Statics Assignment no.2

Q1:

The **Z statistic** (or **Z score**) is a measure used to indicate how many standard deviations an element is from the mean of a distribution. In the context of hypothesis testing, it is primarily associated with the **standard normal distribution**, which has a mean of 0 and a standard deviation of 1.

Z Statistic Formula

The Z statistic is calculated as:

\[

Z = \frac{(X - \mu)}{\sigma}

\]

Where:

- \(X \) = observed value (or sample mean)

- \(\mu \) = population mean (hypothesized mean under the null hypothesis)

- \(\sigma \) = population standard deviation (or standard error for sample means)

This formula standardizes the data, meaning it converts the observed value into a number of standard deviations away from the mean.

Relationship to the Standard Normal Distribution

The Z statistic is a point on the **standard normal distribution**, which is a normal distribution with a mean of 0 and a standard deviation of 1. The Z statistic allows us to map any normal distribution onto this standard normal curve. Once the Z value is computed, we can use standard normal distribution tables (or functions in statistical software) to find probabilities associated with the Z value.

For example:

- A Z score of 0 corresponds to the mean.

- A Z score of +1.96 corresponds to the 97.5th percentile, meaning 97.5% of the data lies below this value in a two-tailed test.

Z Statistic in Hypothesis Testing

In hypothesis testing, the Z statistic helps to determine how far a sample mean (or observed value) deviates from the hypothesized population mean, in units of standard deviation.

1. **Formulate Hypotheses**:

   - Null hypothesis (\(H_0 \)): The sample mean is equal to the population mean.

   - Alternative hypothesis (\(H_A \)): The sample mean is different from the population mean.

2. **Calculate Z Statistic**:

   The Z statistic is calculated to assess whether the observed sample mean differs significantly from the hypothesized mean. This involves comparing the sample mean with the population mean, standardized by the population standard deviation (or the standard error for sample means).

3. **Determine the P-value**:

   Using the Z statistic, a **P-value** can be obtained from the standard normal distribution. The P-value represents the probability of observing a test statistic as extreme as the one calculated, assuming the null hypothesis is true.

4. **Decision Making**:

   - If the **P-value** is smaller than the significance level (commonly \(\alpha = 0.05 \)), the null hypothesis is rejected, suggesting the sample provides enough evidence to support the alternative hypothesis.

   - If the P-value is greater than the significance level, there is insufficient evidence to reject the null hypothesis.

Example of Z Statistic in Use:

In a hypothesis test with \(H_0: \mu = 50 \), if a sample of size \(n = 100 \) has a sample mean of 52, and the population standard deviation is known to be 4, the Z statistic would be:

\[

Z = \frac{(52 - 50)}{4 / \sqrt{100}} = \frac{2}{0.4} = 5

\]

This Z value corresponds to a very small P-value, meaning the null hypothesis would be rejected.

In summary, the Z statistic standardizes the difference between the observed data and the null hypothesis, and its use in hypothesis testing allows for a decision based on the likelihood of observing data under the null hypothesis.

Q2:

A **P-value** in hypothesis testing is the probability of obtaining test results at least as extreme as the observed results, assuming that the **null hypothesis** is true. It quantifies the strength of evidence against the null hypothesis.

P-value in Hypothesis Testing

The P-value helps determine whether to reject or fail to reject the null hypothesis. It does this by comparing the observed results (e.g., sample mean, test statistic) to what would be expected under the assumption that the null hypothesis is true.

Here’s how it fits into the hypothesis testing process:

1. **Formulate Hypotheses**:

   - Null Hypothesis (\(H_0\)): This is the assumption that there is no effect or difference (e.g., no difference between sample mean and population mean).

   - Alternative Hypothesis (\(H_A\)): This is what you suspect might be true instead of the null hypothesis (e.g., the sample mean is different from the population mean).

2. **Collect Data and Calculate a Test Statistic**:

   Depending on the test (e.g., Z-test, t-test), you calculate a statistic that measures the distance between the sample result and what the null hypothesis predicts.

3. **Calculate the P-value**:

   The P-value is derived from the test statistic and indicates how extreme your test statistic is under the assumption that the null hypothesis is true.

