Question1: Define the z-statistic and explain its relationship to the standard normal distribution. How is the z-statistic used in hypothesis testing?

Answer - **Z-Statistic and Its Uses in Hypothesis Testing**
The z-statistic measures how far a data point is from the mean in terms of standard deviations, using the formula:

$$z = \frac{x - \mu}{\sigma}$$

Here, xxx is the data point, μ is the mean, and σ is the standard deviation.

Relationship to Standard Normal Distribution:
The z-statistic follows a standard normal distribution (mean = 0, standard deviation = 1). It tells us how likely it is for a data point to occur.

Relationship to Standard Normal Distribution:
The z-statistic follows a standard normal distribution (mean = 0, standard deviation = 1). It tells us how likely it is for a data point to occur.

Use in Hypothesis Testing:

- **Compare the z-value to critical values from the z-table.**
- **If the z-value falls in the rejection region, reject the null hypothesis.**
- **Commonly used in z-tests for large samples or known population variance.**

Question2 : What is a p-value, and how is it used in hypothesis testing? What does it mean if the p-value is very small (e.g., 0.01)?

Answer - P-Value and Its Role in Hypothesis Testing
The p-value is the probability of obtaining a result as extreme as, or more extreme than, the observed result, assuming the null hypothesis is true.

Use in Hypothesis Testing:

- **Compare the p-value to the significance level (α\alphaα, usually 0.05).**
- **If p-value ≤ α\alphaα, reject the null hypothesis (result is statistically significant).**
- **If p-value > α\alphaα, fail to reject the null hypothesis (insufficient evidence).**

Interpretation of a Small P-Value (e.g., 0.01):

- **A very small p-value means the observed result is highly unlikely under the null hypothesis.**
- **For p=0.01p = 0.01p=0.01, there is only a 1% chance that the observed data occurred due to random variation, leading to rejection of the null hypothesis.**

Question3: Compare and contrast the binomial and Bernoulli distributions.

Answer - Binomial vs. Bernoulli Distributions

1. Definition:

- **Bernoulli Distribution: Represents a single trial with two outcomes (success or failure).**
- **Binomial Distribution: Represents the number of successes in nnn independent Bernoulli trials.**

2. Parameters:

- **Bernoulli:**
  - **p: Probability of success.**
- **Binomial:**
  - **n: Number of trials.**
  - **p: Probability of success in each trial.**

- Bernoulli:
  $P(X = x) = p^x(1-p)^{1-x}$, where $x = 0$ or 1.
- Binomial:
  $P(X = k) = \binom{n}{k}p^k(1-p)^{n-k}$, where $k = 0, 1, ..., n$.

4. Relationship:

- The Bernoulli distribution is a special case of the binomial distribution where n=1

Question 4: Under what conditions is the binomial distribution used, and how does it relate to the Bernoulli distribution?

Answer - **Conditions for Using the Binomial Distribution:**

The **binomial distribution** is used under the following conditions:

1. **Fixed Number of Trials (n)**: The experiment has a fixed number of independent trials.
2. **Two Possible Outcomes**: Each trial results in either success or failure.
3. **Constant Probability (p)**: The probability of success remains the same for each trial.
4. **Independent Trials**: The outcome of one trial does not affect the others.

Relationship to the Bernoulli Distribution:

- A **Bernoulli distribution** describes a single trial with two outcomes (success or failure).

- The **binomial distribution** generalizes this concept by considering multiple (nnn) independent Bernoulli trials and counting the number of successes.
- In essence:
    - **Bernoulli**: For n=1 , the binomial distribution becomes a Bernoulli distribution.

Example:

- **Bernoulli**: Tossing a coin once (n=1), success = heads.
- **Binomial**: Tossing a coin 10 times (n=10), counting the number of heads.

Question5: What are the key properties of the Poisson distribution, and when is it appropriate to use this distribution?

Answer - **Key Properties of the Poisson Distribution:**

1. **Discrete Distribution**: It models the number of events occurring in a fixed interval (time, space, etc.).
2. **Events are Independent**: The occurrence of one event does not affect another.
3. **Constant Rate (λ)**:

The Poisson distribution is suitable when you are modeling the **number of events occurring in a fixed interval** of time, space, or any continuous domain, under the following conditions:

1. **Random Events**: Events occur randomly without predictable patterns.
   Example: Number of cars passing through a toll booth in an hour.
2. **Independent Events**: Each event is independent of others.
   Example: The arrival of customers at a store.
3. **Constant Average Rate**: The average rate of events (λ) is fixed for the interval.
   Example: Number of emails received per hour.
4. **Rare Events**: Events happen infrequently relative to the size of the interval.
   Example: Number of system crashes in a month.
5. **Discrete Counts**: The number of events can only be whole numbers (0,1,2,…0, 1, 2, …0,1,2,…).
   Example: Number of typing errors in a document.

