Chapter 1

Z-Test for Hypothesis Testing

1.1 Introduction

Statistical hypothesis testing provides a formal mechanism for making decisions about population parameters using sample data. One of the most fundamental hypothesis tests is the **Z-test**, which is used to test hypotheses about population means when the population variance is known or when the sample size is sufficiently large.

1.2 Sampling Distribution of the Mean

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with population mean μ and population variance σ^2.

The sample mean is defined as

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

By the Central Limit Theorem, for large n, the sampling distribution of $\bar{X}$ is approximately normal:

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

1.3 Z-Test Statistic

To test a hypothesis about the population mean, we use the Z-statistic:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

Under the null hypothesis, this statistic follows a standard normal distribution:

$$Z \sim \mathcal{N}(0, 1)$$

1.4 One-Sample Z-Test

The one-sample Z-test is used to determine whether the mean of a population differs from a specified value.

1.4.1 Hypotheses

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

1.4.2 Decision Rule

For a significance level $\alpha = 0.05$, the critical value is ± 1.96. The null hypothesis is rejected if $|Z| > 1.96$.

1.5 Two-Sample Z-Test

The two-sample Z-test compares the means of two independent populations when their variances are known.

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

1.6 Summary

The Z-test is a classical hypothesis testing procedure that compares standardized sample statistics to the standard normal distribution. It forms the foundation for many advanced statistical inference techniques.

1.7 Worked Examples of Z-Test

This section presents illustrative examples of the Z-test with detailed calculations to demonstrate the practical application of the theory.

1.7.1 Example 1: One-Sample Z-Test

A manufacturer claims that the mean lifetime of a certain type of bulb is 1000 hours. The population standard deviation is known to be 120 hours. A random sample of $n = 36$ bulbs has a mean lifetime of 1040 hours. Test the claim at the 5% significance level.

Step 1: State the Hypotheses

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

Step 2: Compute the Test Statistic

The Z-statistic is given by:

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Substituting the given values:

$$Z = \frac{1040 - 1000}{120/\sqrt{36}} = \frac{40}{20} = 2.0$$

Step 3: Decision Rule

For a two-tailed test at $\alpha = 0.05$, the critical values are:

$$\pm 1.96$$

Since:

$$|Z| = 2.0 > 1.96$$

we reject the null hypothesis.

Step 4: Conclusion

There is sufficient statistical evidence at the 5% significance level to conclude that the mean lifetime of the bulbs is different from 1000 hours.

—

1.7.2 Example 2: Two-Sample Z-Test

Two independent production processes are compared for average output quality. The population standard deviations are known.

- Process A: $\bar{x}_1 = 75$, $\sigma_1 = 8$, $n_1 = 64$
- Process B: $\bar{x}_2 = 72$, $\sigma_2 = 6$, $n_2 = 49$

Test whether there is a significant difference between the two population means at the 5% significance level.

Step 1: State the Hypotheses

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Step 2: Compute the Test Statistic

The Z-statistic for two independent samples is:

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Substituting the values:

$$Z = \frac{75 - 72}{\sqrt{\frac{8^2}{64} + \frac{6^2}{49}}} = \frac{3}{\sqrt{1 + 0.7347}} = \frac{3}{1.317} \approx 2.28$$

Step 3: Decision Rule

For $\alpha = 0.05$ (two-tailed), the critical values are ± 1.96.

Since:

$$|Z| = 2.28 > 1.96$$

the null hypothesis is rejected.

Step 4: Conclusion

There is sufficient evidence to conclude that the mean outputs of the two processes are significantly different.

—

1.7.3 Example 3: Interpretation Using P-Value

Consider the one-sample Z-test in Example 1, where $Z = 2.0$.

The corresponding p-value is:

$$\text{p-value} = 2P(Z > 2.0) \approx 0.0455$$

Since:

$$\text{p-value} < \alpha = 0.05$$

the null hypothesis is rejected.

Interpretation

Assuming the null hypothesis is true, the probability of observing a sample mean at least as extreme as the one obtained is approximately 4.55%. This result is sufficiently unlikely to reject the null hypothesis at the 5% significance level.

—

1.8 Remarks for Practical Applications

- Larger sample sizes reduce the standard error, increasing the power of the Z-test.
- Statistical significance does not necessarily imply practical significance.
- In real-world applications, population standard deviations are often unknown, in which case the t-test is more appropriate.

