Statistics-Assignment-1

1. Explain the difference between descriptive and inferential statistics. Provide examples of each.

Descriptive statistics:

Descriptive statistics describes the data, it is concerned with the data summarization using graphs, charts and tables. Description refers to what already has happened.

Examples:

1. **Measures of Central Tendency**

➢ Mean : Mean is defined as the sum of all observations in the dataset divided by the number of observations. For example,the average height of students in a class.

$$X = \Sigma X/n$$

➢ Median: Median is the middle most observation when the data is arranged in the ascending order.For example, the median marks scored by a class.

Median=(n+1)th observation/2

➢ Mode: Mode is defined as the most frequently occurring value in the distribution. For example, the most popular car in the city.

2. **Measures of dispersions**

➢ Range: The difference between the largest and the smallest values in a dataset.

Range= X max - X min

➢ Variance: Variance is a measure of the spread of the data. It shows how far the data point is from the mean. If the variance is high, the data is widely spread.

$$\text{Variance for population } (\sigma) = 1/N \sum_{i=1}^{n} (xi - \mu)^2$$

$$\text{Variance for sample (s)} = 1/(n-1) \sum_{1=1}^{n} (xi - \bar{x})^2$$

➢ Standard deviation: It is the square root of the variance.

$$\sigma^2 = 1/N \sum_{i=1}^{n} (xi - \mu)^2$$

$$s^2 = 1/(n-1) \sum_{1=1}^{n} (xi - \bar{x})^2$$

3. Graphical representation

➢ Histograms: shows the distribution of the continuous variable.
➢ Bar charts: shows the frequency of the categorical data.
➢ Pie chart: shows the proportion of different categories within a whole.

Inferential Statistics

Inferential statistics is used to make inferences and predictions about a population based on a sample of the data. Inference refers to what might happen and what might have happened

Examples:

➢ Confidence Intervals
➢ Hypothesis testing
➢ Correlation Analysis

➢ Regression analysis

2. Define the Central Limit Theorem and discuss its significance in statistical inference

Central Limit Theorem: The Central Limit Theorem (CLT) states when sampling is done from a population with mean μ and standard deviation σ , the sampling distribution of the sample mean $\overline{x}$ will tend to a normal distribution with mean μ and standard deviation $\sigma/\sqrt{n}$ as the sample size n becomes large.

❖ CLT theorem states:

- No matter the distribution
- The distribution of $x1$, $x2$, $x3$, $x4$, … , xk would tend to N~ (μ, σ^2/n)
- The more samples, the closer to Normal (k -> ∞)
- The bigger the samples, the closer to Normal (n -> ∞)

❖ CLT allows us to perform tests, solve problems and make inferences using the normal distribution,even when the population is not normally distributed.

❖ The CLT allows us to assume normality for many different variables, which is very useful for confidence intervals, hypothesis testing, and regression analysis.

3. Discuss the concept of sampling and its role in statistical analysis.

Sampling: A sample is randomly chosen from the entire population. A sampling technique where every item in the population has an equal chance of being selected, so the sample is likely to be representative of the population.

- In the real world, it is not always possible to get an exact idea about population parameters, that is why it is important to draw samples from the population.
- Inferential statistics approximates the population parameters by taking samples from it.
- A good sample represents the population well.
- Sample mean varies from one sample to another.

Sampling distribution: The sampling distribution of a statistic is the probability distribution of all possible values the statistic may take when computed from random samples of the same size, drawn from a specified population.

Suppose we are sampling from a population with mean μ and standard deviation σ . Let $\bar{x}$ be the random variable representing the sample mean of n independent observations. Then

- Mean of $\bar{x}$ is equal to μ
- Standard deviation of $\bar{x}$ is equal to $\sigma/\sqrt{n}$ (also called the standard error of $\bar{x}$.
- Even if the population is not normally distributed, for sufficiently large n, $\bar{x}$ is also normally distributed.

4. Explain the process of hypothesis testing and the key components involved.

Hypothesis testing : Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data. Hypothesis testing is basically an assumption that we make about the population parameter.

