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

The **z-statistic** is a standardized score that indicates how many standard deviations a data point is from the mean of a population. It is calculated using the formula:

z=(X−μ)\σ

where:

* X is the observed value,
* μ is the population mean, and
* σ is the population standard deviation.

**Relationship to the Standard Normal Distribution:**

The z-statistic follows a standard normal distribution (mean of 0 and standard deviation of 1). This allows for the comparison of data from different distributions by converting them to a common scale.

**Use in Hypothesis Testing:**

In hypothesis testing, the z-statistic is used to determine how far the sample mean is from the hypothesized population mean under the null hypothesis. It helps to:

1. Calculate the probability of observing a sample mean (or more extreme) if the null hypothesis is true.
2. Compare the z-statistic to critical values from the standard normal distribution to decide whether to reject or fail to reject the null hypothesis. If the p-value associated with the z-statistic is less than the significance level (e.g., 0.05), the null hypothesis is rejected, indicating a statistically significant effect.

**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)?**

A **p-value** is the probability of obtaining results as extreme as, or more extreme than, the observed data, assuming that the null hypothesis is true. It quantifies the evidence against the null hypothesis in hypothesis testing.

**Use in Hypothesis Testing:**

1. **Comparison to Significance Level**: The p-value is compared to a predetermined significance level (e.g., 0.05).
   * If the p-value is less than or equal to the significance level, the null hypothesis is rejected, suggesting a statistically significant effect.
   * If the p-value is greater than the significance level, we fail to reject the null hypothesis.

**Interpretation of a Small p-value (e.g., 0.01):**

A very small p-value (e.g., 0.01) indicates strong evidence against the null hypothesis. It suggests that the observed data is highly unlikely under the assumption that the null hypothesis is true. Consequently, researchers would typically reject the null hypothesis, concluding that there is a statistically significant effect or difference.

**Question3: Compare and contrast the binomial and Bernoulli distributions.**

**Comparison of Binomial and Bernoulli Distributions**:

**Bernoulli Distribution:**

* **Definition**: A Bernoulli distribution represents a single trial with two possible outcomes: success (1) or failure (0).
* **Parameters**: It is characterized by a single parameter p, the probability of success.
* **Support**: The outcomes are {0, 1}.

**Binomial Distribution:**

* **Definition**: A Binomial distribution represents the number of successes in n independent Bernoulli trials.
* **Parameters**: It is characterized by two parameters: n (number of trials) and p (probability of success in each trial).
* **Support**: The outcomes range from 0 to n (i.e., {0, 1, 2, ..., n}).

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

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

1. **Fixed Number of Trials**: The experiment consists of a fixed number of independent trials, denoted as n.
2. **Two Outcomes**: Each trial has only two possible outcomes: success (1) and failure (0).
3. **Constant Probability**: The probability of success, p, remains the same for each trial.
4. **Independent Trials**: The trials are independent; the outcome of one trial does not affect the others.

**Relation to Bernoulli Distribution:**

The binomial distribution is essentially a series of Bernoulli trials. A Bernoulli distribution represents a single trial, while the binomial distribution aggregates the outcomes of n Bernoulli trials. In other words, the binomial distribution counts the number of successes across multiple Bernoulli trials, making it a generalization of the Bernoulli distribution where n can be greater than 1.

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

**Key Properties of the Poisson Distribution:**

1. **Discrete Distribution**: The Poisson distribution models the number of events occurring in a fixed interval of time or space.
2. **Parameter**: It is characterized by a single parameter λ (lambda), which represents the average rate of occurrence of events.
3. **Non-Negative Values**: The possible outcomes are non-negative integers (0, 1, 2, ...).
4. **Memoryless Property**: The probability of an event occurring in a future interval is independent of past occurrences.
5. **Mean and Variance**: Both the mean and variance of a Poisson distribution are equal to λ.

**When to Use the Poisson Distribution:**

The Poisson distribution is appropriate when:

* Events occur independently and randomly.
* The average rate of occurrence λ is known.
* The events are rare relative to the observation period or space (e.g., the number of emails received per hour, the number of phone calls at a call center in a day).

It is commonly used in fields such as telecommunications, traffic flow, and natural events modeling.

