Training Day-111 Report:

Normal Distributions

What is a Normal Distribution?

A **Normal Distribution**, also known as a **Gaussian Distribution**, is a continuous probability distribution characterized by its symmetric, bell-shaped curve. It is one of the most important distributions in statistics and is widely used in various fields due to its natural occurrence in many real-world phenomena.

The Normal Distribution is defined by two parameters:

1. **Mean (μ\mu):** Determines the center of the distribution.
2. **Standard Deviation (σ\sigma):** Determines the spread or width of the distribution.

Properties of Normal Distribution

1. **Symmetry:**
   - The curve is symmetric around the mean (μ\mu).
2. **Bell Shape:**
   - The curve has a peak at the mean and tails off equally on both sides.
3. **Empirical Rule (68-95-99.7 Rule):**
   - About **68%** of the data lies within one standard deviation of the mean (μ±σ\mu \pm \sigma).
   - About **95%** of the data lies within two standard deviations (μ±2σ\mu \pm 2\sigma).
   - About **99.7%** of the data lies within three standard deviations (μ±3σ\mu \pm 3\sigma).
4. **Total Area Under the Curve:**
   - The total area under the curve is equal to 1.

Probability Density Function (PDF)

The PDF of a normal distribution is given by:

f(x)=12πσ2e−(x−μ)22σ2f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}

Where:

- xx is the random variable.
- μ\mu is the mean.

- σ^2 is the variance.
- e is the base of the natural logarithm ($e \approx 2.718$).

Standard Normal Distribution

A **Standard Normal Distribution** is a special case of the normal distribution with:

- Mean (μ) = 0
- Standard Deviation (σ) = 1

To standardize a normal random variable X, use the **Z-score**:

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

Applications of Normal Distribution

1. **Statistics:**
   - Used in hypothesis testing, confidence intervals, and regression analysis.
2. **Natural Sciences:**
   - Models phenomena like heights, weights, and measurement errors.
3. **Finance:**
   - Analyzes stock market returns and risk assessments.
4. **Machine Learning:**
   - Assumes normality in algorithms like Gaussian Naive Bayes.
5. **Quality Control:**
   - Evaluates processes under the assumption of normality in production lines.

Examples

1. Height of Individuals:

Suppose the heights of a population are normally distributed with a mean of 170 cm and a standard deviation of 10 cm. The probability of a randomly selected individual being between 160 and 180 cm is:

$$P(160 \leq X \leq 180) = \int_{160}^{180} f(x) dx$$

Use the Z-score and standard normal tables for calculation.

2. Exam Scores:

If exam scores are normally distributed with a mean of 75 and a standard deviation of 8, the probability of scoring above 85 is:

$$Z = \frac{85 - 75}{8} = 1.25$$

Look up $Z = 1.25$ in standard normal tables to find the probability.

Training Day-112 Report:

Z-Score:

What is Z-Score?

The **Z-Score** (also called the **Standard Score**) measures how many standard deviations a data point is from the mean of a distribution. It allows comparison of data points from different distributions by standardizing them.

A Z-score helps to determine whether a value is typical or unusual in a given dataset.

Formula for Z-Score

Z=X−μσZ = \frac{X - \mu}{\sigma}

Where:

- ZZ = Z-score
- XX = Individual data point
- μ\mu = Mean of the distribution
- σ\sigma = Standard deviation of the distribution

Interpretation of Z-Score

1. **Positive Z-Score (Z>0Z > 0):**
    - Indicates that the data point is above the mean.
2. **Negative Z-Score (Z<0Z < 0):**
    - Indicates that the data point is below the mean.
3. **Z-Score of 0:**
    - Indicates that the data point is exactly at the mean.
4. **Magnitude of Z-Score:**
    - The further the Z-score is from 0, the more unusual the data point is.

Example of Z-Score Calculation

Example 1: Exam Scores

Suppose exam scores are normally distributed with a mean (μ\mu) of 75 and a standard deviation (σ\sigma) of 10. A student scores 85 on the exam.

To calculate the Z-score:

Z=85−7510=1010=1Z = \frac{85 - 75}{10} = \frac{10}{10} = 1

The Z-score is **1**, meaning the student scored 1 standard deviation above the mean.

Applications of Z-Score

1. **Comparing Data Points:**
   - Standardizes values from different distributions for comparison.
   - Example: Comparing test scores from two tests with different scales.
2. **Outlier Detection:**
   - Data points with Z-scores greater than 3 or less than -3 are often considered outliers.
3. **Probability Calculations:**
   - Z-scores are used with standard normal tables to calculate probabilities for normal distributions.
4. **Standardizing Data in Machine Learning:**
   - Z-scores are used to normalize features to have a mean of 0 and standard deviation of 1.

Z-Score and the Standard Normal Table

The **Standard Normal Table** (Z-Table) provides the cumulative probability for a given Z-score.

- Example: A Z-score of **1.25** corresponds to a cumulative probability of approximately **0.8944**.
  This means that 89.44% of the data lies below this Z-score.