Process of the hypothesis testing

Stating the hypothesis: Null hypothesis and Alternate hypothesis are two mutually exclusive statements about the population parameter.

Null Hypothesis(H_o) : The presumed current state of the matter or status quo.

$$H_o: \mu = \mu_o \quad \text{or} \quad H_o: \mu \leq \mu_o \quad \text{or} \quad H_o: \mu \geq \mu_o$$

Alternate Hypothesis (H_a) : An argument or a research question that you want to prove with solid ground obtained from a sample.

$$H_a: \mu \neq \mu_o \quad \text{or} \quad H_a: \mu > \mu_o \quad \text{or} \quad H_a: \mu < \mu_o$$

Identify/Calculate: Depending on the data decide the test

Z-Score - If $population\ standard\ deviation(\sigma)$ is present and n>30

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

T-Score - If the $population\ standard\ deviation(\sigma)$ is absent.

$$T = (x-\mu)/(s/\sqrt{n})$$

Decide the α - **value** : Usually it is 0.05 (as per industry standard).

Calculate p-value:

Calculate the p-value using either Z or T score to check if null hypothesis is true.

Conclusion:

Compare p-value with α - value , if p - value < α- value then reject the H_o and accept the H_a .

Key components:

- ➢ **Level of significance:** Probability of rejecting the null hypothesis when it is true. Fixed before the Hypothesis test.

- ➢ **p-value:** Probability of observing test statistic or more extreme results than the computed test statistic under the null hypothesis.It depends on the sample data. Alpha is pre-fixed but the p-value depends on the value of the test statistic.

- ➢ **Acceptance or Rejection region:** The total area under the distribution curve of the test statistics is divided into Acceptance and Rejection region. Reject the null hypothesis when the test statistic lies in the rejection region, else we fail to reject the null hypothesis.

5. Describe the T-distribution and how it differs from the normal distribution.

T-distribution:

- To use the central limit theorem, we need to know the population standard deviation,σ . When σ is not known, we use its estimator, the sample standard deviation s, in its place. If the population is normally distributed, the standardized statistic

  $$t = (x-\mu)/(s/\sqrt{n}),$$

  has a t-distribution with (n-1) degrees of freedom. The degrees of freedom of the distribution are the degrees of freedom associated with sample standard deviation.

- The t-distribution is also called the student's distribution or student's t-distribution.

- The t-distribution has wider tails than standard normal distribution.

Applied Questions:

6. Calculate the mean, median, and standard deviation for the following dataset: [10, 15, 20, 25, 30].

Mean:

$X = \Sigma X/n$ = 20

Median:

Median=(n+1)th observation/2 = (5+1)/2= 3rd observation= 20

Standard deviation:

$\sigma^2 = 1/N \sum_{i=1}^{n} (xi - \mu)^2$ = 1/5 $\sum_{i=1}^{5}$ (xi- 20) = 50

7. A researcher wants to estimate the average height of students in a university. She samples 50 students and finds the mean height to be 65 inches with a standard deviation of 3 inches. Construct a 95% confidence interval for the population mean height.

To calculate 95% confidence interval for the population mean height, sample mean and sample standard deviation is given and population standard deviation is not given, so we will use t-distribution

Given:

- Sample size (n) = 50
- Sample mean ($\bar{x}$) = 65 inches
- Sample standard deviation (s) = 3 inches
- Confidence level = 95%

Degrees of freedom (df) = n - 1 = 50 - 1 = 49

Finding the t-score for a 95% confidence level with 49 degrees of freedom:

- Using a t-table, we find that the t-score is approximately 2.0096.

Formula for a Confidence Interval using t-distribution:

- $CI = \bar{x} \pm t^* (s / \sqrt{n})$

Calculating the Confidence Interval:

- $CI = 65 \pm 2.0096 (3 / \sqrt{50})$
- $CI \approx 65 \pm 2.0096 (0.4243)$
- $CI \approx 65 \pm 0.8515$

Therefore, the 95% confidence interval for the population mean height is approximately (64.1485, 65.8515) inches.