4. **Compare the P-value to the Significance Level (\(\alpha\))**:

   - The significance level (\(\alpha\)) is a threshold set by the researcher, commonly 0.05, which represents a 5% risk of rejecting the null hypothesis when it is actually true (Type I error).

   - If the P-value is **less than or equal to** \(\alpha\), you reject the null hypothesis.

   - If the P-value is **greater than** \(\alpha\), you fail to reject the null hypothesis.

Interpreting the P-value

- **Small P-value (e.g., 0.01 or less)**: A small P-value indicates that the observed data is very unlikely under the null hypothesis. This suggests strong evidence **against the null hypothesis**, leading to its rejection. For example, a P-value of 0.01 means there is a 1% chance that the observed data would occur if the null hypothesis were true.

- **Large P-value (e.g., 0.20 or more)**: A large P-value indicates that the observed data is more consistent with the null hypothesis, providing insufficient evidence to reject it.

Example of a Small P-value (e.g., 0.01)

If the P-value in a hypothesis test is **0.01**, it means there is a 1% probability of observing a result as extreme as the one obtained (or more extreme) assuming the null hypothesis is true.

Interpretation:

- If the significance level \(\alpha\) is set to 0.05 (a common choice), then a P-value of 0.01 would lead you to **reject the null hypothesis**, as the P-value is smaller than the significance level.

- A P-value of 0.01 indicates strong evidence against the null hypothesis because the probability of observing such an extreme result under the null hypothesis is very low (only 1%).

Conclusion:

- **Small P-value (e.g., 0.01)**: Strong evidence against the null hypothesis. You reject the null hypothesis, suggesting that the alternative hypothesis is more likely to be true.

- **Large P-value**: Weak evidence against the null hypothesis, leading to failure to reject it.

Example:

In a study testing whether a new drug has a different effect than a placebo, if the P-value is 0.01, there is a 1% chance that the observed effect (or a larger one) could have occurred by random chance under the null hypothesis (no difference between the drug and placebo). If this P-value is smaller than the significance level (e.g., 0.05), you would conclude that the drug likely has a significant effect.

Q3:

The **Binomial** and **Bernoulli** distributions are closely related probability distributions that model binary (yes/no, success/failure) outcomes. However, they have distinct characteristics and applications.

1. **Definition**

- **Bernoulli Distribution**:

  - The **Bernoulli distribution** is the probability distribution of a random variable that has **only two possible outcomes**: success (1) or failure (0).

  - It describes a **single trial** of a binary experiment.

  - The probability of success is denoted by \(p \), and the probability of failure is \(1 - p \).

  Example: Flipping a coin once (heads = success, tails = failure).

- **Binomial Distribution**:

  - The **Binomial distribution** describes the number of successes in a fixed number of **independent** Bernoulli trials.

  - It involves repeating a binary experiment (like the Bernoulli process) **multiple times**.

  - The distribution depends on two parameters: \(n \) (the number of trials) and \(p \) (the probability of success in each trial).

  Example: Flipping a coin 10 times and counting the number of heads (successes).

2. **Probability Mass Function (PMF)**

- **Bernoulli Distribution PMF**:

  For a Bernoulli random variable \(X \), the PMF is given by:

\[

P(X = x) =

\begin{cases}

p & \text{if } x = 1 \\

1 - p & \text{if } x = 0

\end{cases}

\]

Where:

- \(p \) = probability of success

- \(1 - p \) = probability of failure

- **Binomial Distribution PMF**:

For a Binomial random variable \(X \), the PMF is:

\[

P(X = k) = \binom{n}{k} p^k (1 - p)^{n - k}

\]

Where:

- \(n \) = number of trials

- \(k \) = number of successes

- \(p \) = probability of success

- \(\binom{n}{k} \) is the binomial coefficient, representing the number of ways to choose \(k \) successes in \(n \) trials.

3. **Key Differences**

| Feature | Bernoulli Distribution | Binomial Distribution |

|-------------------------|---------------------------------------|--------------------------------------|

| **Number of trials** | 1 (single trial) | \(n \) trials (multiple trials) |

| **Random variable** | 0 or 1 (success/failure) | 0 to \(n \) (number of successes) |

| **PMF** | \(P(X = x) = p^x (1 - p)^{1 - x} \) | \(P(X = k) = \binom{n}{k} p^k (1 - p)^{n - k} \) |

| **Mean** (\(\mu \)) | \(\mu = p \) | \(\mu = np \) |

| **Variance** (\(\sigma^2 \)) | \(\sigma^2 = p(1 - p) \) | \(\sigma^2 = np(1 - p) \) |

| **Use case** | Describes a single binary outcome | Describes the total number of successes in multiple trials |

4. **Relationship Between the Two**

- A **Bernoulli distribution** is a special case of the **Binomial distribution** where \(n = 1 \).

