**.

- **Null Hypothesis (H_0):**

$$\mu = 10{,}000$$

- **Alternative Hypothesis (H_1):**

$$\mu \neq 10{,}000$$

Step 2: Choose the Significance Level

$$\alpha = 0.05$$

This means we are allowing a **5% chance of error**.

Step 3: Identify the Z-Test Formula

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Where:

- $\bar{X}$ = sample mean
- μ = population mean
- σ = population standard deviation
- n = sample size

Step 4: Substitute the Given Values

$$Z = \frac{10{,}400 - 10{,}000}{1{,}600/\sqrt{64}}$$

$$Z = \frac{400}{1{,}600/8}$$

$$Z = \frac{400}{200}$$

$$Z = 2.00$$

Step 5: Determine the Critical Z-Value (Decision Rule)
For a **two-tailed test** at $\alpha = 0.05$:

$$Z_{critical} = \pm 1.96$$

❖ **How to Find the Critical Z-Value (±1.96)**

Given
- **Significance level (α) = 0.05**
- **Two-tailed Z-test**

1. Understand the Meaning of α
The **significance level (α)** represents the **probability of rejecting the null hypothesis when it is actually true**.
For:

$$\alpha = 0.05$$

This means:
Total **5% of the area** under the normal curve is in the **rejection region**

2: Split α for a Two-Tailed Test

In a **two-tailed test**, the rejection region is divided equally between **both tails** of the normal distribution.

$$\alpha = 0.05 \Rightarrow \frac{\alpha}{2} = 0.025$$

So:

- **2.5% area** in the **left tail**
- **2.5% area** in the **right tail**

3: Find the Area from the Mean to Z

Most **Z-tables give the area between the mean (0) and Z**.

Area between mean and Z:

$$0.5 - 0.025 = 0.475$$

Always remember:

Mean-to-Z tables → start from **0.5**

Left-tail tables → start from **1.0**

Since the standard normal distribution is symmetric, half of the total probability (0.5) lies on each side of the mean. Z-tables give the area from the mean to Z; therefore, we subtract the tail probability from 0.5.

4: Use the Z-Table

- Look for **0.475** in the **Z-table**
- The corresponding Z-value is **1.96**

This means:

- Area from **0 to 1.96 = 0.475**
- Remaining **0.025** is in the right tail

5: Write the Critical Z-Values

Since it is a **two-tailed test**, we take **both positive and negative values**:

$$Z_{critical} = \pm 1.96$$

Z Table from Mean (0 to Z)

z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07
0	0	0.00399	0.00798	0.01197	0.01595	0.01994	0.02392	0.0279
0.1	0.03983	0.0438	0.04776	0.05172	0.05567	0.05962	0.06356	0.0674
0.2	0.07926	0.08317	0.08706	0.09095	0.09483	0.09871	0.10257	0.1064
0.3	0.11791	0.12172	0.12552	0.1293	0.13307	0.13683	0.14058	0.1443
0.4	0.15542	0.1591	0.16276	0.1664	0.17003	0.17364	0.17724	0.1808
0.5	0.19146	0.19497	0.19847	0.20194	0.2054	0.20884	0.21226	0.2156
0.6	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24537	0.2485
0.7	0.25804	0.26115	0.26424	0.2673	0.27035	0.27337	0.27637	0.2793
0.8	0.28814	0.29103	0.29389	0.29673	0.29955	0.30234	0.30511	0.3078
0.9	0.31594	0.31859	0.32121	0.32381	0.32639	0.32894	0.33147	0.3339
1	0.34134	0.34375	0.34614	0.34849	0.35083	0.35314	0.35543	0.3576
1.1	0.36433	0.3665	0.36864	0.37076	0.37286	0.37493	0.37698	0.379
1.2	0.38493	0.38686	0.38877	0.39065	0.39251	0.39435	0.39617	0.3979
1.3	0.4032	0.4049	0.40658	0.40824	0.40988	0.41149	0.41308	0.4146
1.4	0.41924	0.42073	0.4222	0.42364	0.42507	0.42647	0.42785	0.4292
1.5	0.43319	0.43448	0.43574	0.43699	0.43822	0.43943	0.44062	0.4417
1.6	0.4452	0.4463	0.44738	0.44845	0.4495	0.45053	0.45154	0.4525
1.7	0.45543	0.45637	0.45728	0.45818	0.45907	0.45994	0.4608	0.4616
1.8	0.46407	0.46485	0.46562	0.46638	0.46712	0.46784	0.46856	0.4692
1.9	0.47128	0.47193	0.47257	0.4732	0.47381	0.47441	0.475	0.4755
2	0.47725	0.47778	0.47831	0.47882	0.47932	0.47982	0.4803	0.4807
2.1	0.48214	0.48257	0.483	0.48341	0.48382	0.48422	0.48461	0.485
2.2	0.4861	0.48645	0.48679	0.48713	0.48745	0.48778	0.48809	0.4884