8. A manufacturer claims that the average lifespan of its light bulbs is 1000 hours. A random sample of 50 light bulbs has a

mean lifespan of 980 hours with a standard deviation of 50 hours. Test the manufacturer's claim at a significance level of 0.05 using a right-tailed hypothesis test.

Given

- Manufacturer's claim: Average lifespan (μ) = 1000 hours
- Sample size (n) = 50
- Sample mean ($\bar{x}$) = 980 hours
- Sample standard deviation (s) = 50 hours
- Significance level (α) = 0.05
- Right-tailed test

Hypothesis:

- Null hypothesis (H_0): μ = 1000 hours
- Alternative hypothesis (H_1): μ > 1000 hours (claim)

Test Statistic: Since the population standard deviation is unknown, we use a t-test.

- $t = (\bar{x} - \mu) / (s / \sqrt{n})$
- $t = (980 - 1000) / (50 / \sqrt{50})$
- $t \approx -2.828$

Degrees of Freedom (df):

- df = n - 1 = 50 - 1 = 49

P-value:

- For a Right-tailed test with the calculated t-value of -2.828, we need to find the p-value from the t-distribution with 49 degrees of freedom.

- For t = -2.828 and degrees of freedom 49, p-value is approximately 0.0033

Compare the p-value to the Significance Level

The significance level (α) is 0.05

- **If the p-value ≤ α**: Reject the null hypothesis.
- **If the p-value > α**: Do not reject the null hypothesis.

In this case:

- The p-value (0.0033) is less than the significance level (0.05)

Decision: Since the p-value (0.0033) is less than the significance level (0.05), we reject the null hypothesis.

Conclusion: There is sufficient evidence to reject the manufacturer's claim that the average lifespan of their light bulbs is 1000 hours. The sample data suggests that the average lifespan is less than 1000 hours

9. A pharmaceutical company is testing a new drug for lowering blood pressure. They want to determine if the drug is effective in reducing blood pressure levels. State the null and alternative hypotheses for this study

Hypothesis Testing:

- Null Hypothesis: New drug is not effective in reducing blood pressure.

$$H_0 : \mu = \mu_o$$

- Alternative hypothesis (H_1): New drug is effective in reducing blood pressure.

$$H_1 : \mu < \mu_o$$

10. A quality control manager at a factory wants to ensure that the average weight of products coming off the production line is 500 grams. She takes a random sample of 30 products and finds

the mean weight to be 495 grams with a standard deviation of 10 grams. Test the manager's claim at a significance level of 0.01 using a left-tailed hypothesis test.

Given

- Manager’s claim : Average weight of products (μ) = 500 grams
- Sample size (n) = 30
- Sample mean ($\bar{x}$) = 495 grams
- Sample standard deviation (s) = 10 grams
- Significance level (α) = 0.01
- Left-tailed hypothesis test

Hypothesis:

- Null hypothesis (H_0): $\mu = 500$ grams
- Alternative hypothesis (H_1): $\mu < 500$ grams

Test Statistic: Since the population standard deviation is unknown, we use a t-test.

- $t = (\bar{x} - \mu) / (s / \sqrt{n})$
- $t = (495 - 500) / (10 / \sqrt{30})$
- $t = -5/1.825 = -2.739$

Degrees of Freedom (df):

- $df = n - 1 = 30 - 1 = 29$

P-value:

- For a left-tailed test with the calculated t-value of -2.739, we need to find the p-value from the t-distribution with 29 degrees of freedom.

- For t = -2.739 and degrees of freedom 29, p-value is approximately 0.005

Compare the p-value to the Significance Level

The significance level (α) is 0.01.

- **If the p-value ≤ α**: Reject the null hypothesis.
- **If the p-value > α**: Do not reject the null hypothesis.

In this case:

- The p-value (0.005) is less than the significance level (0.01).

Decision: Since the p-value (0.005) is less than the significance level (0.01), we reject the null hypothesis.

Conclusion: There is sufficient evidence to reject the manager's claim that the average weight of the product is less than 500 grams. The sample data suggests that the average weight of the product is less than 500 grams.