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

Answer:

The **z-statistic** is a measure of how many standard deviations an element is from the mean. It is calculated using:

\[
z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}
\]

Where:
- \(\bar{x}\) = sample mean
- \(\mu\) = population mean
- \(\sigma\) = population standard deviation
- \(n\) = sample size

It follows the **standard normal distribution** (mean = 0, SD = 1) when the population is normally distributed or the sample size is large (n ≥ 30).

In hypothesis testing, the z-statistic helps determine whether to reject the null hypothesis by comparing it to critical values or by finding the corresponding p-value.

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)?

Answer:

A **p-value** is the probability of obtaining test results at least as extreme as the observed data, assuming the null hypothesis is true.

- **Small p-value (e.g., 0.01)** indicates strong evidence against the null hypothesis.
- If **p-value < significance level** (commonly 0.05), we reject the null hypothesis.
- If **p-value ≥ significance level**, we fail to reject the null hypothesis.

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

Answer:

- **Bernoulli Distribution**:
  - Single trial
  - Outcomes: success (1) or failure (0)
  - Mean = \(p\), Variance = \(p(1-p)\)

- **Binomial Distribution**:
  - Multiple independent Bernoulli trials (n trials)
  - Counts number of successes
  - Mean = \(np\), Variance = \(np(1-p)\)

Key Difference: Bernoulli 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:

Binomial distribution is used when:
- There are a fixed number of independent trials (n)
- Each trial has only two outcomes: success or failure
- The probability of success (p) is constant

Relation to Bernoulli: Each trial in a binomial distribution is a Bernoulli trial. Thus, binomial = sum of n Bernoulli trials.

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

Answer:

Poisson distribution models the number of times an event occurs in a fixed interval of time/space.

Properties:
- Events occur independently
- The mean equals the variance (λ)
- Probability mass function:

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

Use case: When events are rare or count-based (e.g., number of calls to a call center per hour).

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

Answer:

- **Probability Distribution**: Describes how probabilities are distributed over values of a random variable.
- **PDF (Probability Density Function)**: Used for continuous variables; the probability that a variable falls within a range is the area under the curve.
- **PMF (Probability Mass Function)**: Used for discrete variables; gives the exact probability of a specific value.

Difference: PDF gives probability over intervals (area), PMF gives probability for specific values.

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

Answer:

The **Central Limit Theorem** states that the sampling distribution of the sample mean approaches a normal distribution as the sample size becomes large, regardless of the population's distribution.

Example: If we take many samples of size 30 or more from a population of exam scores and calculate their means, the distribution of those means will approximate a normal distribution.

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

Answer:

- **Z-score**: Used when population standard deviation is known and sample size is large.
- **T-score**: Used when population standard deviation is unknown and sample size is small (n < 30).

Key difference: T-distribution is wider (accounts for extra uncertainty) and used for smaller samples.

Question 9: 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?

Python Code:
```python
import scipy.stats as stats
import math

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

# Z-score calculation
z = (sample_mean - population_mean) / (std_dev / math.sqrt(sample_size))

# P-value for two-tailed test
p_value = 2 * (1 - stats.norm.cdf(abs(z)))

print(f"Z-score: {z:.2f}")
print(f"P-value: {p_value:.4f}")
```

Output:
```
Z-score: 1.67
P-value: 0.0946
```

Conclusion: Since p-value (0.0946) > 0.05, we **fail to reject** the null hypothesis.

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?

Python Code:
```python
```

```
import numpy as np
import matplotlib.pyplot as plt

# Parameters
n = 10
p = 0.6
samples = 1000

# Generate data
data = np.random.binomial(n, p, samples)

# Plot
plt.hist(data, bins=range(n+2), align='left', rwidth=0.8, color='skyblue', edgecolor='black')
plt.title('Binomial Distribution Simulation')
plt.xlabel('Number of Successes')
plt.ylabel('Frequency')
plt.grid(True)
plt.show()

# Expected mean and variance
mean = np.mean(data)
variance = np.var(data)

print(f"Expected Mean: {mean:.2f}")
print(f"Expected Variance: {variance:.2f}")
```

Output Example:
```
Expected Mean: ~6.00
Expected Variance: ~2.40
```

Theoretical Mean = $np = 10 \times 0.6 = 6$
Theoretical Variance = $np(1-p) = 10 \times 0.6 \times 0.4 = 2.4$