Examples with Probability

Example 1: Finding Probabilities

A dataset of weights is normally distributed with a mean of 70 kg and a standard deviation of 5 kg. What is the probability that a randomly chosen person weighs less than 75 kg?

1. Calculate the Z-score:

$$Z = \frac{75 - 70}{5} = 1$$

2. Use the Z-Table:
   - A Z-score of **1** corresponds to a cumulative probability of **0.8413**.
   - Therefore, there is an **84.13%** chance that a person weighs less than 75 kg.

Example 2: Finding Intervals

What is the probability of a person weighing between 65 kg and 75 kg?

1. Calculate the Z-scores:
   - For 65 kg: $Z = \frac{65 - 70}{5} = -1$.
   - For 75 kg: $Z = \frac{75 - 70}{5} = 1$.

2. Use the Z-Table:
    - Cumulative probability for $Z = -1$: **0.1587**.
    - Cumulative probability for $Z = 1$: **0.8413**.
3. Subtract probabilities:

$$P(65 \leq X \leq 75) = 0.8413 - 0.1587 = 0.6826$$

- Therefore, there is a **68.26%** chance that a person weighs between 65 kg and 75 kg.

Training Day-113 Report:

Central Limit Theorem

What is the Central Limit Theorem?

The **Central Limit Theorem (CLT)** states that the sampling distribution of the sample mean approaches a normal distribution as the sample size increases, regardless of the original distribution of the population.

This theorem is a cornerstone of statistics because it justifies the use of normal distribution in hypothesis testing and confidence interval estimation, even when the underlying population distribution is unknown.

Key Points of the Central Limit Theorem

1. **Normality of Sampling Distribution:**
   - For large sample sizes ($n \geq 30$n \geq 30), the distribution of sample means will be approximately normal.
2. **Mean and Standard Deviation:**
   - The mean of the sampling distribution ($\mu_{\bar{X}}$\mu_{\bar{X}}) equals the population mean (μ\mu).
   - The standard deviation of the sampling distribution ($\sigma_{\bar{X}}$\sigma_{\bar{X}}), also called the **Standard Error (SE)**, is given by: $\sigma_{\bar{X}}=\sigma n$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}
3. **Independence of Sample Size:**
   - The CLT holds true regardless of the population's original shape, as long as the samples are independent and identically distributed.

Applications of Central Limit Theorem

1. **Hypothesis Testing:**
   - Enables the use of Z-tests and T-tests, which assume normality of the sampling distribution.
2. **Confidence Intervals:**
   - Helps calculate intervals to estimate population parameters.
3. **Quality Control:**
   - Used in evaluating sample statistics to infer about population metrics.

Example of CLT

Suppose a population has a mean (μ) of 100 and a standard deviation (σ) of 15. If we take random samples of size $n = 50$:

- The mean of the sampling distribution remains 100.
- The standard error is: $\sigma_{\bar{X}} = \frac{15}{\sqrt{50}} = 2.12$

If the sample means are plotted, their distribution will approximate a normal curve, even if the population distribution is not normal.

Hypothesis Testing

What is Hypothesis Testing?

Hypothesis Testing is a statistical method used to evaluate assumptions (hypotheses) about a population parameter based on sample data.

Key Steps in Hypothesis Testing

1. **State the Hypotheses:**
    - Null Hypothesis (H_0): Assumes no effect or no difference.
    - Alternative Hypothesis (H_a): Contradicts H_0; claims an effect or difference exists.
2. **Set the Significance Level (α):**
    - Commonly used values are 0.05 (5%) or 0.01 (1%).
3. **Choose the Appropriate Test:**
    - Use Z-tests, T-tests, Chi-square tests, or others depending on data type and sample size.
4. **Calculate the Test Statistic:**
    - Example: For Z-test, compute: $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$
5. **Determine the p-value:**
    - Compare the p-value to α to decide whether to reject H_0.
6. **Make a Decision:**
    - If $p \leq \alpha$, reject H_0.
    - If $p > \alpha$, fail to reject H_0.

Types of Errors

1. **Type I Error (α):**
    - Rejecting H_0 when it is true.

2. **Type II Error (β):**
    - Failing to reject H_0 when it is false.

Example of Hypothesis Testing

Example 1: Mean Testing

Suppose the average height of students is claimed to be 170 cm ($H_0: \mu = 170$). A sample of 30 students has a mean height of 172 cm, with a standard deviation of 5 cm. Use a significance level of 0.05.

1. **State Hypotheses:**
    - $H_0: \mu = 170$, $H_a: \mu \neq 170$.
2. **Calculate Test Statistic:**

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{172 - 170}{\frac{5}{\sqrt{30}}} = 2.19$$

3. **Find p-value:**
    - From Z-tables, $p = 0.0282$.
4. **Decision:**
    - Since $p = 0.0282 < \alpha = 0.05$, reject H_0.
    - Conclusion: The mean height is significantly different from 170 cm.

Applications of Hypothesis Testing

1. **Medicine:**
    - Test the efficacy of new drugs.
2. **Manufacturing:**
    - Evaluate quality differences in production.
3. **Market Research:**
    - Determine customer preferences or behavior changes.