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

**Probability Distribution:**

A **probability distribution** is a mathematical function that describes the likelihood of different outcomes in a random experiment. It assigns probabilities to each possible outcome or range of outcomes. Probability distributions can be either discrete or continuous.

**Probability Density Function (PDF):**

A **probability density function (PDF)** is used for continuous random variables. It describes the relative likelihood of a random variable taking on a particular value. The area under the PDF curve over a specified interval represents the probability of the variable falling within that interval.

**Difference Between PDF and PMF:**

* **Probability Mass Function (PMF)**: Used for discrete random variables, the PMF gives the probability that a discrete random variable equals a specific value. The sum of all probabilities in a PMF is 1.
* **PDF**: Used for continuous random variables, the PDF does not give probabilities for specific values (since the probability of any exact value is zero) but rather gives probabilities for ranges of values through integration. The total area under the PDF curve equals 1.

In summary, PMFs apply to discrete outcomes, while PDFs apply to continuous outcomes, reflecting the nature of the random variable being modeled.

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

The **Central Limit Theorem (CLT)** states that, regardless of the original distribution of a population, the distribution of the sample means will approach a normal distribution as the sample size increases, typically around n≥30. This means that for large enough sample sizes, the sampling distribution of the mean will be approximately normally distributed with a mean equal to the population mean and a standard deviation equal to the population standard deviation divided by the square root of the sample size (σ/√n​).

**Example:**

Suppose you have a population of people with heights that are not normally distributed. If you take random samples of size 30 (or more) and calculate the mean height for each sample, the distribution of those sample means will tend to be normal, even if the original height distribution is skewed or otherwise non-normal.

This property of the CLT allows statisticians to make inferences about population parameters using sample data, enabling hypothesis testing and confidence interval estimation even when the underlying population distribution is unknown.

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

**Comparison of z-scores and t-scores:**

* **Z-scores**:
  + **Definition**: A z-score measures how many standard deviations an individual data point is from the mean of a population.
  + **Use Case**: Used when the population standard deviation (σ) is known or when the sample size is large (n≥30).
  + **Distribution**: Follows the standard normal distribution (mean of 0 and standard deviation of 1).
* **T-scores**:
  + **Definition**: A t-score measures how many standard deviations a sample mean is from the population mean, accounting for sample variability.
  + **Use Case**: Used when the population standard deviation is unknown and the sample size is small (n<30).
  + **Distribution**: Follows the t-distribution, which is wider and has heavier tails compared to the normal distribution, especially for smaller sample sizes.

**When to Use:**

* **Use a z-score** when:
  + The population standard deviation is known.
  + The sample size is large (n≥30).
* **Use a t-score** when:
  + The population standard deviation is unknown.
  + The sample size is small (n<30).

In summary, z-scores are preferred for large samples with known population parameters, while t-scores are suitable for small samples with unknown population parameters.

**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.**

Here’s how to calculate the z-score and p-value based on the given data:

**Calculations:**

1. **Z-score**:

z=(X−μ)/(σ/√n)=(105−100)/(15/√25)=5/3≈1.67

1. **P-value**: The p-value can be calculated using the z-score in a two-tailed test:

p-value=2×(1−CDF(z))=2×(1−CDF(1.67))

**Decision:**

Based on a significance level of **0.05**:

* If the p-value is less than or equal to 0.05, we reject the null hypothesis.
* If the p-value is greater than 0.05, we fail to reject the null hypothesis.

Once you calculate the p-value, you can use it to make your decision.

**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.**

Here's how to simulate a binomial distribution with 10 trials and a probability of success of 0.6, and calculate the expected mean and variance:

**Steps:**

1. **Simulate the Binomial Distribution**:
   * **Trials**: 10
   * **Probability of Success**: 0.6
   * **Number of Samples**: 1,000
2. **Expected Mean and Variance**:
   * **Mean**: Mean=n×p=10×0.6=6
   * **Variance**: Variance=n×p×(1−p)=10×0.6×0.4=2.4
3. **Plotting**:
   * Use a histogram to visualize the distribution of successes from the simulated samples.
   * Add vertical lines for the mean and standard deviation.

**Conclusion:**

* The **expected mean** of the binomial distribution is **6**.
* The **expected variance** is **2.4**.

Bottom of Form