- In other words, a Binomial distribution with \(n = 1 \) and probability \(p \) is the same as a Bernoulli distribution with parameter \(p \).

5. **Applications**

- **Bernoulli Distribution**:

  - Used when we are only interested in the outcome of a **single binary experiment**.

  - **Examples**:

    - Did a person purchase a product (yes/no)?

    - Did a student pass an exam (pass/fail)?

- **Binomial Distribution**:

  - Used when we repeat the binary experiment multiple times and want to count the total number of successes.

  - **Examples**:

    - How many students passed out of 10 who took an exam?

    - How many customers made a purchase out of 100 who visited a store?

6. **Example to Illustrate**

- **Bernoulli Example**:

  - Suppose a light bulb is tested to see if it works or not. The probability that it works is \(p = 0.9 \). This is modeled by a Bernoulli distribution.

  - The random variable \(X \) takes values 1 (light works) or 0 (light doesn't work).

- **Binomial Example**:

  - Now suppose we test 10 light bulbs. The number of bulbs that work can be modeled by a Binomial distribution with \(n = 10 \) and \(p = 0.9 \).

  - The random variable \(Y \) takes values from 0 to 10, representing the number of successful outcomes (bulbs that work).

Summary of Key Points:

- The **Bernoulli distribution** models a single trial with two outcomes.

- The **Binomial distribution** models multiple independent trials and counts the number of successes.

- A **Binomial distribution** with \(n = 1 \) is essentially a **Bernoulli distribution**.

Q4:

The **Binomial distribution** is used under specific conditions that involve repeated, independent binary trials. Here's a breakdown of when and how it is used, as well as its relationship to the **Bernoulli distribution**.

Conditions for Using a Binomial Distribution

A random variable \(X \) is said to follow a **Binomial distribution** if **all** of the following conditions are met:

1. **Fixed Number of Trials**:

   - The experiment must be repeated a **fixed number of times**, denoted by \(n \).

   - Each trial represents an independent instance of the experiment.

   Example: Flipping a coin 10 times or conducting 20 quality checks in a factory.

2. **Binary Outcomes (Success/Failure)**:

   - Each trial must have only **two possible outcomes**: "success" (often coded as 1) or "failure" (often coded as 0).

   Example: Success = getting heads on a coin flip, Failure = getting tails.

3. **Constant Probability of Success**:

   - The probability of success, \(p \), remains **constant** for every trial.

   - Similarly, the probability of failure is \(1 - p \) and is also constant across trials.

   Example: The probability of getting heads on a fair coin flip is always \(p = 0.5 \) across all flips.

4. **Independent Trials**:

   - The outcome of any trial must be **independent** of the outcomes of other trials. In other words, the result of one trial does not affect the probability of outcomes in other trials.

   Example: The outcome of one coin flip doesn't influence the outcome of the next flip.

Relationship to the Bernoulli Distribution

The **Bernoulli distribution** is the foundation of the **Binomial distribution**, and they are closely related in the following ways:

1. **Single Trial vs. Multiple Trials**:

   - The **Bernoulli distribution** models the outcome of a **single** trial, which has only two possible outcomes (success or failure).

   - The **Binomial distribution** generalizes this to **multiple** independent Bernoulli trials, where the random variable \(X \) counts the number of successes in \(n \) trials.

2. **Binomial as a Sum of Bernoulli Trials**:

   - A **Binomial distribution** can be thought of as the sum of independent **Bernoulli** trials. If you perform \(n \) independent Bernoulli trials, each with success probability \(p \), then the total number of successes follows a Binomial distribution.

Mathematically, if \(X_i \) are independent Bernoulli random variables representing the outcome of the \(i \)-th trial, then:

\[

X = X_1 + X_2 + \dots + X_n

\]

Where \(X \) is the Binomial random variable representing the total number of successes.