1.9 AI-Oriented Examples of Z-Test

This section presents additional worked examples of the Z-test motivated by applications in Artificial Intelligence, Machine Learning, and Data Science.

1.9.1 Example 4: Model Inference Latency Analysis

An AI team claims that the average inference latency of a deployed model is 200 milliseconds. Historical monitoring shows that the population standard deviation of latency is 30 milliseconds.

A random sample of $n = 64$ inference requests shows an average latency of 212 milliseconds. Test the claim at the 5% significance level.

Step 1: Hypotheses

$$H_0 : \mu = 200$$

$$H_1 : \mu \neq 200$$

Step 2: Test Statistic

$$Z = \frac{212 - 200}{30/\sqrt{64}} = \frac{12}{3.75} = 3.2$$

Step 3: Decision

For $\alpha = 0.05$, the critical value is ± 1.96.

Since:

$$|Z| = 3.2 > 1.96$$

the null hypothesis is rejected.

Conclusion

There is strong statistical evidence that the average inference latency is different from the claimed 200 milliseconds.

—

1.9.2 Example 5: A/B Testing of Recommendation Models

Two recommendation models are compared based on average user engagement time.

- Model A (Baseline): $\bar{x}_1 = 4.6$ minutes, $\sigma_1 = 1.1$, $n_1 = 500$
- Model B (New): $\bar{x}_2 = 4.9$ minutes, $\sigma_2 = 1.2$, $n_2 = 500$

Test whether the new model significantly improves engagement at $\alpha = 0.05$.

Step 1: Hypotheses

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

Step 2: Test Statistic

$$Z = \frac{4.9 - 4.6}{\sqrt{\frac{1.1^2}{500} + \frac{1.2^2}{500}}}$$
$$Z = \frac{0.3}{\sqrt{0.00242 + 0.00288}} = \frac{0.3}{0.0728} \approx 4.12$$

Step 3: Decision

Since:

$$|Z| > 1.96$$

the null hypothesis is rejected.

Conclusion

The new recommendation model produces a statistically significant improvement in user engagement.

—

1.9.3 Example 6: AI Model Accuracy Drift Detection

An image classification model historically achieves an average accuracy of 92% with a known standard deviation of 3%.

After deployment, a monitoring system collects $n = 100$ samples and reports an average accuracy of 90.8%.

Test whether the model accuracy has degraded at $\alpha = 0.05$.

Step 1: Hypotheses

$$H_0 : \mu = 92$$

$$H_1 : \mu < 92$$

(This is a one-tailed test.)

Step 2: Test Statistic

$$Z = \frac{90.8 - 92}{3/\sqrt{100}} = \frac{-1.2}{0.3} = -4.0$$

Step 3: Decision

For a one-tailed test at $\alpha = 0.05$, the critical value is -1.645.

Since:

$$Z = -4.0 < -1.645$$

the null hypothesis is rejected.

Conclusion

There is strong evidence that the model accuracy has significantly degraded after deployment.

—

1.9.4 Example 7: Training Time Benchmark Validation

A deep learning framework claims that the average training time per epoch is 18 seconds. The population standard deviation is 2.5 seconds.

A benchmark run of $n = 49$ epochs yields an average training time of 19 seconds.

Step 1: Hypotheses

$$H_0 : \mu = 18$$

$$H_1 : \mu \neq 18$$

Step 2: Test Statistic

$$Z = \frac{19 - 18}{2.5/\sqrt{49}} = \frac{1}{0.357} \approx 2.80$$

Step 3: Decision

Since:

$$|Z| = 2.80 > 1.96$$

the null hypothesis is rejected.

Conclusion

The observed training time is statistically different from the claimed value.

—

1.10 Key Observations from AI Applications

- Z-tests are widely used in AI systems for monitoring, benchmarking, and A/B testing.
- Large sample sizes in production systems make Z-tests particularly appropriate.
- Statistical significance should always be interpreted alongside practical impact and business relevance.

Chapter 2

t-Test for Hypothesis Testing

2.1 Introduction

In statistical inference, hypothesis tests about population means often require knowledge of the population variance. In practice, however, the population variance is rarely known and must be estimated from the sample itself. This additional uncertainty invalidates the direct use of the Z-test, especially for small sample sizes.

The **t-test**, introduced by William Sealy Gosset, addresses this problem by incorporating the variability introduced by estimating the population variance. The test statistic follows a *t-distribution*, rather than a standard normal distribution.