6: Decision Rule (Exam Format)

- If **|Z calculated| > 1.96 → Reject H_0**
- If **|Z calculated| ≤ 1.96 → Fail to reject H_0**

Visual Understanding (Conceptual)

```
   Rejection    Acceptance      Rejection
     2.5%          95%             2.5%
------|----------|------------|------
    -1.96          0            +1.96
```

For a two-tailed Z-test at α = 0.05, the significance level is divided into two equal parts (0.025 each). From the Z-table, the Z-value corresponding to an area of 0.475 between the mean and Z is 1.96. Hence, the critical Z-values are ±1.96.

Step 6: Make the Decision

- Calculated Z = **2.00**
- Critical Z = **±1.96**

Since:

$$|2.00| > 1.96$$

☞ **Reject the null hypothesis**

Step 7: Conclusion (Statistical Decision)

There is a **statistically significant difference** between the sample mean and the claimed population mean.

Step 8: Interpretation (Data Science Context)

The analysis shows that the **average daily website traffic is significantly different from 10,000 users**.

This insight can help:

- Marketing teams evaluate campaign impact
- Product teams analyze user growth
- Data scientists update traffic forecasting models

PROTOTYPING:

How to Find the Critical Z-Value for a One-Tailed Z-Test

Given

- **Significance level (α) = 0.05**
- **One-tailed test** (either right-tailed or left-tailed)

Step 1: Understand What “One-Tailed” Means

In a **one-tailed test**, the **entire rejection region (α)** is placed **only on one side** of the normal distribution.

- **Right-tailed test** → rejection on the **right**
- **Left-tailed test** → rejection on the **left**

Step 2: Use the Full α in One Tail

Since it is **one-tailed**, we do **not split α**.

$$\alpha = 0.05$$

So:

- **5% area** is in **one tail only**

Step 3: Find the Area from Mean to Z

Most Z-tables give the **area between the mean (0) and Z**.

Area between mean and Z:

$$0.5 - 0.05 = 0.45$$

Step 4: Use the Z-Table

•Look for **0.45** in the Z-table

•The corresponding Z-value is **1.645**

Step 5: Write the Critical Z-Value

Right-Tailed Test

$$Z_{critical} = +1.645$$

Left-Tailed Test

$$Z_{critical} = -1.645$$

Step 6: Decision Rule

Right-Tailed Test

- If **Z calculated > 1.645 → Reject H_0**
- Otherwise → Fail to reject H_0

Left-Tailed Test

- If **Z calculated < −1.645 → Reject H_0**
- Otherwise → Fail to reject H_0

Visual Representation (Concept)

Right-Tailed Test

```
Acceptance Region     Rejection Region
------------------|-------------------
                 1.645
                  5%
```

Left-Tailed Test

```
Rejection Region     Acceptance Region
------------------|-------------------
       5%        -1.645
```

Right-Tailed Z-Test

Decision rule ($\alpha = 0.05$):

- If **$Z > 1.645$ → Reject H_0**
- Otherwise → Fail to reject H_0

Here:

$$Z = 2.00 > 1.645$$

✔ Conclusion:

👉 **Reject the null hypothesis (H_0)**

Interpretation:

There is **sufficient statistical evidence** to support the alternative hypothesis in the right-tailed test.

