Module III: Statistical Inference 1

Topic 1: Introduction to Statistical Inference, Sampling Distribution, Standard Error, Testing of Hypothesis, Levels of Significance, Test of Significance, and Confidence Limits

Outline

1. Sampling in Statistics
2. Statistical Inference
3. Sampling Distribution
4. Standard Error
5. Hypothesis Testing
6. Type I, Type II Errors and Levels of Significance
7. Confidence Interval

Sampling in Statistics

- **Sampling** is a statistical method of obtaining representative data (observations) from a group. We often use sampling concepts in everyday life without realizing it. For example, Checking the quality of rice by taking a handful is an example of random sampling from a large population.
- **Population (Universe):**The group of objects (or individuals) under study is called the population or universe. A population can be either Finite or Infinite.
- **Sample:** A part of the population that contains selected objects (or individuals) is called a sample. The number of individuals in a sample is called the **sample size**.
- **Random sampling** is the selection of objects (individuals) from the population in such a way that each object has the same chance of being selected. The lottery system is a common example of random sampling.
- **Simple sampling** is a special case of random sampling.Each event in

Statistical Inference

A Statistical inference is the process of drawing conclusions about populations based on samples. It involves two main activities:

(i). Estimation: Determining population parameters from sample statistics. Estimating a single value for a population parameter is called a point estimation. Providing a range of values for a population parameter is called an interval estimation.

(ii). Hypothesis testing: Making decisions about populations based on sample data

Key concepts:

- **Population:** The entire group we want to study
- **Sample:** A subset of the population
- **Parameter:** A characteristic of the population (e.g., mean, variance)
- **Statistic:** An estimate of the parameter based on the sample.

Population

(a) Entire group of interest
(b) Often too large to study entirely
(c) Described by parameters (e.g., μ, σ)

Sample

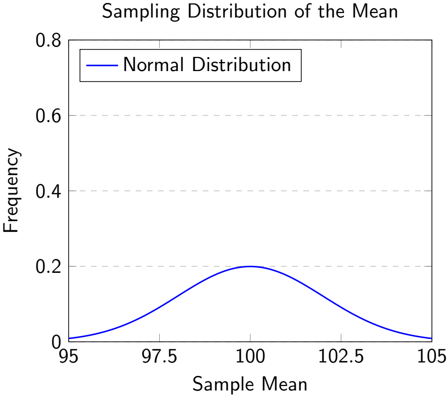
(a) Subset of the population
(b) Used to make inferences
(c) Described by statistics (e.g., $\bar{x}$, s)

Example: In AI/ML: We often use statistical inference to understand and make predictions about large datasets. Predicting house prices based on features like size, location, etc.

- **Population:** All houses in a city
- **Sample:** Dataset of houses with known prices and features
- **Parameter:** True relationship between features and price
- **Statistic:** Estimated relationship from our model (e.g., coefficients in linear regression)
- **Inference:** Using the model to predict prices of new houses

Sampling Distribution

A sampling distribution is the distribution of a statistic over many samples. It describes the variability of the statistic. Most commonly used: sampling distribution of the sample mean ($\bar{X}$).

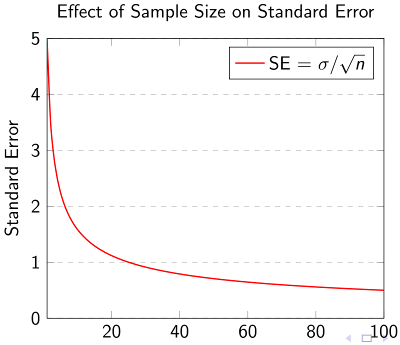


- Properties:
  - Center: Expected value of the statistic

Standard Error

i. **The Standard Error (SE)** is the standard deviation of a sampling distribution. The most commonly used one is Standard Error of the Mean.

ii. **Formula:** $SE = \frac{\sigma}{\sqrt{n}}$, where σ is the population standard deviation and n is the sample size. If σ is unknown, we estimate it with the sample standard deviation s. The reciprocal of standard error is called the **precision**.



Hypothesis Testing

- Hypothesis testing is a statistical method used to make inferences about a population based on sample data. It involves making an assumption (hypothesis) about a population parameter and then testing this assumption using sample data.
- **Key components:**
  - Null hypothesis (H_0): The assumption we start with
  - Alternative hypothesis (H_1 or H_a): The competing claim
  - Test statistic: A value calculated from the sample data
  - Decision rule: Criteria for rejecting or failing to reject H_0

Steps in Hypothesis Testing:

1. State the null and alternative hypotheses
2. Choose a significance level (α)
3. Select the appropriate test statistic
4. Determine the critical region
5. Calculate the test statistic from sample data
6. Make a decision: Reject H_0 or fail to reject H_0
7. Interpret the results

Type I, Type II Errors and Levels of Significance

- **Type I Error** (False Positive): Rejecting the null hypothesis when it's actually true.Probability $= \alpha$ (significance level)
- **Type II Error** (False Negative): Failing to reject the null hypothesis when it's actually false. Probability $= \beta$
- Power of a test = 1 - β (probability of correctly rejecting a false null hypothesis)
- **The level of significance (α)** is the probability of rejecting the null hypothesis when it is actually true. It represents the risk of making a Type I error.
- **Common levels:** 0.05 (5%), 0.01 (1%), 0.10 (10%)
- Smaller α means Lower risk of Type I error and Higher risk of Type II error (failing to reject a false null hypothesis)
- **Choosing the Level of Significance** is depends on the nature of the problem & consequences of errors. For (i) Critical applications like healthcare diagnosis: Lower α (0.01 or 0.001) & (ii) Exploratory Data Analysis: Higher α (0.05 or 0.10)
- Decreasing α typically increases β

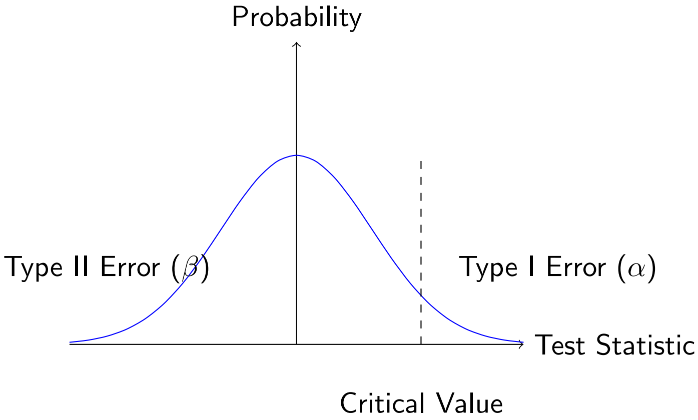


Figure: Relationship between significance level and errors

Confidence Interval

- A confidence interval is a range of values that is likely to contain the true population parameter
- It quantifies the uncertainty in our point estimate
- Interpretation: If we repeated the sampling process many times, the true parameter would be within the interval in X% of the cases
- Common confidence levels: 95%, 99%, 90%
- General form: Point Estimate ± (Critical Value × Standard Error)
- For a population mean (large sample):

$$\bar{X} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

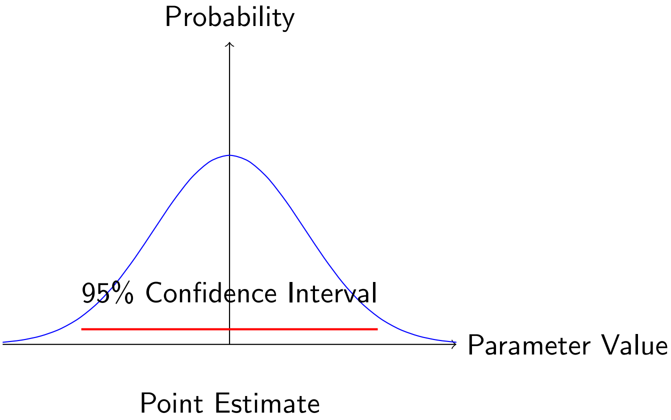


Figure: 95% Confidence Interval

Question 1: Explain the following terms:
(i).Standard Error
(ii). Statistical Hypothesis
(iii). Critical Region of a Statistical Test
(iv). Test of Significance

Answer:

i. Standard Error: The standard deviation of the sampling distribution of a statistic, usually the mean. It measures the accuracy with which a sample represents the population.

ii. Statistical Hypothesis: A statement about a population parameter that can be tested using statistical methods. Common hypotheses include null (H_0) and alternate (H_1).

iii. Critical Region of a Statistical Test: The range of values for which the null hypothesis is rejected. If the test statistic falls in this region, it indicates that the result is statistically significant.

iv. Test of Significance: A method to determine if the observed data provide enough evidence to reject a null hypothesis. Common tests include the Z-test, t-test, and chi-square test.