3. **Binomial Distribution for \(n = 1 \)**:

- A **Bernoulli distribution** is a special case of the **Binomial distribution** where the number of trials \(n = 1 \).

- In this case, the Binomial distribution with \(n = 1 \) becomes a Bernoulli distribution with probability \(p \).

Example:

Suppose we are flipping a fair coin.

- **Bernoulli distribution**: Flipping the coin once, where "success" (heads) occurs with \(p = 0.5 \) and "failure" (tails) occurs with \(1 - p = 0.5 \), is a Bernoulli process.

- **Binomial distribution**: If we flip the coin 10 times, the number of heads (successes) we get out of 10 flips follows a Binomial distribution with parameters \(n = 10 \) and \(p = 0.5 \).

Summary

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

- The **Binomial distribution** describes the **number of successes** in a fixed number of **independent Bernoulli trials**.

- The Binomial distribution generalizes the Bernoulli process by allowing multiple trials, and for \(n = 1 \), the Binomial distribution is equivalent to the Bernoulli distribution.

Q5:

The **Poisson distribution** is a probability distribution used to model the number of events that occur in a fixed interval of time, space, or any other dimension, provided these events occur **randomly and independently**. It is particularly useful for modeling rare or infrequent events.

Key Properties of the Poisson Distribution

1. **Discrete Distribution**:

   - The Poisson distribution applies to **discrete random variables**, meaning it models countable events (e.g., number of accidents, number of emails received, etc.).

2. **Probability Mass Function (PMF)**:

   The PMF of the Poisson distribution is given by:

   \[

   P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}

   \]

   Where:

   - \(X \) = random variable representing the number of events.

   - \(k \) = the number of occurrences of the event.

   - \(\lambda \) = the **average rate** (mean number of events) per time unit, spatial area, etc.

   - \(e \) = Euler's number (\(\approx 2.718 \)).

3. **Mean and Variance**:

   - The mean of the Poisson distribution is \(\lambda \), which represents the **average number of occurrences** of the event in the given interval.

   - The variance of the Poisson distribution is also \(\lambda \). This unique property implies that for a Poisson distribution, the mean and variance are equal.

   \[

   \text{Mean} = \lambda, \quad \text{Variance} = \lambda

   \]

4. **Memorylessness**:

- The Poisson distribution assumes that events occur **independently** of one another. The occurrence of an event at one point in time does not affect the likelihood of another event occurring in the next time period.

5. **Skewness**:

- The Poisson distribution is positively **skewed**, especially when \(\lambda \) is small. As \(\lambda \) increases, the distribution becomes more symmetric and resembles a normal distribution (due to the Central Limit Theorem).

6. **Additivity**:

- If two independent Poisson random variables have rates \(\lambda_1 \) and \(\lambda_2 \), then the sum of these two Poisson variables is also a Poisson-distributed variable with rate \(\lambda_1 + \lambda_2 \).

\[

X_1 \sim \text{Poisson}(\lambda_1), \quad X_2 \sim \text{Poisson}(\lambda_2) \quad \Rightarrow \quad X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)

\]

When to Use the Poisson Distribution

The Poisson distribution is appropriate to use when the following conditions are met:

1. **Events Occur Randomly**:

- The events you are modeling must occur randomly in time or space. Each event happens without a predictable pattern.

2. **Independent Events**:

- The occurrence of an event should not influence or affect the occurrence of another event. For example, the fact that one car accident has occurred should not affect the likelihood of another one occurring in a different time interval or place.

3. **Constant Average Rate**:

- The average number of occurrences, \(\lambda \), must be constant across the fixed interval of time or space being studied. This means that the events happen at a constant rate on average (e.g., 5 emails per hour).

4. **Rare or Low-Probability Events**:

   - The Poisson distribution is commonly used to model **rare events** in a large population or over a large period of time. For example, the number of phone calls a company receives in a minute, or the number of accidents occurring in a city in a month, when these events are rare.

5. **Small Probability per Event, Large Number of Trials**:

   - The Poisson distribution is often an approximation for a **Binomial distribution** when the probability of success in each trial is very small, and the number of trials is large. This is known as the **Poisson limit theorem**.