Left-Tailed Z-Test

Decision rule ($\alpha = 0.05$):

•If **$Z < -1.645$ → Reject H_0**

•Otherwise → Fail to reject H_0

Here:

$$Z = 2.00 \ (not\ less\ than\ -1.645)$$

✔**Conclusion:**

☞ **Fail to reject the null hypothesis (H_0)**

Interpretation:

There is **no sufficient evidence** to support the alternative hypothesis in the left-tailed test.

“Fail to Reject the Null Hypothesis”

Fail to reject the null hypothesis means that the **sample data does not provide enough statistical evidence** to support the alternative hypothesis at the chosen level of significance.

1. One-Tailed Test

TESTING:

A **one-tailed test** checks if the parameter is **either greater than or less than** a specific value.

(a) Right-Tailed Test

Used when we want to test whether the value is **greater than** the hypothesized value.

Alternative Hypothesis:

$$H_1: \mu > \mu_0$$

Keywords in the question:

- greater than
- higher than
- increased
- more than
- above

Example:

A company claims the average battery life is **10 hours**. A researcher wants to test if the battery life is **greater than 10 hours**.

- $H_0: \mu = 10$
- $H_1: \mu > 10$

➞ This is a **Right-Tailed Test**.

Decision rule at **α = 0.05**: Reject H_0 if **Z > 1.645**.

(b) Left-Tailed Test

Used when we want to test whether the value is **less than** the hypothesized value.

Alternative Hypothesis:

$$H_1: \mu < \mu_0$$

Keywords in the question:

- less than
- lower than
- decreased
- below
- smaller than

Example:

A company claims the average packet weight is **500 g**. A customer suspects the weight is **less than 500 g**.

- $H_0: \mu = 500$
- $H_1: \mu < 500$

➡This is a **Left-Tailed Test**.

Decision rule at **α = 0.05**:

Reject H_0 if **Z < −1.645**.

2. Two-Tailed Test

A **two-tailed test** checks whether the value is **different** (either greater or smaller).

Alternative Hypothesis:

$$H_1 : \mu \neq \mu_0$$

Keywords in the question:

- different from
- not equal to
- changed
- any difference

Example:

A manufacturer claims the average bulb life is **1000 hours**. A researcher wants to test whether the bulb life is **different from 1000 hours**.

- $H_0 : \mu = 1000$
- $H_1 : \mu \neq 1000$

➞This is a **Two-Tailed Test**.

Decision rule at **α = 0.05**:

Reject H_0 if **|Z| > 1.96**.

Example Problem: One-Tailed Z-Test

A company claims that the **average lifetime of its bulbs is 1000 hours**.

A random sample of **36 bulbs** has a **mean lifetime of 980 hours**.

The **population standard deviation is 120 hours**.

At **5% level of significance**, test whether the bulbs last **less than 1000 hours**.

Step 1: State the Hypotheses

This is a **left-tailed test**.

- **Null Hypothesis (H_0):**

$\mu = 1000$ hours

- **Alternative Hypothesis (H_1):**

$\mu < 1000$ hours

Step 2: Given Data

- Population mean, $\mu = 1000$
- Sample mean, $\bar{x} = 980$
- Population standard deviation, $\sigma = 120$
- Sample size, $n = 36$
- Significance level, $\alpha = 0.05$

Step 3: Test Statistic Formula

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Step 4: Calculate Z Value

$$Z = \frac{980 - 1000}{120/\sqrt{36}}$$

$$Z = \frac{-20}{120/6}$$

$$Z = \frac{-20}{20}$$

$$Z = -1.0$$

Step 5: Critical Value

For a **one-tailed test at 5% level**:

•Critical Z value = **−1.645**

Step 6: Decision Rule

- If **$Z \le -1.645$**, reject H_0
- If **$Z > -1.645$**, fail to reject H_0

Here:

$$Z = -1.0 > -1.645$$

Step 7: Conclusion

❌ **Fail to reject the null hypothesis**

☞ There is **no sufficient evidence** to conclude that the average bulb life is **less than 1000 hours** at 5% significance level.

Final Answer:

At 5% level of significance, the calculated Z value (−1.0) is greater than the critical value (−1.645). Hence, the null hypothesis is accepted.

There is no significant evidence to support the claim that the mean lifetime of the bulbs is less than 1000 hours.