Question 2: Define the following terms:
(i).Alternate Hypothesis
(ii). A Statistic
(iii). Level of Significance
(iv). Two-Tailed Test

Answer:

i. Alternate Hypothesis (H_1): A hypothesis that proposes a change or difference from the null hypothesis. It represents the conclusion that is accepted if the null hypothesis is rejected.

ii. A Statistic: A quantity calculated from sample data, used to estimate a population parameter. Examples: sample mean, sample variance.

iii. Level of Significance (α): The probability of rejecting the null hypothesis when it is actually true (Type I error). Common values are 0.05 (5%) and 0.01 (1%).

iv. Two-Tailed Test: A test of significance where the critical region is in both tails of the probability distribution. It checks for deviation in either direction from the hypothesized value.

Hypothesis Testing problems

(Problems based on Binomial distributions and Proportions)

Hypothesis Testing problems based on Binomial distribution

The binomial distribution can be approximated by the normal distribution when the sample size is large. Normal approximation is valid if $np \geq 5$, where p is the probability of success.

Steps of Hypothesis Testing:

1. **State the Hypotheses:**
   $H_0 : p = p_0$ (Null Hypothesis)
   $H_1 : p \neq p_0$ (Alternative Hypothesis) for a two-tailed test.
2. **Choose the Significance Level:**
   Common values: $\alpha = 0.05$ or $\alpha = 0.01$.
3. **Calculate the Test Statistic:**
   Use the formula $z = \frac{x-np}{\sqrt{npq)}}$.
4. **Determine the Critical Value :**
   For a two-tailed test, compare z with $\pm z_{\alpha/2}$ (e.g., ± 1.96 for $\alpha = 0.05$).
5. **Make a Decision:**

Problem 1: A coin was tossed 400 times, and the head turned up 216 times. Test the hypothesis that the coin is unbiased at a 5% level of significance.

Solution:

Step 1:

Null hypothesis (H_0): The coin is unbiased ($p = 0.5$).

Alternative hypothesis (H_1): The coin is biased ($p \neq 0.5$).

Step 2: Expected number of heads

$$E(\text{heads}) = \frac{1}{2} \times 400 = 200 = np$$

Observed number of tails $= 216$.

Step 3: Standard Deviation

$$\text{S.D.} = \sqrt{npq} = \sqrt{400 \times \frac{1}{2} \times \frac{1}{2}} = \sqrt{100} = 10$$

Step 4:
The z-test statistic formula is:

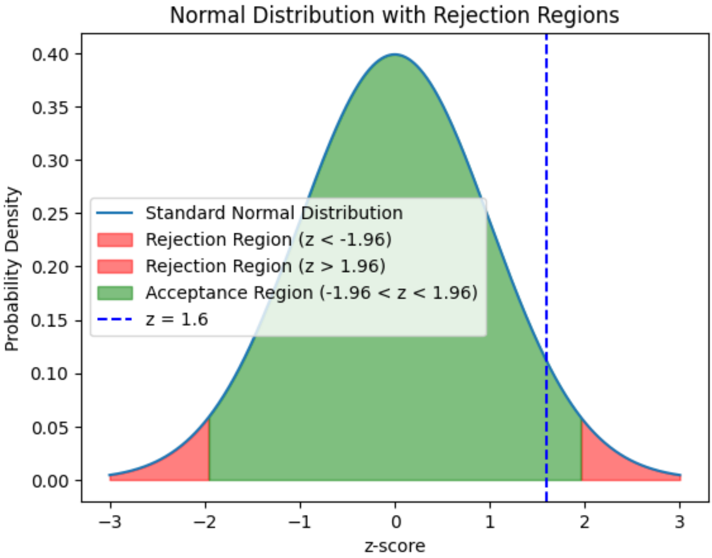
$$z = \frac{x - np}{\sqrt{npq}}$$

Substituting the values:

$$z = \frac{216 - 200}{10} = \frac{16}{10} = 1.6$$

Step 5:
At the 5% level of significance, the critical value for a two-tailed test is $z = 1.96$. Since $z = 1.6$ is less than 1.96, we fail to reject the null hypothesis.

Conclusion: The coin is likely **unbiased**.



The critical values at a 5% significance level are $z = -1.96$ and $z = 1.96$. Our calculated z-score (1.6) lies in the acceptance region.

Problem 2: A coin was tossed 1600 times, and the tail turned up 864 times. Test the hypothesis that the coin is unbiased at a 1% level of significance.

Solution:

Step 1:

Null hypothesis (H_0): The coin is unbiased ($p = 0.5$).

Alternative hypothesis (H_1): The coin is biased ($p \neq 0.5$).

Step 2: Expected number of tails

$$E(\text{tails}) = \frac{1}{2} \times 1600 = 800 = np$$

Observed number of tails $= 864$.

Step 3:

$$\textbf{S.D.} = \sqrt{npq} = \sqrt{1600 \times \frac{1}{2} \times \frac{1}{2}} = \sqrt{400} = 20$$

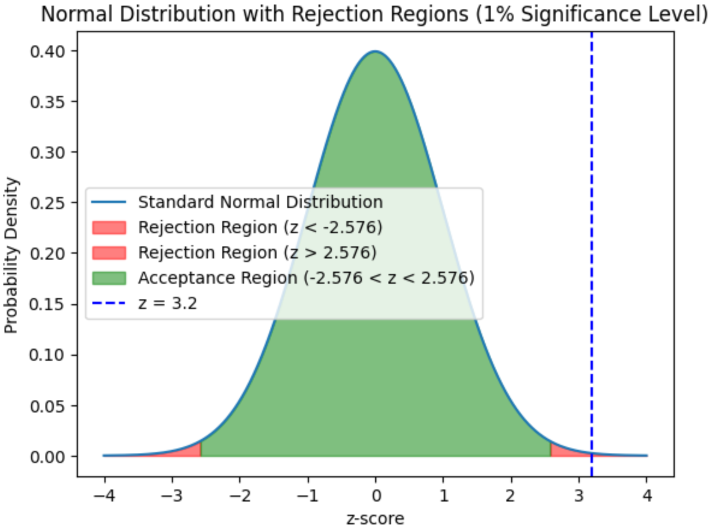
Step 4:
The z-test statistic formula is:

$$z = \frac{x - np}{\sqrt{npq}}$$

Substituting the values:

$$z = \frac{864 - 800}{20} = \frac{64}{20} = 3.2$$

Step 5:
At the 1% level of significance, the critical value for a two-tailed test is $z = 2.576$. Since $z = 3.2$ is greater than 2.576, we reject the null hypothesis.
Conclusion: The coin is **biased** at the 1% significance level.



The critical values at a 1% significance level are $z = -2.576$ and $z = 2.576$. Our calculated z-score (3.2) lies in the rejection region.

Problem 3: In 324 throws of a six-faced die, an odd number turned up 181 times. Test the hypothesis that the die is unbiased at the 1% level of significance.

Solution:

Step 1:

Null hypothesis (H_0): The die is unbiased ($p = 0.5$), i.e., the probability of getting an odd number (1, 3, 5) is the same as the probability of getting an even number (2, 4, 6).

Alternative hypothesis (H_1): The die is biased ($p \neq 0.5$).

