Assignment 3

Normal Distribution and Empirical Rule

Domain: Monthly Salary Distribution of IT Employees

1. Introduction

In statistics and data science, understanding how data is distributed is essential for analysis, prediction, and decision-making. One of the most important probability distributions in statistics is the Normal Distribution, also known as the Gaussian Distribution.

The normal distribution is widely used in economics, engineering, social sciences, machine learning, quality control, and financial modelling. Many real-world phenomena approximately follow a normal distribution due to the Central Limit Theorem (CLT).

In this assignment, we analyse the Monthly Salary Distribution of IT Employees and model it using a normal distribution. We also explain the Empirical Rule (68–95–99.7 Rule) and its statistical significance.

2. Domain Description: Monthly Salary of IT Employees

The IT industry consists of professionals such as:

- Software Developers

- Data Scientists

- System Analysts

- Network Engineers

- DevOps Engineers

- Cybersecurity Analysts

In a large IT organization, salaries are influenced by:

- Education level

- Years of experience

- Skill specialization

- Location

- Company performance

When we collect salary data from hundreds or thousands of employees, the distribution often becomes approximately normal.

3. Definition of Normal Distribution

A Normal Distribution is a continuous probability distribution defined by its:

- Mean (μ)
- Standard Deviation (σ)

Mathematical Definition

The Probability Density Function (PDF) of a normal distribution is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Where:

- x = random variable
- μ = mean
- σ = standard deviation
- e = Euler's number (≈ 2.718)
- π = 3.14159

4. Assumed Statistical Parameters

For this study, we assume:

- Mean Salary (μ) = ₹50,000
- Standard Deviation (σ) = ₹8,000

These are hypothetical values used to demonstrate the empirical rule.

5. Key Properties of Normal Distribution

5.1 Symmetry

The normal curve is perfectly symmetric about the mean.

Mean=Median=Mode

5.2 Bell-Shaped Curve

The curve has a bell shape where:

- Peak occurs at the mean

- Tails approach the x-axis asymptotically

5.3 Total Area Under Curve = 1

The total probability under the curve equals 1:

$$\int_{-\infty}^{\infty} f(x), dx = 1$$

This represents 100% probability.

5.4 Continuous Distribution

Salary is treated as a continuous variable because it can take any real value within a range.

6. Standard Deviation and Variance

6.1 Variance (σ^2)

$$\sigma^2 = \mathbb{E}[(X - \mu)^2]$$

Variance measures the spread of salary values from the mean.

6.2 Standard Deviation (σ)

$$\sigma = \sqrt{\sigma^2}$$

Standard deviation tells us how much salaries typically deviate from ₹50,000.

Higher $\sigma \rightarrow$ more spread Lower $\sigma \rightarrow$ more concentrated salaries

7. Empirical Rule (68–95–99.7 Rule)

The Empirical Rule applies only to normal distributions.

7.1 Within ±1σ (68%)

$$\mu - \sigma \text{ to } \mu + \sigma$$

₹42,000 to ₹58,000

Approximately 68% of employees earn within this range.

7.2 Within ±2σ (95%)

₹34,000 to ₹66,000

Approximately 95% of employees earn within this range.

7.3 Within ±3σ (99.7%)

₹26,000 to ₹74,000

Approximately 99.7% of employees earn within this range.

Very few employees fall outside this range.