   If \(n \) is large and \(p \) (the probability of success in each trial) is small, and the product \(np \) is finite (i.e., \(np = \lambda \)), then a Binomial distribution \(B(n, p) \) can be approximated by a Poisson distribution \(\text{Poisson}(\lambda) \).

Practical Applications of the Poisson Distribution

1. **Count of Events Over Time**:

   - **Example**: The number of phone calls received at a call center per hour, the number of emails arriving in your inbox per day, or the number of accidents on a highway in a month.

2. **Count of Events in Space**:

   - **Example**: The number of trees in a forest per square kilometer, or the number of bacteria colonies in a petri dish per square centimeter.

3. **Rare Event Modeling**:

   - **Example**: The number of equipment failures in a factory during a week, or the number of customers arriving at a store in an hour.

4. **Modeling Arrival Processes**:

   - The Poisson distribution is often used in queuing theory to model the number of arrivals (e.g., customers, packets in a network, etc.) to a system in a fixed amount of time.

Example of Use:

Imagine a **customer support center** receives, on average, 3 phone calls per hour. We want to know the probability of receiving exactly 5 calls in the next hour. Since the number of calls per hour is a random, discrete event, we can use the Poisson distribution with $\lambda = 3$.

Using the Poisson formula:

$$
P(X = 5) = \frac{3^5 e^{-3}}{5!} = \frac{243 \times e^{-3}}{120} \approx 0.1008
$$

So, there is a **10.08%** chance of receiving exactly 5 calls in the next hour.

Summary

- The **Poisson distribution** models the number of events occurring in a fixed interval of time or space, with a constant rate and independent events.
- It is particularly useful when dealing with **rare events** or counts of events in fixed intervals.
- The key conditions are random, independent occurrences, with a constant average rate of events per interval.

Q6:

Probability Distribution

A **probability distribution** describes how the values of a random variable are distributed. It gives the probabilities or likelihoods of all possible outcomes of a random variable in a certain range or set of values. Probability distributions can be divided into two types, depending on whether the random variable is **discrete** or **continuous**.

- For **discrete** random variables, the distribution specifies the probability for each possible outcome.
- For **continuous** random variables, the distribution specifies the probability density over a range of values.

In both cases, the total probability for all possible outcomes must sum (or integrate) to 1.

Probability Density Function (PDF)

A **Probability Density Function (PDF)** is used for **continuous** random variables. It describes the likelihood (density) of the random variable taking on a particular value within a continuous range. A PDF does not directly give probabilities for specific values but rather indicates how "dense" the probability is around any point. To find the probability of the variable falling within a certain interval, you integrate the PDF over that interval.

For a random variable X with a PDF $f(x)$, the probability that X falls between a and b is given by:

$$
P(a \leq X \leq b) = \int_a^b f(x) \, dx
$$

Key properties of a PDF:

- $f(x) \geq 0$ for all x.
- The total area under the PDF curve equals 1:

  $$
  \int_{-\infty}^{\infty} f(x) \, dx = 1
  $$

Probability Mass Function (PMF)

A **Probability Mass Function (PMF)** is used for **discrete** random variables. It gives the probability that a discrete random variable takes on a specific value. Unlike a PDF, a PMF assigns a probability to each possible discrete value directly.

For a discrete random variable X with a PMF $P(X = x_i) = p(x_i)$, the probabilities for each possible value x_i must satisfy:

$$
0 \leq p(x_i) \leq 1, \quad \text{and} \quad \sum_{i} p(x_i) = 1
$$

Key properties of a PMF:

- It gives exact probabilities for each discrete outcome.
- The sum of all probabilities for all possible outcomes must equal 1.

Difference Between PDF and PMF

Feature	PDF (Probability Density Function)	PMF (Probability Mass Function)
Type of Random Variable	Continuous	Discrete
Probability for a Specific Value	$P(X = x) = 0$ for any single value (since continuous)	$P(X = x_i) > 0$ for specific values
Describes	Probability density over an interval	Probability of specific, discrete outcomes
Total Probability	The area under the curve equals 1	The sum of probabilities equals 1
Finding Probabilities	Requires integration over an interval	Sum probabilities directly for each outcome

Example to Illustrate:

- **PDF (Continuous Random Variable)**: The height of adult males in a population is often modeled as a continuous random variable, say X, with a normal distribution. The PDF shows the likelihood of different height ranges, and to find the probability that a man is between 170 cm and 180 cm tall, you would integrate the PDF over that interval.