Step 2:Expected Number of 3's or 4's

The probability of getting an odd number (1, 3, or 5) on a fair die is:

$$P(\text{odd number}) = \frac{3}{6} = 0.5$$

Hence, the expected number of odd numbers in 324 throws is:

$$E(\text{odd numbers}) = 0.5 \times 324 = 162 = np$$

Observed number of odd numbers = 181.

Step 3:

The standard deviation is calculated using the formula for binomial distribution:

$$\text{S.D.} = \sqrt{npq}$$

$$\therefore \text{S.D.} = \sqrt{324 \times 0.5 \times 0.5} = \sqrt{81} = 9$$

Step 4:
The z-test statistic formula is:

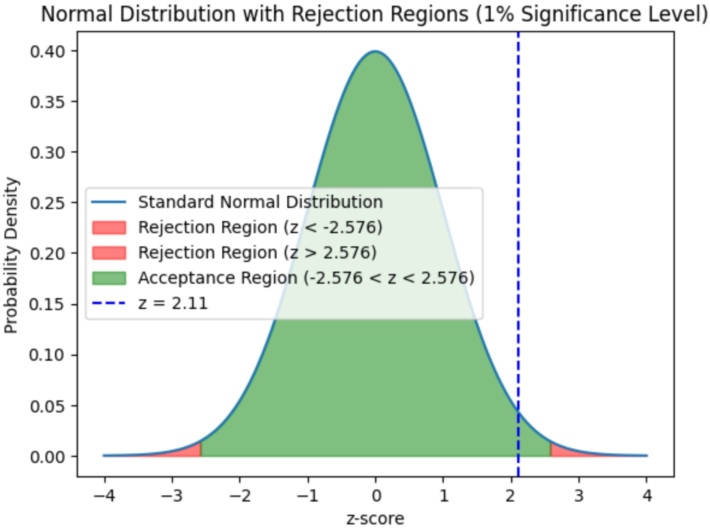
$$z = \frac{x - np}{\sqrt{npq}}$$

Substituting the values:

$$z = \frac{181 - 162}{9} = \frac{19}{9} = 2.11$$

Step 5: At the 1% level of significance for a two-tailed test, the critical value is $z = 2.576$.Since $z = 2.11$ is less than 2.576, we fail to reject the null hypothesis.

Conclusion: The die is **not significantly biased** at the 1% significance level.



The critical values at a 1% significance level are $z = -2.576$ and $z = 2.576$. Our calculated z-score (2.11) lies within the acceptance region.

Problem 4: A die is thrown 9000 times, and a throw of 3 or 4 was observed 3240 times. Test whether the die can be regarded as unbiased.

Solution:

Step 1:

Null hypothesis (H_0): The die is unbiased ($p = \frac{2}{6} = \frac{1}{3}$), i.e., the probability of getting a 3 or 4 is $1/3$.

Alternative hypothesis (H_1): The die is biased ($p \neq \frac{1}{3}$).

Step 2: Expected Number of 3's or 4's

The probability of getting a 3 or 4 on a fair die is:

$$P(3 \text{ or } 4) = \frac{2}{6} = \frac{1}{3}$$

Hence, the expected number of 3's or 4's in 9000 throws is:

$$E(3 \text{ or } 4) = \frac{1}{3} \times 9000 = 3000$$

Observed number of 3's or 4's $= 3240$.

Step 3:
The standard deviation is calculated using the formula for binomial distribution:

$$\text{S.D.} = \sqrt{npq}$$

$$\therefore \text{S.D.} = \sqrt{9000 \times \frac{1}{3} \times \frac{2}{3}} = \sqrt{2000} \approx 44.72$$

Step 4:
The z-test statistic formula is:

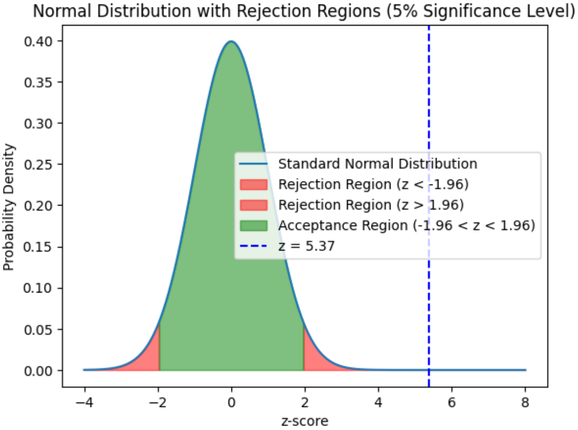
$$z = \frac{x - np}{\sqrt{npq}}$$

Substituting the values:

$$z = \frac{3240 - 3000}{44.72} = \frac{240}{44.72} \approx 5.37$$

Step 5: At the 5% level of significance for a two-tailed test, the critical value is $z = 1.96$. Since $z = 5.37$ is much greater than 1.96, we reject the null hypothesis.

Conclusion: The die is significantly **biased**.



The critical values at a 5% significance level are $z = -1.96$ and $z = 1.96$. Our calculated z-score (5.37) lies far in the rejection region.

Hypothesis test problems based on Proportions

A hypothesis test for a single proportion is used to determine if the sample proportion differs significantly from a hypothesized proportion in the population.

Step 1:

Null Hypothesis (H_0): The population proportion is equal to the hypothesized proportion. $H_0 : p = p_0$

Alternative Hypothesis (H_a): The population proportion is different from the hypothesized proportion. $H_a : p \neq p_0$

Step 2: Choose Significance Level (α) as 0.05 or 0.01.

Step 3: Formula for the test statistic Z:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

$\hat{p}$: Sample proportion, p_0: Hypothesized proportion, n: Sample size **Step 4:** Compare the calculated Z to the critical value(s) and draw a conclusion.

Problem 5: A coin is tossed 400 times, and it turns up heads 216 times. Test whether the coin may be regarded as an unbiased one at the 5% significance level.

Solution:

Step 1: Null Hypothesis (H_0):

The coin is unbiased, meaning the proportion of heads is 0.5.

$$H_0 : p = 0.5$$

Alternative Hypothesis (H_1):

The coin is biased, meaning the proportion of heads is not equal to 0.5.

$$H_1 : p \neq 0.5$$

This is a two-tailed test.

Step 2:

Observed Proportion of Heads:

$$\hat{p} = \frac{216}{400} = 0.54$$

Expected Proportion under H_0:

$$p_0 = 0.5$$

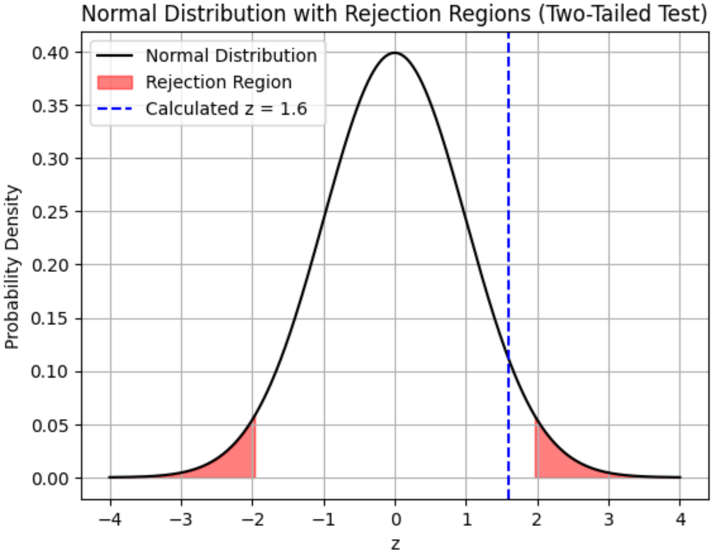
Step 3 Test Statistic:

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

Substituting the values:

$$z = \frac{0.54 - 0.5}{\sqrt{\frac{0.5 \times 0.5}{400}}} = \frac{0.04}{\sqrt{\frac{0.25}{400}}} = \frac{0.04}{0.025} = 1.6$$

Step 4:
For a two-tailed test at the 5% significance level, the critical values are $z = \pm 1.96$ and the calculated z-value is 1.6, which is within the acceptance region $(-1.96, 1.96)$.
Conclusion: Since the calculated z-value does not exceed the critical value, we **fail to reject** the null hypothesis. Therefore, there is no evidence to suggest the coin is biased.