Question6: Define the terms "probability distribution" and "probability density function" (PDF). How does a PDF differ from a probability mass function (PMF)?

Answer - **Probability Distribution**

A **probability distribution** describes how probabilities are assigned to all possible outcomes of a random variable. It provides a mathematical function that gives the likelihood of each outcome.

- **Discrete Variables**: Outcomes are distinct (e.g., rolling a die).
- **Continuous Variables**: Outcomes are within a range (e.g., height of people).

Probability Density Function (PDF)

A **Probability Density Function (PDF)** is used for **continuous random variables**. It shows the likelihood of a variable falling within a specific range of values.

- The area under the PDF curve between two points gives the probability of the variable being in that range.
- The **total area under the curve is always 1**.
-

Probability Mass Function (PMF)

A **Probability Mass Function (PMF)** is used for **discrete random variables**. It assigns probabilities to individual outcomes.

- Each probability is a specific value.
- The **sum of all probabilities is 1**.

Key Differences Between PDF and PMF

Aspect	PDF	PMF
Type of Variable	Continuous	Discrete
Output	Density (not probability directly)	Exact probability
Example	Heights of people	Rolling a die
Probability Calculation	Area under curve	Sum of probabilities

Question7: Explain the Central Limit Theorem (CLT) with example.

Answer - Central Limit Theorem (CLT)

The Central Limit Theorem (CLT) states that, regardless of the shape of the population distribution, the sampling distribution of the sample mean will approximate a normal distribution as the sample size becomes large (typically n>30).

Scenario:

Suppose the heights of a population are right-skewed, with a mean height (μ\muμ) of 165 cm and a standard deviation (σ\sigmaσ) of 10 cm.

Steps:

1. **Randomly select a sample of 50 people (n=50n = 50n=50).**
2. **Compute the sample mean for this group.**
3. **Repeat this process several times to create a distribution of sample means.**

Result:

- **The distribution of these sample means will be approximately normal, even though the original population was skewed.**
- **The mean of the sample means will still be 165 cm.**

Question8: Compare z-scores and t-scores. When should you use a z-score, and when should a t-score be a pplied instead?

Z-Scores vs. T-Scores: Simple Explanation

Z-Score

- **What it is: Measures how far a value is from the average, using the population's standard deviation.**
- **When to use:**
  - **When the population standard deviation is known.**
  - **When you have a large sample (usually more than 30 data points).**

T-Score

- **What it is: Similar to a z-score, but used when the population's standard deviation is unknown and we use the sample's standard deviation instead.**
- **When to use:**
  - **When the population standard deviation is unknown.**
  - **When you have a small sample (usually 30 or fewer data points).**

Quick Comparison

Factor	**Z-Score**	**T-Score**
Population Std Dev	**Known**	**Unknown**
Sample Size	**Large (n > 30)**	**Small (n ≤ 30)**

Summary:

- **Z-score: Use for large samples or when you know the population's standard deviation.**
- **T-score: Use for small samples or when you don't know the population's standard deviation.**

Question9: Given a sample mean of 105, a population mean of 100, a standard deviation of 15, and a sample size of 25, calculate the z-score and p-value. Based on a significance level of 0.05, do you reject or fail to reject the null hypothesis? Task: Write Python code to calculate the z-score and p-value for the given data. Objective: Apply the formula for the z-score and interpret the p-value for hypothesis testing.

Answer - https://github.com/Abhishek-D8mik3/Assignments

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Z-Score Calculation: The formula is applied directly to calculate the z-score.

P-Value Calculation: The cumulative distribution function (CDF) from scipy.stats.norm.cdf() gives the p-value based on the z-score.

Decision: If the p-value is less than the significance level (0.05), we reject the null hypothesis; otherwise, we fail to reject it.

Question10: Simulate a binomial distribution with 10 trials and a probability of success of 0.6 using Python. Generate 1,000 samples and plot the distribution. What is the expected mean and variance? Task: Use Python to generate the data, plot the distribution, and calculate the mean and variance. Objective: Understand the properties of a binomial distribution and verify them through simulation.

Answer – https://github.com/Abhishek-D8mik3/Assignments

Explanation:

1. **Simulate the Data:**
   - **We use np.random.binomial(n_trials, prob_success, n_samples) to generate 1,000 samples where each sample has 10 trials with a 60% chance of success.**
2. **Plotting:**
   - **We create a histogram to visualize the distribution of successes across the 1,000 samples.**
3. **Calculating the Mean and Variance:**
   - **The mean of the binomial distribution should be $n \times p = 10 \times 0.6 = 6$**
   - **The variance should be $n \times p \times (1-p) = 10 \times 0.6 x 0.4 = 2.4$**
   - **We use np.mean() and np.var() to calculate these values from the generated samples.**