Question 1: Define the z-statistic and explain its relationship to the standard normal distribution. How is the z-statistic used in hypothesis testing?

A **z-statistic** (also called a z-score) measures how many standard deviations a data point or sample mean is from the population mean. The formula for the z-statistic is:

Z=X−μσZ = \frac{X - \mu}{\sigma}Z=σX−μ

Where:

- XXX = value or sample mean
- μ\muμ = population mean
- σ\sigmaσ = population standard deviation

The z-statistic follows the **standard normal distribution**, which has a mean of 0 and a standard deviation of 1. The area under this curve represents probabilities, and the z-statistic is used to determine how likely or unlikely a value is, assuming a normal distribution.

In **hypothesis testing**, the z-statistic helps assess the difference between a sample statistic and a population parameter. It compares the observed value to what is expected under the null hypothesis, helping to determine if the difference is statistically significant.

Question 2: 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 observing data at least as extreme as the current data, assuming that the null hypothesis is true. It provides evidence against the null hypothesis.

In **hypothesis testing**, the p-value is compared to a significance level (α\alphaα, commonly 0.05). If the p-value is less than α\alphaα, it suggests that the observed result is unlikely under the null hypothesis, and we reject the null hypothesis.

A **very small p-value** (e.g., 0.01) means that there is strong evidence against the null hypothesis, indicating that the result is highly unlikely to occur by chance, and we may reject the null hypothesis with high confidence.

Question 3: Compare and contrast the binomial and Bernoulli distributions.

- **Bernoulli distribution** is a special case of the binomial distribution where there is only **one trial**. It models a single experiment with two possible outcomes: success (with probability ppp) or failure (with probability 1−p1 - p1−p).

- **Binomial distribution** models the number of successes in **n independent Bernoulli trials**. It requires two parameters: the number of trials nnn and the probability of success ppp. The probability mass function is given by:

P(X=k)=(nk)pk(1−p)n−kP(X = k) = \binom{n}{k} p^k (1 - p)^{n - k}P(X=k)=(kn)pk(1−p)n−k

Where kkk is the number of successes.

Key differences:

- **Bernoulli**: Single trial, outcomes are success or failure.
- **Binomial**: Multiple trials, counting the number of successes in nnn independent Bernoulli trials.

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

The **binomial distribution** is used when:

1. There are a fixed number of trials (nnn).
2. Each trial has two possible outcomes: success or failure.
3. The probability of success ppp is constant across trials.
4. The trials are independent of each other.

The binomial distribution is essentially the **sum of multiple Bernoulli trials**. If you perform nnn independent Bernoulli experiments with the same probability of success ppp, the distribution of the number of successes follows a binomial distribution.

Question 5: What are the key properties of the Poisson distribution, and when is it appropriate to use this distribution?

The **Poisson distribution** models the number of events occurring within a fixed interval of time or space when these events happen independently, and the rate at which they occur is constant. The key properties are:

- Defined by the **rate parameter** λ\lambdaλ, which represents the average number of events in the interval.
- The mean and variance of the distribution are both equal to λ\lambdaλ.
- The probability mass function is:

P(X=k)=λke−λk!P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}P(X=k)=k!λke−λ

Where kkk is the number of occurrences.

It is appropriate to use the Poisson distribution when:

1. Events occur independently.
2. Events happen at a constant average rate.
3. The probability of more than one event occurring in a very small interval is negligible.

Examples include modeling the number of phone calls received by a call center in an hour or the number of emails received in a day.

Question 6: Define the terms "probability distribution" and "probability density function" (PDF). How does a PDF differ from a probability mass function (PMF)?

- A **probability distribution** describes how the values of a random variable are distributed. It assigns probabilities to different outcomes or ranges of outcomes.
- A **probability density function (PDF)** is used for **continuous random variables** and gives the relative likelihood of a value falling within a range. The area under the PDF curve for a range represents the probability of the random variable falling in that range.

P(a≤X≤b)=∫abf(x)dxP(a \leq X \leq b) = \int_a^b f(x) dxP(a≤X≤b)=∫abf(x)dx

- A **probability mass function (PMF)** is used for **discrete random variables**. It gives the probability that a random variable takes on a specific value.

The key difference: **PDF** applies to continuous variables and deals with ranges (areas under the curve), while **PMF** applies to discrete variables and assigns specific probabilities to individual outcomes.

Question 7: Explain the Central Limit Theorem (CLT) with an example.

The **Central Limit Theorem (CLT)** states that the distribution of the sample mean (or sum) of a large number of independent, identically distributed (i.i.d.) random variables approaches a **normal distribution**, regardless of the original distribution of the variables, as the sample size increases.

Example: Suppose you are rolling a fair die repeatedly. The outcomes (1 to 6) are not normally distributed. However, if you take the average of the outcomes from many rolls (say 50 or 100 rolls), the distribution of these averages will approximate a normal distribution, even though the die outcomes are not normal.

The CLT is critical in inferential statistics because it allows us to use normal distribution-based methods for hypothesis testing and confidence intervals, even when the population distribution is not normal.