The shaded areas represent the rejection regions for a two-tailed test with $\alpha = 0.05$. The calculated $z = 1.6$ falls within the acceptance region.

Problem 6: A coin is tossed 1600 times, and tails turn up 864 times. Test the hypothesis that the coin is unbiased at a 1% level of significance.

Step 1: Null Hypothesis (H_0):

The coin is unbiased, meaning the proportion of tails is 0.5.

$$H_0 : p = 0.5$$

Alternative Hypothesis (H_1):

The coin is biased, meaning the proportion of tails is not equal to 0.5.

$$H_1 : p \neq 0.5$$

This is a two-tailed test.

Step 2: Significance level $(\alpha) = 0.01$.

Step 3: Observed Proportion of Tails:

$$\hat{p} = \frac{864}{1600} = 0.54$$

Expected Proportion under H_0:

$$p_0 = 0.5$$

Step 4: Test Statistic is

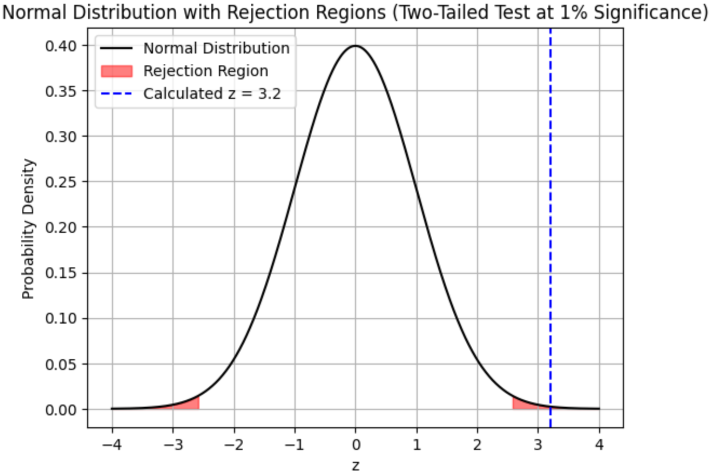
$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Substituting the values:

$$z = \frac{0.54 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{1600}}} = \frac{0.04}{\sqrt{\frac{0.25}{1600}}} = \frac{0.04}{0.0125} = 3.2$$

Step 5: For a two-tailed test at the 1% significance level, the critical values are $z = \pm 2.575$. The calculated z-value is 3.2, which lies in the rejection region.

Conclusion: Since the calculated $z = 3.2$ exceeds the critical value 2.575, we reject the null hypothesis. Therefore, the coin is not unbiased.



The shaded areas represent the rejection regions for a two-tailed test with $\alpha = 0.01$. The calculated $z = 3.2$ falls within the rejection region.

Problem 7: In 324 throws of a six-faced die, an odd number turned up 181 times. Test the hypothesis that the die is unbiased at a 1% level of significance.

Solution:

Step 1:

Null Hypothesis (H_0): The die is unbiased, meaning the proportion of odd numbers is 0.5.

$$H_0 : p = 0.5$$

Alternative Hypothesis (H_1): The die is biased, meaning the proportion of odd numbers is not equal to 0.5.

$$H_1 : p \neq 0.5$$

Step 2: Significance level $(\alpha) = 0.01$.

Step 3: Observed Proportion of Odd Numbers:

$$\hat{p} = \frac{181}{324} = 0.5586$$

Expected Proportion under H_0:

$$p_0 = 0.5$$

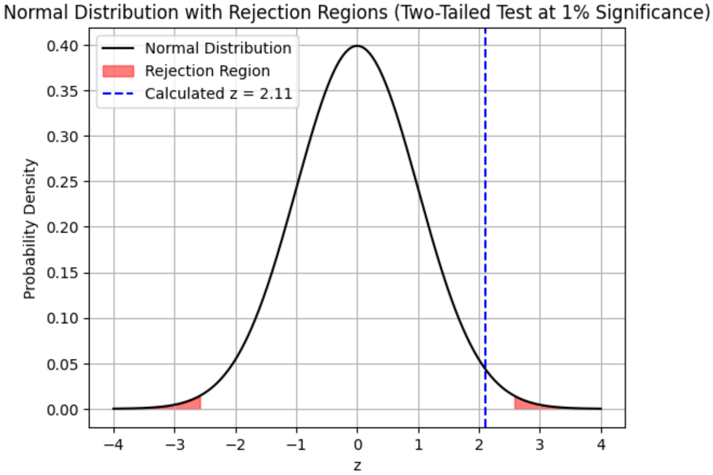
Step 4: The Test Statistic is

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Substituting the values:

$$z = \frac{0.5586 - 0.5}{\sqrt{\frac{0.5(1-0.5)}{324}}} = \frac{0.0586}{\sqrt{\frac{0.25}{324}}} = \frac{0.0586}{0.0278} \approx 2.11$$

Step 5: For a two-tailed test at the 1% significance level, the critical values are $z = \pm 2.575$. The calculated z-value is approximately 2.11, which does not lie in the rejection region.
Conclusion: Since the calculated $z = 2.11$ is less than the critical value 2.575, we fail to reject the null hypothesis. Therefore, there is insufficient evidence to conclude that the die is biased.



The shaded areas represent the rejection regions for a two-tailed test with $\alpha = 0.01$. The calculated $z = 2.11$ falls within the acceptance region.

Problem 8 A die is thrown 9000 times, and a throw of 3 or 4 was observed 3240 times. Test whether the die can be regarded as unbiased using a hypothesis test for proportions.

Solution:

Step 1:

Null Hypothesis (H_0): The die is unbiased, meaning the proportion of throws resulting in 3 or 4 is $\frac{2}{6} = \frac{1}{3}$.

$$H_0 : p = \frac{1}{3}$$

Alternative Hypothesis (H_1): The die is biased, meaning the proportion of throws resulting in 3 or 4 is not equal to $\frac{1}{3}$.

$$H_1 : p \neq \frac{1}{3}$$

Step 2:

We will conduct the test at the 5% significance level ($\alpha = 0.05$). This means there is a 5% risk of rejecting the null hypothesis when it is true.

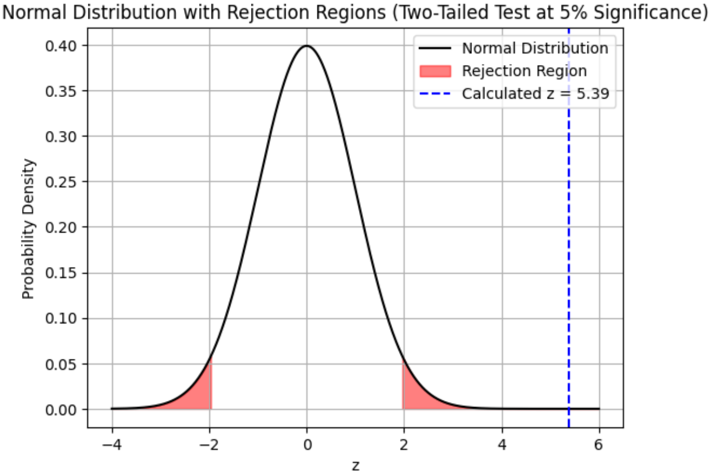
Step 3: Observed Proportion of Throws with 3 or 4: $\hat{p} = \frac{3240}{9000} = 0.36$
Expected Proportion under H_0: $p_0 = \frac{1}{3} = 0.3333$
Step 4: The Test Statistic is

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Substituting the values:

$$z = \frac{0.36 - 0.3333}{\sqrt{\frac{0.3333(1-0.3333)}{9000}}} = \frac{0.0267}{\sqrt{\frac{0.2222}{9000}}} = \frac{0.0267}{0.00495} \approx 5.39$$

Step 5: For a two-tailed test at the 5% significance level, the critical values are $z = \pm 1.96$. The calculated z-value is approximately 5.39, which lies in the rejection region.
Conclusion: Since the calculated $z = 5.39$ is greater than the critical value 1.96, we reject the null hypothesis. Therefore, we conclude that the die is biased.



