Statistics and Probability Theory Assignment:

I. Foundational Knowledge

1. Familiarize Yourself with Basic Statistical Concepts

- **Mean** : The average of a dataset, calculated as the sum of all values divided by the number of values.
- **Median** : The middle value in a dataset when arranged in ascending order. If there’s an even number of values, the median is the average of the two middle values.
- **Mode** : The most frequently occurring value in a dataset.
- **Standard Deviation** : A measure of how spread out the data is from the mean. It quantifies variability or dispersion.

2. Understand Descriptive vs. Inferential Statistics

- **Descriptive Statistics** : Summarizes and describes data using measures like mean, median, mode, and standard deviation. Example: Calculating the average height of students in a class.
- **Inferential Statistics** : Uses sample data to make generalizations about a population. Example: Estimating the average height of all students in a university based on a sample.

3. Importance of Probability Theory

Probability theory helps us quantify uncertainty and randomness in data. It forms the foundation for statistical inference, hypothesis testing, and decision-making under uncertainty.

II. Theoretical Questions

1. Difference Between Descriptive and Inferential Statistics

- **Descriptive Statistics** :
  - Purpose: To summarize and describe data.
  - Tools: Mean, median, mode, standard deviation, histograms, etc.
  - Example: A teacher calculates the average test score of a class (mean = 75).
- **Inferential Statistics** :
  - Purpose: To draw conclusions about a population based on sample data.
  - Tools: Confidence intervals, hypothesis testing, regression analysis, etc.
  - Example: A researcher estimates the average test score of all students in a school based on a sample of 30 students.

2. Central Limit Theorem

- **Definition** : 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 shape of the population distribution.
- **Significance** :
  - Allows us to use the normal distribution for hypothesis testing and confidence intervals, even if the population is not normally distributed.
  - Justifies the use of z-tests and t-tests for large samples.

3. Sampling and Its Role in Statistical Analysis

- **Sampling** : The process of selecting a subset of individuals or items from a population to represent the whole.

- **Role** :
  - Reduces cost and time compared to studying the entire population.
  - Enables estimation of population parameters (e.g., mean, proportion) using sample statistics.
  - Helps minimize bias if random sampling methods are used.

4. Hypothesis Testing Process

- **Key Components** :
  1. **Null Hypothesis (H_0)** : The default assumption (e.g., no effect, no difference).
  2. **Alternative Hypothesis (H_1)** : The claim being tested (e.g., there is an effect or difference).
  3. **Test Statistic** : A value calculated from the sample data (e.g., z-score, t-score).
  4. **Significance Level (α)** : The threshold for rejecting H_0 (commonly 0.05).
  5. **P-value** : The probability of observing the test statistic or something more extreme, assuming H_0 is true.
  6. **Decision Rule** : Reject H_0 if p-value $\leq \alpha$; otherwise, fail to reject H_0.

5. T-Distribution vs. Normal Distribution

- **T-Distribution** :
  - Used when the sample size is small ($n < 30$) or the population standard deviation is unknown.
  - Has heavier tails than the normal distribution, accounting for more variability in small samples.

- **Normal Distribution** :
  - Used when the sample size is large ($n \geq 30$) or the population standard deviation is known.
  - Symmetrical and bell-shaped.

III. Applied Questions

6. Calculate Mean, Median, and Standard Deviation

Dataset: [10, 15, 20, 25, 30]

- **Mean** :

Mean=Number of valuesSum of all values=510+15+20+25+30=5100 =20

- **Median** : Arrange the data in ascending order: [10, 15, 20, 25, 30]. The middle value is 20.
- **Standard Deviation** : Formula:

$\sigma=n\sum(xi-\bar{x})2$

Where xi are the data points, $\bar{x}$ is the mean, and n is the number of values.

  - Step 1: Calculate deviations from the mean: [-10, -5, 0, 5, 10].
  - Step 2: Square the deviations: [100, 25, 0, 25, 100].
  - Step 3: Sum the squared deviations: 100+25+0+25+100=250.
  - Step 4: Divide by n: 5250=50.
  - Step 5: Take the square root: 50≈7.07.

Results :

  - Mean = 20

- Median = 20
- Standard Deviation ≈ 7.07

7. Construct a 95% Confidence Interval

Given:

- Sample mean ($\bar{x}$) = 65 inches
- Sample standard deviation (s) = 3 inches
- Sample size (n) = 50
- Confidence level = 95%
- **Step 1** : Find the critical value ($t*$): For a 95% confidence level and $df=n-1=49$, $t*\approx 2.01$ (from t-table).
- **Step 2** : Calculate the margin of error (ME):

$ME=t*\cdot ns=2.01\cdot 503\approx 2.01\cdot 0.424\approx 0.85$

- **Step 3** : Construct the interval:

$\text{CI}=\bar{x}\pm ME=65\pm 0.85=(64.15,65.85)$

Result : The 95% confidence interval is approximately **(64.15, 65.85)** .

8. Test the Manufacturer's Claim

Given:

- Claimed mean ($\mu 0$) = 1000 hours
- Sample mean ($\bar{x}$) = 980 hours
- Sample standard deviation (s) = 50 hours
- Sample size (n) = 50

- Significance level (α) = 0.05
- Right-tailed test
- **Step 1** : State hypotheses:
  - $H0:\mu=1000$
  - $H1:\mu>1000$
- **Step 2** : Calculate the test statistic (t):

$t=s/nx^{-}-\mu0=50/50980-1000=7.07-20\approx-2.83$

- **Step 3** : Find the critical value: For $df=49$ and $\alpha=0.05$, the critical value is $t$critical≈1.68.
- **Step 4** : Compare t to $t$critical: Since $t=-2.83$ is less than $t$critical $=1.68$, we fail to reject $H0$.

Conclusion : There is insufficient evidence to support the manufacturer's claim at the 0.05 significance level.

9. Null and Alternative Hypotheses

For the pharmaceutical company:

- $H0:\mu=\mu0$ (The drug has no effect on blood pressure).
- $H1:\mu<\mu0$ (The drug reduces blood pressure).

10. Left-Tailed Hypothesis Test

Given:

- Claimed mean ($\mu0$) = 500 grams
- Sample mean (x^{-}) = 495 grams
- Sample standard deviation (s) = 10 grams

- Sample size (n) = 30
- Significance level (α) = 0.01
- **Step 1** : State hypotheses:
  - $H0{:}\mu=500$
  - $H1{:}\mu<500$
- **Step 2** : Calculate the test statistic (t):

$t=s/n\bar{x}-\mu 0=10/30495-500=1.826-5\approx-2.74$

- **Step 3** : Find the critical value: For $df=29$ and $\alpha=0.01$, the critical value is $t$critical≈-2.46.
- **Step 4** : Compare t to $t$critical: Since $t=-2.74$ is less than $t$critical $=-2.46$, we reject $H0$.

Conclusion : There is sufficient evidence to conclude that the average weight is less than 500 grams at the 0.01 significance level.