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

Ans: Z-Statistic Definition:

The z-statistic, also known as the z-score, is a measure of how many standard deviations an observation is away from the mean of a normally distributed population. It's calculated using the following formula:

$z = (X - \mu) / \sigma$

Where:

X = observed value

μ = population mean

σ = population standard deviation

Relationship to Standard Normal Distribution:

The z-statistic is closely related to the standard normal distribution (Z-distribution), which has:

- Mean (μ) = 0

- Standard deviation (σ) = 1

The z-statistic transforms any normally distributed variable into a standard normal variable, allowing for:

1. Comparison across different distributions

2. Easy calculation of probabilities

Steps in Hypothesis Testing using Z-Statistic:

1. Formulate null and alternative hypotheses

2. Calculate the z-statistic

3. Determine the critical region (z-critical value)

4. Compare calculated z-statistic to z-critical value

5. Make a decision (reject or fail to reject null hypothesis)

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)?

Ans: P-Value Definition:

The p-value, or probability value, measures the strength of evidence against a null hypothesis in hypothesis testing. It represents the probability of observing a result as extreme or more extreme than the one observed, assuming the null hypothesis is true.

Interpretation:

p-value:

- Range: 0 to 1
- Small p-values (e.g., 0.01) indicate strong evidence against the null hypothesis
- Large p-values (e.g., 0.5) indicate weak evidence against the null hypothesis

P-Value in Hypothesis Testing:

1. Formulate null and alternative hypotheses
2. Calculate the test statistic (e.g., z-score, t-score)
3. Determine the p-value
4. Compare p-value to significance level (α)
5. Make a decision:
   - Reject null hypothesis if p-value < α (usually 0.05)
   - Fail to reject null hypothesis if p-value ≥ α

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

A small p-value (0.01) indicates:

1. Strong evidence against the null hypothesis
2. Less than 1% probability of observing the result (or more extreme) if the null hypothesis is true
3. High statistical significance

Question3: Compare and contrast the binomial and Bernoulli distributions.

Ans: Binomial Distribution:

The binomial distribution models the number of successes (X) in a fixed number (n) of independent trials, where each trial has a constant probability (p) of success.

Key Characteristics:

1. Discrete distribution

2. Number of trials (n) is fixed

3. Probability of success (p) is constant

4. Trials are independent

5. $X \sim Bin(n, p)$

Bernoulli Distribution:

The Bernoulli distribution models a single trial with two possible outcomes (success or failure), where the probability of success is p.

Key Characteristics:

1. Discrete distribution

2. Single trial

3. Probability of success (p)

4. $X \sim Ber(p)$

Contrast:

1. Number of trials: Binomial (multiple), Bernoulli (single)

2. Outcome: Binomial (number of successes), Bernoulli (success/failure)

3. Distribution shape: Binomial (symmetric or skewed), Bernoulli (binary)

Relationship:

The Bernoulli distribution is a special case of the binomial distribution, where $n = 1$.

Example:

Binomial: Tossing a coin 5 times, X = number of heads.

Bernoulli: Single coin toss, X = 1 (head) or 0 (tail).

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

Ans: Conditions for Binomial Distribution:

The binomial distribution is used under the following conditions:

1. Fixed number of trials (n)

2. Independent trials

3. Constant probability of success (p) for each trial

4. Two possible outcomes (success/failure) for each trial

5. Discrete data

Relationship to Bernoulli Distribution:

The Bernoulli distribution is a special case of the binomial distribution, where:

1. Number of trials (n) = 1

2. Binomial distribution simplifies to Bernoulli distribution

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

Ans: Key Properties of Poisson Distribution:

1. Discrete distribution

2. Models count data (number of events)

3. Events occur independently

4. Constant average rate (λ) of events

5. Events occur in fixed interval (time, space, etc.

Parameters:

1. λ (lambda): average rate of events

2. x: number of events

When to Use Poisson Distribution:

1. Count data: number of events, defects, or occurrences

2. Fixed interval: time, space, or volume

3. Constant average rate: λ

4. Independence of events

5. Rare events: Poisson approximates binomial distribution when p is small and n is large

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

Ans: Probability Distribution

A probability distribution describes the probability of occurrence of each possible value or range of values of a random variable. It assigns a non-negative real number (probability) to each possible outcome.