The t-test is one of the most widely used tools in statistics and is fundamental to experimental analysis in machine learning, data science, and artificial intelligence.

2.2 Motivation for the t-Test

Recall the Z-test statistic:

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

The Z-test assumes that the population standard deviation σ is known. When σ is unknown, it is replaced by the sample standard deviation s, leading to the statistic:

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

Since s is a random variable, the resulting statistic no longer follows a normal distribution. Instead, it follows a **t-distribution** with a specified number of degrees of freedom.

2.3 The t-Distribution

The t-distribution is a family of probability distributions indexed by the **degrees of freedom (df)**.

2.3.1 Properties of the t-Distribution

- Symmetric about zero
- Heavier tails than the normal distribution
- Depends on degrees of freedom
- Approaches the standard normal distribution as $df \to \infty$

For a one-sample t-test, the degrees of freedom are:

$$df = n - 1$$

The heavier tails account for the extra uncertainty introduced by estimating the population variance.

2.4 Assumptions of the t-Test

The validity of the t-test relies on the following assumptions:

1. Observations are independent
2. Data are drawn from a normally distributed population, or sample size is sufficiently large
3. Measurements are on an interval or ratio scale

The t-test is robust to moderate deviations from normality, particularly for larger sample sizes.

2.5 One-Sample t-Test

2.5.1 Objective

The one-sample t-test evaluates whether the mean of a population differs from a specified value when the population variance is unknown.

2.5.2 Hypotheses

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

2.5.3 Test Statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where:

- $\bar{x}$ is the sample mean
- s is the sample standard deviation
- n is the sample size

2.5.4 Degrees of Freedom

$$df = n - 1$$

2.5.5 Decision Rule

At significance level α:

- Reject H_0 if $|t| > t_{\alpha/2,df}$
- Otherwise, fail to reject H_0

2.6 Example: One-Sample t-Test

An AI model's training time is claimed to be 50 seconds per epoch. A sample of $n = 16$ epochs yields a mean training time of 53 seconds with a sample standard deviation of 4 seconds. Test the claim at $\alpha = 0.05$.

Solution

$$t = \frac{53 - 50}{4/\sqrt{16}} = \frac{3}{1} = 3.0$$

Degrees of freedom:

$$df = 15$$

Critical value:

$$t_{0.025,15} \approx 2.131$$

Since $|t| > 2.131$, the null hypothesis is rejected.

2.7 Two-Sample t-Test (Independent Samples)

2.7.1 Objective

To compare the means of two independent populations when variances are unknown.

2.7.2 Equal Variance (Pooled) t-Test

This test assumes equal population variances.

Pooled Variance

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Test Statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Degrees of Freedom

$$df = n_1 + n_2 - 2$$

2.7.3 Unequal Variance (Welch's) t-Test

Welch's t-test does not assume equal variances and is preferred in practice.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The degrees of freedom are approximated using the Welch–Satterthwaite equation.

2.8 Example: Two-Sample t-Test

Two machine learning models are compared for accuracy.

- Model A: $\bar{x}_1 = 88$, $s_1 = 5$, $n_1 = 20$
- Model B: $\bar{x}_2 = 84$, $s_2 = 6$, $n_2 = 22$

Using Welch's t-test:

$$t = \frac{88 - 84}{\sqrt{\frac{25}{20} + \frac{36}{22}}} \approx 2.29$$

The difference is statistically significant at $\alpha = 0.05$.

2.9 Paired t-Test

2.9.1 Objective

The paired t-test is used when observations occur in pairs, such as before-and- after measurements or same-user comparisons.

2.9.2 Method

Define differences:

$$d_i = X_i - Y_i$$

Apply a one-sample t-test on the differences:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

2.10 Example: Paired t-Test

An AI model's accuracy is measured before and after fine-tuning on the same dataset.

$$\bar{d} = 1.8, \quad s_d = 1.2, \quad n = 12$$

$$t = \frac{1.8}{1.2/\sqrt{12}} \approx 5.20$$

The improvement is statistically significant.

2.11 Confidence Intervals and the t-Test

A $(1 - \alpha)$ confidence interval for the mean is given by:

$$\bar{x} \pm t_{\alpha/2,df} \frac{s}{\sqrt{n}}$$

Confidence intervals provide more information than hypothesis tests by quantifying the range of plausible parameter values.