The shaded areas represent the rejection regions for a two-tailed test with $\alpha = 0.05$. The calculated $z = 5.39$ falls within the rejection region.

Assignment Problems

1. In a city, a sample of 500 people is taken, out of which 280 are tea drinkers, and the rest are coffee drinkers. Can we assume that both coffee and tea are equally popular in this city at a 5% level of significance?

2. A manufacturing company claims that at least 95% of its products supplied conform to the specifications. Out of a sample of 200 products, 18 are found to be defective. Test the claim at a 5% level of significance.

Sampling and Significance Tests

(Simple sampling of attributes. Test of significance for large samples, comparison of large samples)

Simple Sampling of Attributes

- **Simple Sampling:** A sampling method where every individual or attribute in the population has an equal chance of being selected.This method ensures a random and unbiased sample, essential for statistical inference in AI and ML models.
- **Scenario in AI/ML:** When working with datasets, simple sampling can be used to create training and test sets. This helps ensure that the model generalizes well to unseen data.

  For **example**, If we want to sample 100 data points from a dataset of 10,000, we use simple random sampling to ensure all data points have an equal chance of being selected.

Test of Significance for Large Samples

- **Test of significance:** Used to determine whether the observed data significantly deviates from what is expected under the null hypothesis.For large samples, when the sample size $n > 30$, we can apply the Z-test for large samples.
- **Z-test:**

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \tag{1}$$

  where: $\bar{x}$ is the sample mean, μ is the population mean, σ is the standard deviation, n is the sample size.
- **Critical value:** Compare the Z-value to the critical value from the Z-table at a given significance level (e.g., $\alpha = 0.05$).
- **Scenario in AI/ML:** Significance can be used for feature selection or model comparison. Comparing two model predictions to see if one model is significantly better than another. Testing whether a feature Significantly affects the output of a model.

Large Samples

- When comparing two large samples, we test if the means of two populations are significantly different.
- **Hypotheses:**
  - Null Hypothesis H_0: The two sample means are equal.
  - Alternate Hypothesis H_1: The two sample means are different.
- **Two-sample Z-test formula:**

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

  where, $\bar{x}_1, \bar{x}_2$ are the sample means, σ_1, σ_2 are the population standard deviations, n_1, n_2 are the sample sizes.
- **Scenario:** We have two datasets with different feature values, and we want to test whether their mean feature values differ significantly.

Test of Significance of Difference between Two Sample Proportions

- Used to test whether two population proportions are significantly different.
- Null Hypothesis (H_0): $P_1 = P_2$

  Alternate Hypothesis (H_1): $P_1 \neq P_2$
- Test statistic:

$$Z = \frac{p_1 - p_2}{\sqrt{P(1-P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{3}$$

  where: p_1 and p_2 are the sample proportions, $P = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$ is the pooled sample proportion,n_1 and n_2 are the sample sizes.
- Compare the Z-value to the critical value at the chosen significance level.
- **Scenario in AI / ML:** Testing whether the proportion of successful outcomes in two models is significantly different.

Problem 1: In an examination, the mean grade of students across various schools was 74.5 with a standard deviation of 8. At one particular school, 200 students took the exam, and their mean grade was 75.9. Test the significance of this result at the 5% and 1% significance levels.

Solution:

Step 1: Hypotheses

Null hypothesis: $H_0 : \mu = 74.5$

Alternative hypothesis: $H_A : \mu \neq 74.5$ (two-tailed test)

Step 2: $\mu = 74.5, \sigma = 8$, $\bar{x} = 75.9$, $n = 200$

Step 3: Z-test formula

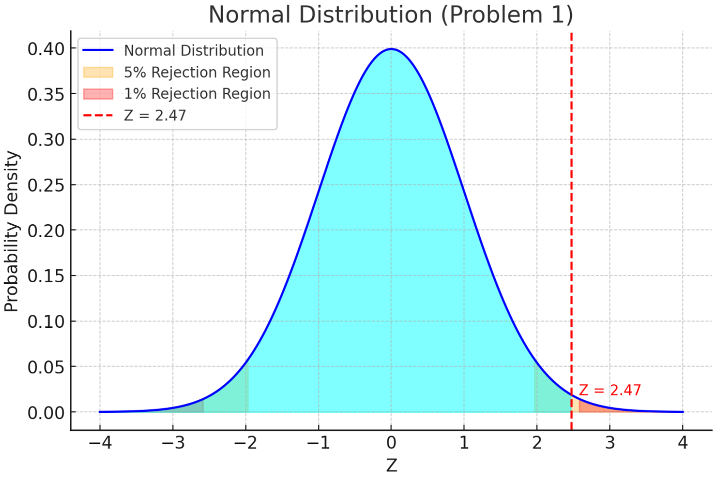
$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{75.9 - 74.5}{\frac{8}{\sqrt{200}}} \approx 2.47$$

Step 4: Compare Z-value

- Critical Z-value at 5%: $Z = \pm 1.96$
- Critical Z-value at 1%: $Z = \pm 2.58$

Conclusion:

- Significant at 5% level, as $Z = 2.47$ exceeds 1.96.
- Not significant at 1% level, as $Z = 2.47$ is less than 2.58.


Normal Distribution (Problem 1)
Normal Distribution
5% Rejection Region
1% Rejection Region
Z = 2.47
Probability Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Z = 2.47
−4
−3
−2
−1
0
1
2
3
4
Z

Problem 2: In a coding competition, the national mean score was 82 with a standard deviation of 10. At a particular university, 150 students participated, and their mean score was 85. Is the difference significant at 5% and 1% significance levels?

Solution:

Step 1:

Null hypothesis: $H_0 : \mu = 82$

Alternative hypothesis: $H_A : \mu \neq 82$ (two-tailed test)

Step 2: $\mu = 82, \sigma = 10$, $\bar{x} = 85$, $n = 150$

Step 3: Z-test formula

$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{85 - 82}{\frac{10}{\sqrt{150}}} \approx 3.67$$

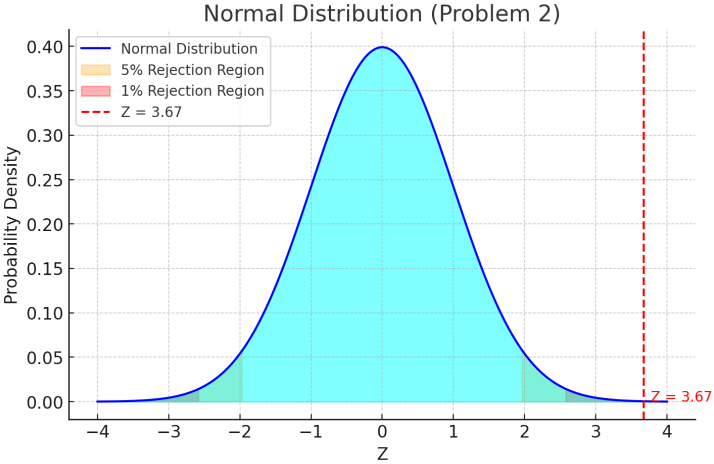
Step 4:

Critical Z-value at 5%: $Z = \pm 1.96$

Critical Z-value at 1%: $Z = \pm 2.58$

Conclusion:

Significant at both 5% and 1% levels, as $Z = 3.67$ exceeds both critical values.


Normal Distribution (Problem 2)
Normal Distribution
5% Rejection Region
1% Rejection Region
Z = 3.67
Probability Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−4
−3
−2
−1
0
1
2
3
4
Z
Z = 3.67

Problem 3: Intelligence tests were given to two groups of boys and girls. Their respective means, standard deviations, and sample sizes are given below:

Group	Mean ($\bar{x}$)	Standard Deviation (σ)	Sample Size (n)
Girls	75	8	60
Boys	73	10	100

We need to determine if the two means significantly differ at the 5% level of significance.