Question 8: Compare z-scores and t-scores. When should you use a z-score, and when should a t-score be applied instead?

- A **z-score** measures how many standard deviations a data point is from the population mean, and it is used when the population standard deviation σ\sigmaσ is known, or the sample size is large (typically n>30n > 30n>30).
- A **t-score** is used when the population standard deviation is **unknown**, and the sample size is **small** (n≤30n \leq 30n≤30). It accounts for the additional variability introduced by estimating the population standard deviation from the sample.

Use a z-score:

- When the population standard deviation is known.
- When the sample size is large.

Use a t-score:

- When the population standard deviation is unknown.
- When the sample size is small.

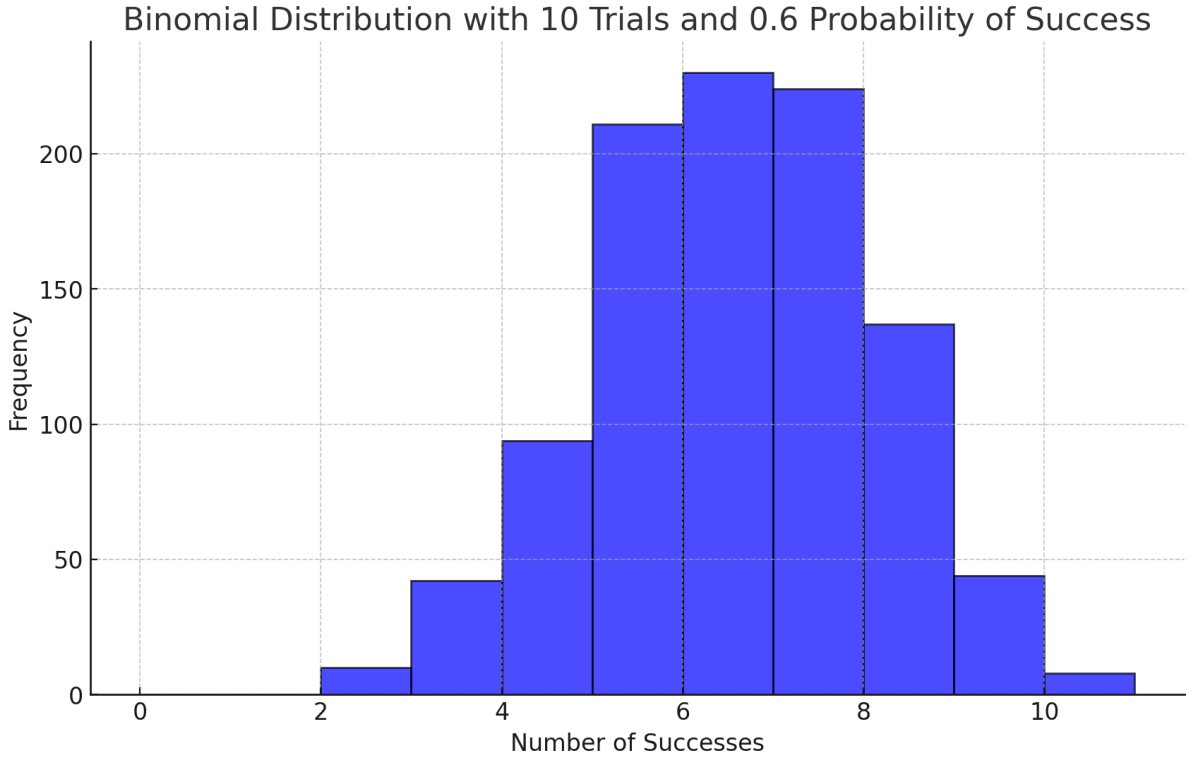
Question9: Given a sample mean of 105, a population mean of 100, a standard deviation of 15, and a sample size of 25, calculate the z-score and p-value. Based on a significance level of 0.05, do you reject or fail to reject the null hypothesis? Task: Write Python code to calculate the z-score and p-value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

ANSWER

the z-score is approximately **1.67**, and the p-value is **0.096**. Since the p-value (0.096) is greater than the significance level (α=0.05\alpha = 0.05α=0.05), we **fail to reject** the null hypothesis. This means that there is not enough evidence to conclude that the sample mean is significantly different from the population mean.

Question 10 : 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.

Now, I'll simulate a binomial distribution with 10 trials and a probability of success of 0.6 for Question 10 and calculate the expected mean and variance.



after simulating a binomial distribution with 10 trials and a probability of success of 0.6, the following values were obtained:

- Mean: 6.10 (expected mean: $n \times p = 10 \times 0.6 = 6$n \times p = 10 \times 0.6 = 6$n×p=10×0.6=6$)
- Variance: 2.41 (expected variance: $n \times p \times (1-p) = 10 \times 0.6 \times 0.4 = 2.4$n \times p \times (1 - p) = 10 \times 0.6 \times 0.4 = 2.4$n×p×(1−p)=10×0.6×0.4=2.4$)

The simulated values are very close to the theoretical values for the binomial distribution.