2.12 Comparison of Z-Test and t-Test

Feature	Z-Test	t-Test
Variance	Known	Unknown
Sample Size	Large	Small/Any
Distribution	Normal	t
Degrees of Freedom	Not used	Required
Practical Usage	Rare	Very Common

2.13 Applications in AI and Data Science

- Model performance comparison
- Offline experiment analysis

- Feature impact studies
- A/B testing with limited samples
- Model drift detection

2.14 Summary

The t-test extends the Z-test by accounting for uncertainty in variance estimation. It is the default choice for hypothesis testing about means in real-world applications. Understanding the t-test is essential for rigorous experimental analysis in artificial intelligence and machine learning.

2.15 Types of t-Tests and Worked Examples

The **t-test** is a family of statistical hypothesis tests used to determine whether there is a significant difference between means. They are widely used when the population standard deviations are unknown and sample sizes are small. :contentReference[oaicite:1]index=1

2.15.1 One-Sample t-Test

Used to determine whether the mean of a single sample differs from a known or hypothesized population mean.

Formulation

$$H_0 : \mu = \mu_0, \quad H_1 : \mu \neq \mu_0$$

Test statistic:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}},$$

with degrees of freedom $df = n - 1$. :contentReference[oaicite:2]index=2

Example

A data scientist measures the average prediction latency of an AI model on a dataset and obtains:

$$\bar{x} = 110 \text{ ms}, \quad s = 10 \text{ ms}, \quad n = 20,$$

and hypothesizes $\mu_0 = 100$ ms.

Compute the test statistic:

$$t = \frac{110 - 100}{10/\sqrt{20}} = \frac{10}{10/\sqrt{20}} \approx 4.47.$$

For $\alpha = 0.05$ (two-tailed) with $df = 19$, the critical value $t_{0.025,19} \approx 2.093$. Since $|t| > 2.093$, reject H_0. There is significant evidence that the model latency differs from 100 ms.

2.15.2 Two-Sample (Independent) t-Test

Used to compare means from two independent groups. :contentReference[oaicite:3]index=3

Formulation

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 \neq \mu_2$$

When variances are assumed equal, the pooled estimate is:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$

and

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

with $df = n_1 + n_2 - 2$. When variances differ, the Welch's t-test is used. :contentReference[oaicite:4]index=4

Example

Two development teams measure model accuracy:

Group A:

$$\bar{x}_1 = 0.85, \quad s_1 = 0.03, \quad n_1 = 15$$

Group B:

$$\bar{x}_2 = 0.80, \quad s_2 = 0.04, \quad n_2 = 15$$

Assume unequal variances. The Welch t-statistic:

$$t = \frac{0.85 - 0.80}{\sqrt{\frac{0.03^2}{15} + \frac{0.04^2}{15}}} = \frac{0.05}{\sqrt{0.00006 + 0.000107}} \approx \frac{0.05}{0.01345} \approx 3.72.$$

Degrees of freedom are approximated using the Welch–Satterthwaite formula (not shown), and for large enough samples the critical two-tailed value at $\alpha = 0.05$ is near $t \approx 2.04$. Since $|t| > 2.04$, reject H_0.

2.15.3 Paired t-Test

Used when samples are related (e.g., before–after measurements) and each observation in one sample pairs with one in the other. :contentReference[oaicite:5]index=5

Formulation

Given difference scores:

$$d_i = X_i - Y_i,$$

test:

$$H_0 : \mu_d = 0, \quad H_1 : \mu_d \neq 0,$$

with:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}, \quad df = n - 1.$$

Example

A model's performance is measured before and after a calibration routine on 10 datasets. Mean difference: $\bar{d} = 2.5$, standard deviation of differences: $s_d = 1.8$, $n = 10$:

$$t = \frac{2.5}{1.8/\sqrt{10}} \approx \frac{2.5}{0.569} \approx 4.39.$$

For $\alpha = 0.05$, $df = 9$, the critical value is $t_{0.025,9} \approx 2.262$. Since $4.39 > 2.262$, reject H_0. The calibration step significantly affects performance.

2.15.4 Summary of t-Test Types

Test	**Purpose**	**When used**
One-Sample t-test	Compare sample mean to known value	Single sample
Two-Sample t-test	Compare means of two independent groups	Independent samples
Paired t-test	Compare means of related samples	Paired/related data