Solution:

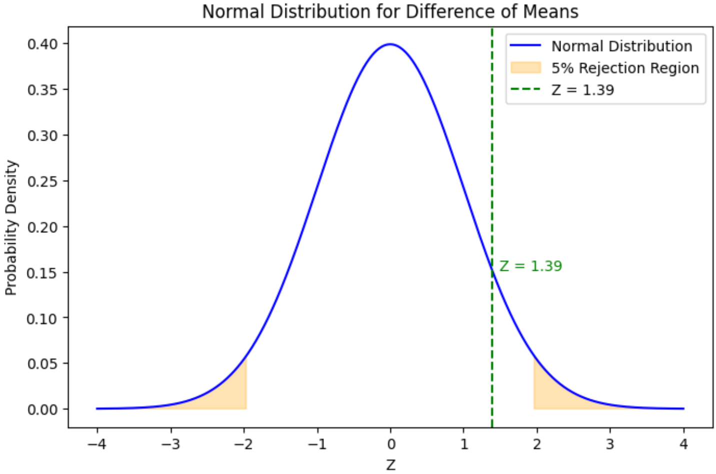
Step 1: Null hypothesis (H_0): $\mu_1 = \mu_2$

Alternative hypothesis (H_A): $\mu_1 \neq \mu_2$ (two-tailed test)

Step 2: $\bar{x}_1 = 75$, $\bar{x}_1 = 73$, $\sigma_1 = 8$, $\sigma_2 = 10$, $n_1 = 60$, $n_2 = 100$

Step 3: Z-test formula for difference of means:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(75 - 73)}{\sqrt{\frac{8^2}{60} + \frac{10^2}{100}}} \approx 1.39$$



Since the calculated Z-value (1.39) is less than 1.96, we **fail to reject the null hypothesis** at the 5% level. Hence, there is no significant difference between the means of boys and girls.

Problem 4: A group of researchers tested two machine learning algorithms (Algorithm A and Algorithm B) on a benchmark dataset. The results of their accuracy scores are as follows:

Algorithm	Mean ($\bar{x}$)	Standard Deviation (σ)	Sample Size (n)
Algorithm A	85	5	50
Algorithm B	82	6	80

Using a significance level of 5%, determine if there is a significant difference in the mean accuracy scores of the two algorithms.

Solution:

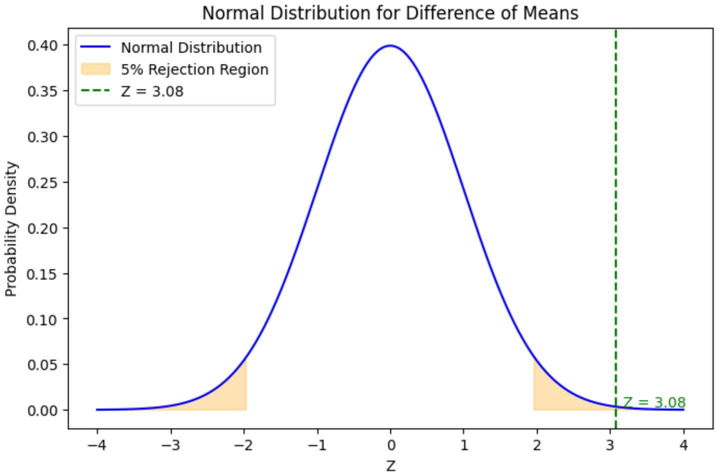
Step 1: Null hypothesis (H_0): $\mu_A = \mu_B$

Alternative hypothesis (H_A): $\mu_A \neq \mu_B$ (two-tailed test)

Step 2: $\bar{x}_A = 85$, $\bar{x}_B = 82$, $\sigma_A = 5$, $\sigma_B = 6$, $n_A = 50$, $n_B = 80$

Step 3: Z-test formula for difference of means:

$$Z = \frac{(\bar{x}_A - \bar{x}_B)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} \frac{(85 - 82)}{\sqrt{\frac{5^2}{50} + \frac{6^2}{80}}} \approx 3.08$$



Since the calculated Z-value (3.08) is greater than 1.96, we **reject the null hypothesis** at the 5% level. Hence, there is a significant difference between the mean accuracy scores of Algorithm A and Algorithm B.

Hypothesis Testing problems - II

Difference of Two-Proportion

Problem 1: A sample of 300 units of a manufactured product contains 65 defective units. In another sample of 200 units, 35 units were found defective. At the 5% level of significance, we want to test if there is a significant difference in the proportion of defectives between the two samples.

Solution:

Step 1:

Null hypothesis (H_0): $H_0 : p_1 = p_2$

Alternative hypothesis (H_1): $H_1 : p_1 \neq p_2$

Step 2: The sample proportions are calculated as follows:

$$\hat{p}_1 = \frac{65}{300} = 0.2167$$

$$\hat{p}_2 = \frac{35}{200} = 0.175$$

Step 3: The pooled proportion (P) is the combined proportion of defectives from both samples:

$$P = \frac{x_1 + x_2}{n_1 + n_2} = \frac{65 + 35}{300 + 200} = \frac{100}{500} = 0.2$$

Step 4: The Z-test statistic is calculated using the formula:

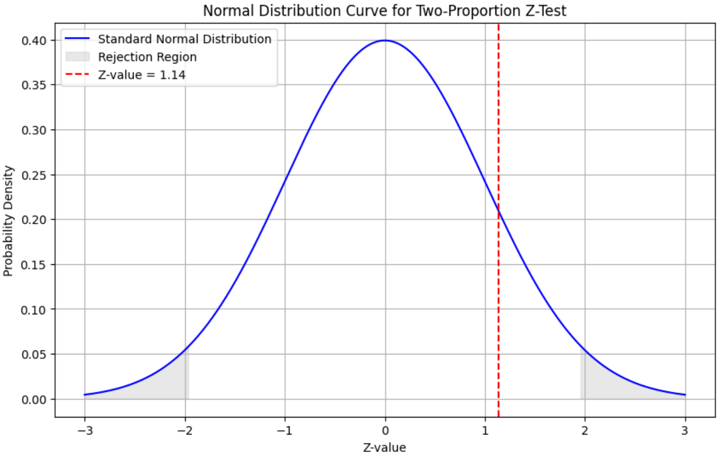
$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{P(1-P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Substituting the values:

$$z = \frac{(0.2167 - 0.175)}{\sqrt{0.2(1-0.2)\left(\frac{1}{300} + \frac{1}{200}\right)}} \approx 1.14$$

For a two-tailed test at the 5% level of significance, the critical z-value is ±1.96. Since the calculated z-value (1.14) is less than the critical value (1.96), we fail to reject the null hypothesis.

Conclusion: There is no significant difference in the proportions of defectives in the two samples at the 5% level of significance.



The shaded areas represent the rejection regions. The red dashed line marks the calculated z-value.

Problem 2: In a large city A, 20% of a random sample of 900 school boys had a slight physical defect. In another large city B, 18.5% of a random sample of 1600 school boys had the same defect. At the 5% level of significance, we want to test if the difference between the proportions is significant.