- **PMF (Discrete Random Variable)**: The number of heads in 10 flips of a fair coin is a discrete random variable, say Y, which can take values from 0 to 10. The PMF gives the probability for each specific outcome (e.g., $P(Y = 5)$, the probability of getting exactly 5 heads, is calculated directly).

Q7:

The **Central Limit Theorem (CLT)** is a fundamental concept in statistics. It states that, given a sufficiently large sample size, the distribution of the sample mean will approach a **normal distribution** (Gaussian distribution), **regardless of the shape of the population distribution**, as long as the population has a finite mean and variance.

Key Points of CLT:

1. **Sample Size**: The sample size must be large enough (usually \(n \geq 30 \) is considered sufficient).

2. **Independence**: The samples should be independent of each other.

3. **Population Distribution**: The population from which the samples are drawn can be of any shape (skewed, uniform, etc.), but the sample mean will follow a normal distribution as the sample size increases.

Why is CLT important?

The CLT allows statisticians to make inferences about population parameters using sample data. Even if the original population distribution is unknown or non-normal, the sample mean will tend to follow a normal distribution for large sample sizes, enabling the use of standard statistical techniques (like confidence intervals or hypothesis testing).

Example:

Imagine we want to study the average height of all students in a university. The true distribution of heights might be skewed (for example, more students might be taller). Now, if we:

1. Take a random sample of 10 students and calculate the average height (sample mean).

2. Repeat this process of sampling and calculating the sample mean many times.

3. Plot the distribution of these sample means.

What the CLT tells us is that **even if the original height distribution is skewed**, the distribution of the sample means will approximate a normal distribution **as the number of samples increases**.

Simulation Example:

Suppose we have a **right-skewed** population distribution, such as a population with incomes (where a few very high incomes skew the data). We could:

- Randomly sample 50 people from this population.

- Calculate the average income of this sample.

- Repeat this many times (say 1000 times).

As we plot the distribution of the sample means (from each sample), we would observe that this distribution becomes increasingly **bell-shaped** (i.e., normal distribution), even though the original population was skewed.

Thus, the **CLT** provides the theoretical foundation for using sample means and variances in inference, enabling us to apply normal distribution techniques (like Z-tests) to a wide range of problems.

Q8:

Z-scores and **T-scores** (or t-values) are both used in statistics to understand how far a data point or sample mean is from the population mean, measured in terms of standard deviations or standard errors. However, they are used in different scenarios based on sample size and whether the population standard deviation is known.

Z-score:

A **Z-score** tells you how many standard deviations a data point (or sample mean) is from the population mean. It is used when:

- The population standard deviation (\(\sigma \)) is **known**.

- The sample size (\(n \)) is **large** (usually \(n \geq 30 \)).

- The data is normally distributed or the Central Limit Theorem applies (for large sample sizes).

Formula for Z-score:

\[

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

\]

Where:

- X is the data point (or sample mean).

- μ is the population mean.

- σ is the population standard deviation.

T-score:

A **T-score** is used in a similar way to the Z-score but is applied when the population standard deviation (σ) is **unknown**, and especially when the sample size is **small** ($n < 30$). It compensates for the increased variability and uncertainty when using smaller samples.

The T-distribution is similar to the normal distribution but has heavier tails (i.e., more extreme values) to account for the increased variability in smaller samples. As the sample size increases, the T-distribution approaches the Z-distribution.

Formula for T-score:

$$
T = \frac{X - \mu}{s / \sqrt{n}}
$$

Where:

- X is the sample mean.

- μ is the population mean (or assumed mean under a null hypothesis).

- s is the sample standard deviation (used as an estimate of the population standard deviation).

- n is the sample size.

Key Differences:

Z-score	**T-score**
Used when population standard deviation is **known**.	Used when population standard deviation is **unknown**.
Used with **large samples** (usually $n \geq 30$).	Used with **small samples** ($n < 30$).
Follows a **normal distribution**.	Follows a **T-distribution** with $n-1$ degrees of freedom.

| Has **thinner tails** (less likely to observe extreme values). | Has **heavier tails** (accounts for more variability in small samples). |

When to use Z-score vs T-score:

- **Z-score**: Use when the population standard deviation (\(\sigma \)) is known, and the sample size is large (\(n \geq 30 \)).