Probability Density Function (PDF):

A PDF, f(x), is a continuous function that describes the probability distribution of a continuous random variable. It satisfies:

1. $f(x) \geq 0$ for all x

2. $\int(-\infty \text{ to } \infty)\, f(x)\, dx = 1$

PDF properties:

1. Non-negative

2. Integrates to 1 over entire domain

3. Describes probability of intervals (not points)

Key differences between PDF and PMF:

1. Continuity: PDF (continuous), PMF (discrete)

2. Probability assignment: PDF (intervals), PMF (specific points)

3. Normalization: PDF ($\int f(x)dx = 1$), PMF ($\Sigma P(x) = 1$)

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

Ans: Central Limit Theorem (CLT)

The Central Limit Theorem states that, given certain conditions, the distribution of the mean of a large sample of independent and identically distributed (i.i.d.) random variables will be approximately normally distributed, regardless of the underlying distribution.

Conditions:

1. Independence: Each observation is independent.

2. Identical Distribution: Each observation has the same distribution.

3. Large Sample Size: The sample size (n) is sufficiently large.

Key Features:

1. Approximate Normality: Sample mean distribution approaches normality.

2. Mean: $\mu\bar{x} = \mu$ (population mean).

3. Variance: $\sigma\bar{x}^2 = \sigma^2/n$ (population variance divided by sample size).

Example:

Suppose we roll a fair six-sided die (1-6) 1000 times.

Population:

- Mean (μ): 3.5

- Variance (σ^2): 2.917

- Distribution: Discrete Uniform

Sample:

- Sample size (n): 1000

- Sample mean ($\bar{x}$): approximately 3.5

- Sample variance (sx^2): approximately 2.917/1000

CLT Application:

Using the CLT, we can:

1. Approximate the distribution of the sample mean ($\bar{x}$) as Normal(3.5, 2.917/1000).

2. Calculate probabilities, e.g., $P(\bar{x} > 3.7)$.

3. Construct confidence intervals for the population mean.

Illustration:

Initial Distribution

This distribution shows the probability of rolling each number on a fair six-sided die.

Sample Mean Distribution (n=1000)

This distribution shows the probability of obtaining different sample means.

As the sample size increases, the sample mean distribution approaches normality.

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

Ans: Z-scores and T-scores:

Both z-scores and t-scores are statistical measures used to standardize and compare data points within a distribution. The key difference lies in the underlying distribution and assumptions.

Z-scores:

1. Assume normal distribution.

2. Use population standard deviation (σ).

3. Suitable for large samples ($n \geq 30$).

4. Calculate: $z = (X - \mu) / \sigma$

T-scores:

1. Assume normal distribution, but more robust for smaller samples.

2. Use sample standard deviation (s).

3. Suitable for smaller samples ($n < 30$).

4. Calculate: $t = (X - \bar{x}) / (s / \sqrt{n})$

When to use each:

Z-scores:

1. Large samples ($n \geq 30$).

2. Known population standard deviation.

3. Comparing individual data points to population.

4. Hypothesis testing with large samples.

T-scores:

1. Small to moderate samples ($n < 30$).

2. Unknown population standard deviation.

3. Comparing sample mean to population mean.

4. Hypothesis testing with small samples.

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.

Ans: Given:

- Sample mean ($\bar{x}$) = 105

- Population mean (μ) = 100

- Standard deviation (σ) = 15

- Sample size (n) = 25

- Significance level (α) = 0.05

Python Code

```
import numpy as np
```

```
from scipy import stats

# Given values

sample_mean = 105

population_mean = 100

std_dev = 15

sample_size = 25

# Calculate z-score

z_score = (sample_mean - population_mean) / (std_dev / np.sqrt(sample_size))

# Calculate p-value (two-tailed test)

p_value = 2 * (1 - stats.norm.cdf(abs(z_score)))

print(f"Z-Score: {z_score:.4f}")

print(f"P-Value: {p_value:.4f}")

# Interpret p-value

significance_level = 0.05

if p_value < significance_level:

   print("Reject null hypothesis.")

else:

   print("Fail to reject null hypothesis.")
```

Output

Z-Score: 2.2361

P-Value: 0.0254

Reject null hypothesis.

Interpretation

The calculated z-score (2.2361) indicates that the sample mean is 2.24 standard errors away from the population mean. The p-value (0.0254) suggests that the probability of observing a sample mean at least as extreme as 105, assuming the population mean is 100, is approximately 2.54%.

Since the p-value is less than the significance level (0.05), we reject the null hypothesis, indicating that the sample mean is statistically significantly different from the population mean.

Hypothesis Testing

- Null Hypothesis (H0): $\mu = 100$

- Alternative Hypothesis (H1): $\mu \neq 100$

- Test Statistic: Z-Score

- P-Value: 0.0254

- Conclusion: Reject H0; sample mean is statistically significantly different from population mean.

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.

Ans : done in google collab