Solution:

Step 1:

Null hypothesis (H_0): $H_0 : p_1 = p_2$

Alternative hypothesis (H_1):$H_1 : p_1 \neq p_2$

Step 2: The sample proportions are:

$$\hat{p}_1 = 0.20 \quad \text{(City A)}$$

$$\hat{p}_2 = 0.185 \quad \text{(City B)}$$

Step 3: The pooled proportion (P) is calculated as:

$$P = \frac{x_1 + x_2}{n_1 + n_2} = \frac{0.20 \times 900 + 0.185 \times 1600}{900 + 1600} = 0.19$$

Step 4: The Z-test statistic is calculated using the formula:

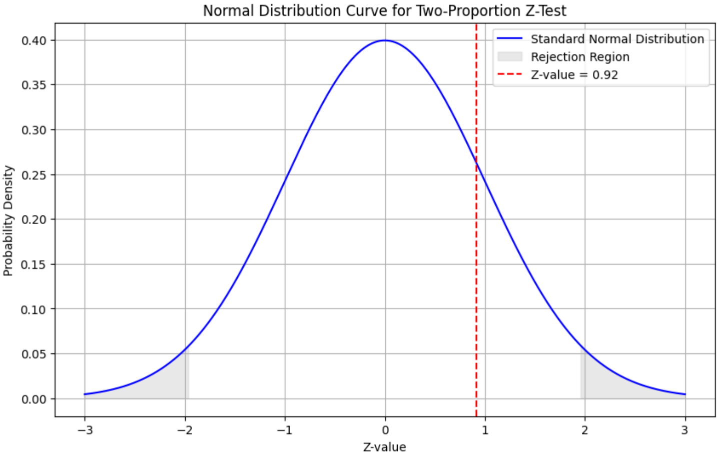
$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{P(1-P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Substituting the values:

$$z = \frac{(0.20 - 0.185)}{\sqrt{0.19(1-0.19)\left(\frac{1}{900} + \frac{1}{1600}\right)}} \approx 0.94$$

For a two-tailed test at the 5% level of significance, the critical z-value is ±1.96. Since the calculated z-value (0.94) is less than the critical value (1.96), we fail to reject the null hypothesis.

Conclusion: There is no significant difference in the proportions of boys with physical defects in cities A and B at the 5% level of significance.



The shaded areas represent the rejection regions. The red dashed line marks the calculated z-value.

Problem 3: Before an increase in excise duty on tea, 800 out of 1000 people were tea drinkers. After the increase, 800 people were tea drinkers out of 1200 people sampled. At the 5% level of significance, we want to test if the difference in tea consumption before and after the excise.

Solution:

Step 1:

Null hypothesis (H_0): $H_0 : p_1 = p_2$ Alternative hypothesis (H_1): $H_1 : p_1 \neq p_2$

Step 2: The sample proportions are calculated as follows:

$$\hat{p}_1 = \frac{800}{1000} = 0.80 \quad \text{(Before excise duty increase)}$$

$$\hat{p}_2 = \frac{800}{1200} = 0.6667 \quad \text{(After excise duty increase)}$$

Step 3: The pooled proportion (P) is the combined proportion of tea drinkers from both samples:

$$P = \frac{x_1 + x_2}{n_1 + n_2} = \frac{800 + 800}{1000 + 1200} = 0.7273$$

Step 4: The Z-test statistic is calculated using the formula:

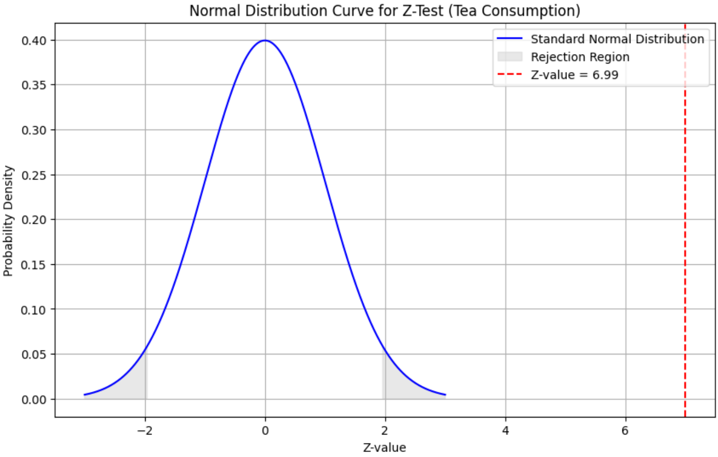
$$z = \frac{(p_1 - p_2)}{\sqrt{P(1-P)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Substituting the values:

$$z = \frac{(0.80 - 0.6667)}{\sqrt{0.7273(1 - 0.7273)\left(\frac{1}{1000} + \frac{1}{1200}\right)}} \approx 7.00$$

For a two-tailed test at the 5% level of significance, the critical z-values are ±1.96. Since the calculated z-value (7.00) is much greater than the critical value (1.96), we reject the null hypothesis.

Conclusion: There is a significant difference in the proportion of tea drinkers before and after the excise duty increase at the 5% level of significance.



The shaded areas represent the rejection regions. The red dashed line marks the calculated z-value.

Assignment Questions

(1) In a sample of 600 men from a certain city, 450 are found smokers. In another sample of 900 men from another city, 450 are smokers. Do the data indicate that the cities are significantly different with respect to the habit of smoking among men? Test at 5% significance level.
$z = 6.38$

(2) A sample of 100 tyres is taken from a lot. The mean life of a tyre is found to be 39350 kms with a SD of 3260. Can it be considered as the true random sample from a population with a mean life of 40000 kms? (Use 5% significance level).

(3) In two large populations there are 30% and 25% respectively of fair haired people. Is this difference likely to be hidden in samples of 1200 and 900 respectively from the two populations?

(4) A stenographer claims that she can type at the rate of 120 words per minute. Can we reject her claim on the basis of 100 trials in which she demonstrates a mean of 116 words with a standard deviation of 15 words? Use 5% level of significance.

Confidence Intervals for Means and Proportions

(i) Confidence Interval for Mean:
If the population standard deviation σ is known, the confidence interval for the population mean μ is given by:

$$\bar{x} \pm Z \times \frac{\sigma}{\sqrt{n}}$$

Where $\bar{x}$ is the sample mean, Z is the desired confidence level, σ is the population standard

(ii) Confidence Interval for Proportion:
The confidence interval for a population proportion p is given by:

$$\hat{p} \pm Z \times \sqrt{\frac{\hat{p}\hat{q})}{n}}$$

Where $\hat{p}$ is the sample proportion, Z is the desired confidence level, n is the sample size.

(iii) Confidence Interval for Difference of Two Means:

For two independent samples, the confidence interval for the difference in means $\mu_1 - \mu_2$ is given by:

$$(\bar{x}_1 - \bar{x}_2) \pm Z \times \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Where $\bar{x}_1$ and $\bar{x}_2$ are the sample means, σ_1 and σ_2 are the population standard deviations for the two groups, n_1 and n_2 are the sample sizes.

(iv) Confidence Interval for Difference of Two Proportions:

The confidence interval for the difference in proportions $p_1 - p_2$ is given by:

$$(\hat{p}_1 - \hat{p}_2) \pm Z \times \sqrt{\frac{\hat{p}_1 \hat{q}_1)}{n_1} + \frac{\hat{p}_2 \hat{q}_2)}{n_2}}$$

Where $\hat{p}_1$ and $\hat{p}_2$ are the sample proportions, Z is the Z-value corresponding to the desired confidence level, n_1 and n_2 are the sample sizes.

Problem 1: To know the mean weights of all 10-year-old boys in Delhi, a sample of 225 was taken. The mean weight of the sample was found to be 67 pounds, with a standard deviation of 2 pounds (considered as the population standard deviation). Find the 95% confidence interval for the mean weight of the population.