- **T-score**: Use when the population standard deviation (\(\sigma \)) is unknown, and especially when the sample size is small (\(n < 30 \)). If the sample size is large, even without the population standard deviation, the T-distribution converges toward the Z-distribution, so Z-scores can be used in practice.

Example:

1. **Z-score scenario**: Suppose you have data on test scores for 500 students, and you know the population standard deviation of test scores is 15. You want to know how unusual a student's score of 85 is. You would use a Z-score because you know the population standard deviation and have a large sample.

2. **T-score scenario**: If you are conducting an experiment with only 15 participants, and you want to determine how the sample mean compares to a hypothesized population mean but don't know the population standard deviation, you would use a T-score because the sample is small, and the population standard deviation is unknown.

Q9:

Problem Breakdown:

You are given the following information:

- **Sample mean (\(\bar{X} \))** = 105

- **Population mean (\(\mu \))** = 100

- **Population standard deviation (\(\sigma \))** = 15

- **Sample size (\(n \))** = 25

- **Significance level (α)** = 0.05

We want to:

1. Calculate the **Z-score** using the given data.

2. Calculate the **p-value** based on this Z-score.

3. Perform hypothesis testing to decide whether to **reject or fail to reject the null hypothesis** at the 0.05 significance level.

Steps:

Step 1: Hypothesis Setup

We will perform a **two-tailed test** as we are comparing whether the sample mean is significantly different from the population mean.

- **Null Hypothesis \(H_0 \)**: The sample mean is equal to the population mean, i.e., \(\bar{X} = \mu \).

- **Alternative Hypothesis \(H_a \)**: The sample mean is not equal to the population mean, i.e., \(\bar{X} \neq \mu \).

Step 2: Z-Score Calculation

The formula for the Z-score is:

\[

Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}

\]

Where:

- \(\bar{X} \) is the sample mean.

- \(\mu \) is the population mean.

- \(\sigma \) is the population standard deviation.

- \(n \) is the sample size.

Let’s calculate the Z-score using the given values:

\[

Z = \frac{105 - 100}{15 / \sqrt{25}} = \frac{5}{15 / 5} = \frac{5}{3} = 1.67

\]

So, the **Z-score** is **1.67**.

Step 3: P-Value Calculation

The **p-value** tells us the probability of obtaining a result as extreme as, or more extreme than, the observed result under the null hypothesis.

Since this is a **two-tailed test**, we need to calculate the area in both tails. The p-value is calculated as:

$$
p = 2 \times (1 - \text{CDF}(Z))
$$

Where CDF is the cumulative distribution function of the standard normal distribution.

Using Python, we calculate:

```python
import scipy.stats as stats

# Given data
sample_mean = 105
population_mean = 100
std_dev = 15
sample_size = 25

# Step 1: Calculate the Z-score
z_score = (sample_mean - population_mean) / (std_dev / (sample_size ** 0.5))

# Step 2: Calculate the p-value (two-tailed test)
p_value = 2 * (1 - stats.norm.cdf(abs(z_score)))

z_score, p_value
```

The **p-value** is calculated to be approximately **0.096**.

Step 4: Hypothesis Testing Decision

- We compare the **p-value** (0.096) with the **significance level** (0.05).

- Since the p-value (0.096) is **greater than** 0.05, we **fail to reject the null hypothesis**.

Conclusion:

At the 0.05 significance level, we do not have enough evidence to conclude that the sample mean (105) is significantly different from the population mean (100). Thus, we **fail to reject the null hypothesis**.

This means that the observed difference between the sample mean and the population mean could be due to random sampling variation rather than a significant effect.

Q10:

Problem Breakdown:

We are asked to:

1. Simulate a **binomial distribution** with the following parameters:
    - **Number of trials (n)**: 10
    - **Probability of success (p)**: 0.6
    - **Number of samples**: 1000
2. Plot the distribution of the simulated data.
3. Calculate the **expected mean** and **variance** of the binomial distribution based on its theoretical properties.
4. Compare these theoretical values with the actual mean and variance derived from the simulated data.