8. Graphical Representation of Salary Distribution

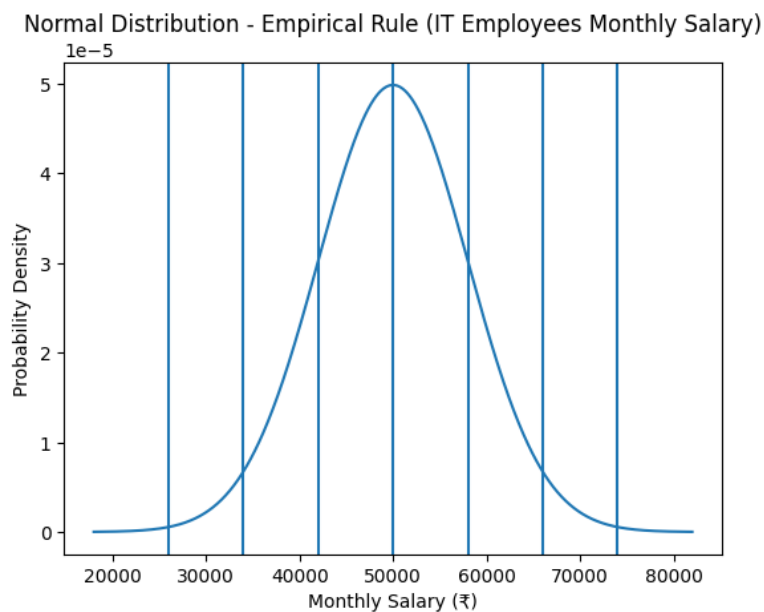**Figure 1: Normal Distribution of IT Employees' Monthly Salary**

Figure 1 shows the bell-shaped normal distribution curve of IT employees' monthly salary with mean (μ = ₹50,000) and standard deviation (σ = ₹8,000). The vertical lines indicate ±1σ, ±2σ, and ±3σ ranges representing 68%, 95%, and 99.7% of the data respectively.

9. Z-Score and Standardization

To standardize the distribution, we use:

$$Z = \frac{X - \mu}{\sigma}$$

Where:

- Z = Standard score

- X = observed value

Example:

If salary = ₹66,000:

$$Z = \frac{X - \mu}{\sigma} = \frac{80000 - 60000}{10000} = 2$$

This means the salary is 2 standard deviations above the mean.

10. Central Limit Theorem (CLT)

The Central Limit Theorem states:

The sampling distribution of the sample mean approaches a normal distribution as sample size increases, regardless of population distribution.

This explains why salary data often becomes normally distributed when the sample size is large.

11. Skewness and Kurtosis

11.1 Skewness

Normal distribution has:

Skewness=0

Meaning it is perfectly symmetric.

11.2 Kurtosis

Normal distribution has:

Kurtosis=3

It is called Mesokurtic Distribution.

12. Practical Applications in IT Industry

12.1 HR Salary Analysis

- Determine salary benchmarks

- Detect salary outliers

- Maintain pay equity

12.2 Financial Forecasting

- Predict salary growth

- Budget allocation

12.3 Machine Learning

Normal distribution assumptions are used in:

- Gaussian Naive Bayes

- Linear Regression

- Statistical Hypothesis Testing

13. Hypothesis Testing Example

Suppose management claims:

- "Average salary is ₹52,000."

We test using:

- Null Hypothesis (H_0): $\mu = 50,000$

- Alternative Hypothesis (H_1): $\mu \neq 50,000$

Using Z-test, we check statistical significance.

14. Real-World Limitations

Although salary data is often approximately normal:

- Executive salaries may cause right skew

- Entry-level bulk hiring may cause left skew

- Economic recession may distort distribution

Thus, real data may not be perfectly normal.

15. Graph Interpretation

The generated graph shows:

- Peak at ₹50,000

- Symmetry on both sides

- Gradual decline toward extremes

- Vertical lines marking $\pm1\sigma, \pm2\sigma, \pm3\sigma$

The area under the curve represents probability density.

16. Advantages of Using Normal Distribution

- Simplifies complex data

- Enables probability calculations

- Useful in predictive modelling

- Forms basis for inferential statistics

17. Conclusion

The monthly salary distribution of IT employees can be effectively modelled using a normal distribution. The empirical rule helps us understand how salary values are distributed around the mean.

Key Observations:

- Most employees earn near ₹50,000

- Extreme salaries are rare

- Standard deviation controls spread

- Z-score enables comparison

The normal distribution plays a fundamental role in statistics, economics, data science, and engineering.