Solution: Given:

Sample size $n = 225$, Sample mean $\bar{x} = 67$

Population standard deviation $\sigma = 2$, Confidence level $= 95\%$

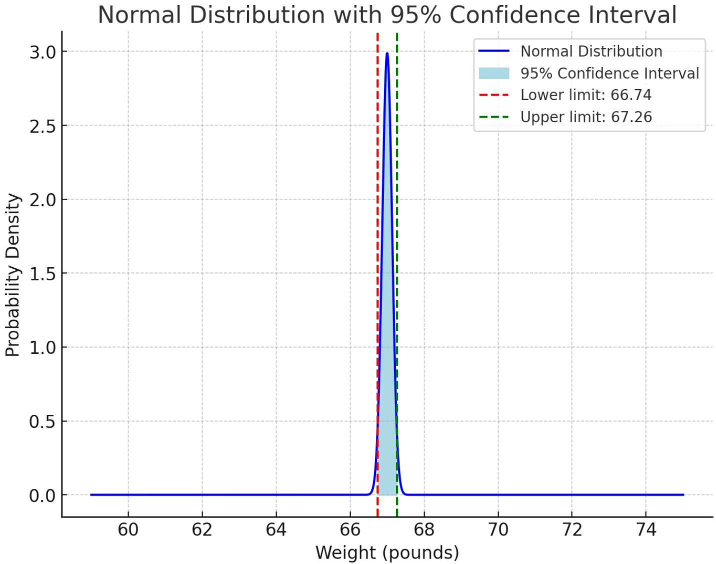
Standard error (SE) is calculated as:

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{2}{15} = 0.1333$$

The 95% confidence interval is:

$$67 \pm 1.96 \times 0.1333 = (66.74, 67.26)$$

Thus, we are 95% confident that the true mean weight lies between 66.74 pounds and 67.26 pounds.


Normal Distribution with 95% Confidence Interval
Normal Distribution
95% Confidence Interval
Lower limit: 66.74
Upper limit: 67.26
Probability Density
3.0
2.5
2.0
1.5
1.0
0.5
0.0
60
62
64
66
68
70
72
74
Weight (pounds)

Problem 2: The heights of a random sample of 50 college students showed a mean of 174.5 centimeters and a standard deviation of 6.9 centimeters. Construct a 99% confidence interval for the mean height of all college students.

Solution: Given:

Sample size $n = 50$, Sample mean $\bar{x} = 174.5\,\text{cm}$

Sample standard deviation $s = 6.9\,\text{cm}$

Confidence level $= 99\%$

Standard error (SE) is calculated as:

$$SE = \frac{s}{\sqrt{n}} = \frac{6.9}{\sqrt{50}} = 0.9756\,\text{cm}$$

The 99% confidence interval is:

$$174.5 \pm 2.576 \times 0.9756 = (171.99\,\text{cm}, 177.01\,\text{cm})$$

Thus, we are 99% confident that the true mean height of college students lies between 171.99 cm and 177.01 cm.

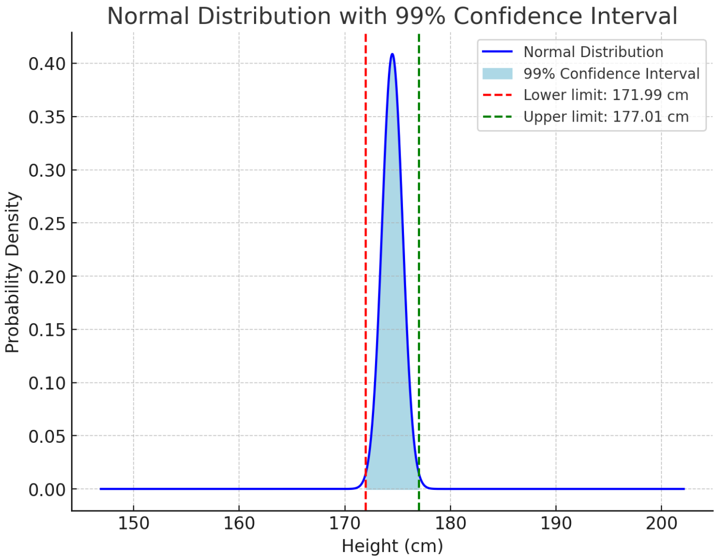


Figure: Normal Distribution Curve showing the 99% Confidence Interval

Problem 3: A random sample of 500 apples was taken from a large consignment, and 65 were found to be bad. Estimate the proportion of bad apples in the consignment as well as the standard error of the estimate. Also, find the percentage of bad apples in the consignment.

Solution: Given:

Sample size $n = 500$, Number of bad apples in the sample $x = 65$

The estimated proportion $\hat{p}$ is calculated as:

$$\hat{p} = \frac{x}{n} = \frac{65}{500} = 0.13$$

The standard error (SE) of the proportion is given by:

$$SE = \sqrt{\frac{\hat{p}\hat{q})}{n}} = \sqrt{\frac{0.13 \times 0.87}{500}} = 0.0154$$

The percentage of bad apples is:

$$\hat{p} \times 100 = 0.13 \times 100 = 13\%$$

Thus, the estimated proportion of bad apples in the consignment is 0.13, with a standard error of 0.0154, and 13% of the apples are bad.

Problem 4: In a locality of 18,000 families, a sample of 840 families was selected at random. Of these 840 families, 206 families were found to have a monthly income of Rs. 2500 or less. Estimate how many of the 18,000 families have a monthly income of Rs. 2500 or less. Also, find the limits within which this estimate would lie.

Solution: Given:
Total number of families $N = 18,000$
Sample size $n = 840$
Number of families with income Rs. 2500 or less $x = 206$
The sample proportion $\hat{p}$ is calculated as:

$$\hat{p} = \frac{206}{840} = 0.2452$$

Estimation of Families with income Rs. 2500 or less in the locality:

$$\text{Estimate} = \hat{p} \times N = 0.2452 \times 18,000 = 4,413.6 \approx 4,414$$

Standard Error (SE) is given by:

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.2452 \times (1-0.2452)}{840}} \approx 0.015$$

Confidence Interval for 95% confidence level (Z = 1.96):

$$CI = \hat{p} \pm 1.96 \times SE = 0.2452 \pm 1.96 \times 0.015$$

$$CI = (0.216, 0.274)$$

Multiplying by the total population N:

$$\text{Limits} = (0.216 \times 18,000, 0.274 \times 18,000) = (3,888, 4,932)$$

Thus, the estimate is that between 3,888 and 4,932 families in the locality have a monthly income of Rs. 2500 or less.

Problem 5: The mean and standard deviation of the maximum loads supported by 60 cables are 11.09 tonnes and 0.73 tonnes, respectively. Find: (i) 95% confidence limits for the mean of the maximum loads of all cables produced by the company. (ii) 99% confidence limits for the mean of the maximum loads of all cables produced by the company.

Solution: Given:

Sample size $n = 60$

Sample mean $\bar{x} = 11.09\,\text{tonnes}$

Standard deviation $s = 0.73\,\text{tonnes}$

Standard Error (SE):

$$SE = \frac{s}{\sqrt{n}} = \frac{0.73}{\sqrt{60}} = 0.0943\,\text{tonnes}$$

(i) 95% Confidence Limits: For 95% confidence level, the Z-value is 1.96. The confidence limits are:

$$\bar{x} \pm Z \times SE = 11.09 \pm 1.96 \times 0.0943$$

$$= 11.09 \pm 0.1848$$

$$= (10.9052\,\text{tonnes}, 11.2748\,\text{tonnes})$$

Thus, the 95% confidence limits for the mean are $(10.91, 11.27)$ tonnes.
(ii) 99% Confidence Limits: For 99% confidence level, the Z-value is 2.576. The confidence limits are:

$$\bar{x} \pm Z \times SE = 11.09 \pm 2.576 \times 0.0943$$

$$= 11.09 \pm 0.2429$$

$$= (10.8471\,\text{tonnes}, 11.3329\,\text{tonnes})$$

Thus, the 99% confidence limits for the mean are $(10.85, 11.33)$ tonnes.

Assignment Questions

(1). A survey was conducted in a slum locality of 2000 families by selecting a sample of size 800. It was revealed that 180 families were illiterates. Find the probable limits of the illiterate families in the population of 2000.
(2). A sample of 900 days was taken in a coastal town and it was found that on 100 days the weather was very hot. Obtain the probable limits of the percentage of very hot weather.
(3) The mean and S.D of the maximum loads supported by 60 cables are 11.09 tonnes and 0.73 tonnes respectively. Find (a) 95% (b) 99% confidence limits for the mean of the maximum loads of all cables produced by the company.
(4) 400 children are chosen in an industrial town and 150 are found to be underweight. Assuming the conditions of simple sampling, estimate the percentage of children who are underweight in the industrial town and assign limits within which the percentage probably lies.