Binomial Distribution Recap:

A **binomial distribution** models the number of successes in nnn independent trials, where each trial has two possible outcomes (success or failure) and the probability of success remains constant across trials.

Theoretical Mean and Variance of a Binomial Distribution:

For a binomial distribution with parameters nnn (number of trials) and ppp (probability of success), the theoretical mean μ\muμ and variance σ2\sigma^2σ2 are given by:

- **Mean (μ\muμ)**: μ=n×p\mu = n \times pμ=n×p
- **Variance (σ2\sigma^2σ2)**: σ2=n×p×(1−p)\sigma^2 = n \times p \times (1 - p)σ2=n×p×(1−p)

Step-by-Step Solution:

Step 1: Simulate the Binomial Distribution

We will use the numpy library to simulate the binomial distribution. We will generate 1000 samples where each sample represents the number of successes in 10 trials with a probability of success of 0.6.

Step 2: Calculate the Theoretical Mean and Variance

Using the formulas for the mean and variance of a binomial distribution:

- Theoretical mean = n×pn \times pn×p
- Theoretical variance = n×p×(1−p)n \times p \times (1 - p)n×p×(1−p)

Given:

- n=10n = 10n=10
- p=0.6p = 0.6p=0.6

We can calculate the theoretical mean and variance.

Step 3: Calculate the Simulated Mean and Variance

Using the simulated data, we calculate the mean and variance using Python's numpy functions mean() and var().

Step 4: Plot the Distribution

Using the matplotlib library, we will create a histogram to visualize the frequency of successes in the 1000 samples.

Python Code Implementation:

```
import numpy as np

import matplotlib.pyplot as plt


# Given parameters

n_trials = 10

p_success = 0.6

n_samples = 1000


# Step 1: Simulate 1000 samples of a binomial distribution

binom_samples = np.random.binomial(n=n_trials, p=p_success, size=n_samples)
```

```
# Step 2: Calculate the expected mean and variance
mean_theoretical = n_trials * p_success
variance_theoretical = n_trials * p_success * (1 - p_success)

# Step 3: Calculate the mean and variance of the simulated data
mean_simulated = np.mean(binom_samples)
variance_simulated = np.var(binom_samples)

# Step 4: Plot the distribution of the simulated data
plt.hist(binom_samples, bins=range(0, n_trials+2), edgecolor='black', alpha=0.7)
plt.title("Binomial Distribution (n=10, p=0.6) - 1000 Samples")
plt.xlabel("Number of Successes")
plt.ylabel("Frequency")
plt.show()

(mean_theoretical, variance_theoretical, mean_simulated, variance_simulated)
```

Output:

- **Theoretical Mean**: 6.0
- **Theoretical Variance**: 2.4
- **Simulated Mean**: 5.944
- **Simulated Variance**: 2.319

Analysis of the Results:

Theoretical vs Simulated Values:

- **Theoretical Mean**: The theoretical mean is calculated as n×p=10×0.6=6.0n \times p = 10 \times 0.6 = 6.0n×p=10×0.6=6.0. This is the expected number of successes in 10 trials if the probability of success in each trial is 0.6.
- **Theoretical Variance**: The theoretical variance is n×p×(1−p)=10×0.6×0.4=2.4n \times p \times (1 - p) = 10 \times 0.6 \times 0.4 = 2.4n×p×(1−p)=10×0.6×0.4=2.4. This represents the expected variation in the number of successes.

From the simulation:

- The **simulated mean** is very close to the theoretical mean (5.944 vs 6.0).
- The **simulated variance** is also close to the theoretical variance (2.319 vs 2.4).

Visualizing the Distribution:

- The histogram displays the distribution of the number of successes in 1000 trials. We see a distribution that is centered around 6, which corresponds to the theoretical mean.
- The binomial distribution is discrete, with integer outcomes from 0 to 10. The bars in the histogram show how frequently each number of successes occurred in the 1000 samples.

Conclusion:

The simulation matches well with the theoretical properties of the binomial distribution. The simulated mean and variance are close to their theoretical values, demonstrating that the binomial distribution follows its expected behavior in practice. The histogram provides a visual confirmation of the distribution's shape, with the center around the expected number of successes (6 in this case). This helps in understanding how the binomial distribution models real-world processes where we are counting successes across multiple trials.