Statistics Advance-1

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

=> z statistic and standard Normal distribution

The z-statistic, or z-score, are measures how many standard deviations a data point (sample mean) is from the population mean. it is a way of standardizing data allowing for comparison between different distributions by converting them to a common scale.

The formula for z-statistics is:

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

→ x is the observed value or sample mean
→ μ is the population mean
→ σ is the population standard deviation.

If you are dealing with a sample instead of a population, and the population's standard deviation is unknown, you would use the sample standard deviation and adjust the formula accordingly.

When working with the mean of a sample ($\bar{x}$) from a population, the formula is:

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

→ $\bar{x}$ is the sample mean.
→ μ is the population mean.
→ σ is the population standard deviation,
→ n is the sample size.

& Relationship to the standard Normal distribution.
The z-statistic follows a standard normal distribution, also known as the z-distribution, which has a mean of 0 and a standard deviation of 1. When you calculate a z-score you're converting raw data into a value on this standard normal distribution, which allows you to assess probabilities and make comparisons.
→ A z-score of 0 means the value is exactly at the mean.
→ positive z-scores indicate values above the mean.
→ negative z-scores indicate values below the mean.

The standard normal distribution is essential in statistics because it provides a reference for understanding how extreme or typical a data point is within a distribution.

☆ Z statistic in Hypothesis testing
In hypothesis testing, the z statistic is used when testing the means of large

samples ($n > 30$) or when the population standard deviation is known.

1. Null Hypothesis (H_0): A statement that there is no effect or no difference (e.g. $\mu = \mu_0$).

2. Alternative Hypothesis (H_1 or H_A): A statement that there is an effect or a difference. (e.g. $\mu \neq \mu_0$).

3. Calculate the Z-statistic: using the formula above to determine how far the sample mean is from the population mean under the null hypothesis.

4. Compare with critical values: the Z-statistic is compared against critical values from the standard normal distribution. Common critical values for a two-tailed test are ± 1.96 for a 95% confidence level (5% significance level).

→ If $|Z| > 1.96$ you reject the null hypothesis.

→ If $|Z| \leq 1.96$, you fail to reject the null hypothesis.

5. P-value: Alternatively, you can use the Z-statistics to calculate the P-value, which tells you the probability of obtaining the test statistic as extreme as the one observed, assuming the null hypothesis is true. If the P-value is less than the significance level (commonly 0.05), you reject the null hypothesis.

Example

Suppose you want to test whether the mean height of a population is 170 cm. You collect a sample of 100 people and calculate a sample mean of 172 cm, with a population standard deviation of 10 cm.

Using the Z-statistic formula:

$$Z = \frac{172 - 170}{10/\sqrt{100}} = \frac{2}{1} = 2$$

with a z-value of 2 and using a significance level of 0.05 (critical Z-value of 1.96 for a two-tailed test), you would reject the null hypothesis because $|Z| = 2 > 1.96$, suggesting that the mean height is significantly different from 170 cm.

Q. 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 obtaining test results at least as extreme as the observed results under the assumption that the null hypothesis is true. In other words, it quantifies how likely it is to observe the data purely by

random chance if the null hypothesis is to hold.

The p-value helps us assess whether the observed data provides enough evidence to reject the null hypothesis in favor of the alternative hypothesis.

P-value in Hypothesis Testing

In hypothesis testing, the p-value is used to determine whether the observed data is consistent with the null hypothesis.

1. Null Hypothesis (H_0): Assumes that there is no effect or no difference. (e.g., the population mean is equal to some hypothesized value).
2. Alternative Hypothesis (H_1): Assumes that there is an effect or difference. (e.g. the population mean is different from the hypothesized value).
3. Calculate the Test Statistic: A test statistic, or t-statistic, is calculated based on the sample data.
4. Compute the p-value: the p value is derived from the test statistic and shows the likelihood of observing the sample data (or more extreme data) if the null hypothesis is true.
5. Compare with significance level (α): The p-value is compared with a predetermined significance level (commonly $\alpha = 0.05$).

→ if p-value < α: reject the null hypothesis. This suggests that the observed data is unlikely under the null hypothesis, providing evidence in favor of the alternative hypothesis.

→ if p-value > α: fail to reject the null hypothesis. This means the data does not provide strong enough evidence to support the alternative hypothesis.

* Looking at a small p-value (eg. 0.01)
if the p-value is very small, such as 0.01, it indicates that the observed data is highly unlikely under the null hypothesis. Specifically, if the null hypothesis were true, there would be only a 1% chance of obtaining the observed results by random chance. In such cases, we reject the null hypothesis and conclude that there is strong evidence for the alternative hypothesis.

Interpretation of a small p-value

⇔ p-value ≥ 0.01: There is only e.g. probability that the observed data could occur under the null hypothesis.

→ Practical interpretation: A small p-value suggests that the observed effect

or difference is statistically significant. The smaller the p-value, the stronger the evidence against the null hypothesis.

Example
Suppose you are testing whether a new drug is more effective than the current standard drug. After conducting an experiment, you compute a p-value of 0.01. This means there is only a 1% chance that the observed improvement in patients outcomes could happen by random chance if the new drug is no better than the standard drug.

→ Since the p-value is smaller than the commonly used significance level of 0.05, you would reject the null hypothesis and conclude that the new drug is significantly more effective.

Q.3. Compare and contrast the binomial and Bernoulli distributions.

=> The binomial and Bernoulli distributions are closely related but they are used in different contexts and have some key differences.

1. Definition

☆ Bernoulli distribution

→ A Bernoulli distribution models a single trial of a random experiment

(with two possible outcomes: "success" (with Probability: p) and "failure" (with Probability 1-p).
→ it is the simplest discrete Probability distribution.
→ Examples: Flipping a coin once (success = heads, failure = tails).

☆ Binomial distribution
→ A binomial distribution models the number of successes in a fixed number of independent Bernoulli trials.
→ it Generalizes the Bernoulli distribution by considering multiple trials.
→ Example: Flipping a coin 10 times and counting the number of heads.

2. Number of Trials
→ Bernoulli: A single trial (i.e one flip of a coin).
→ Binomial: involves a fixed number of trials (i.e. flipping the coin n times.)

3. Parameters
Bernoulli: Has only one parameter:
→ p: The Probability of success.
→ The Probability of failure is 1-p.
Binomial: Has two parameters.
→ n: the number of independent Bernoulli trials.
→ p: the probability of success in each trial.

4. Probability mass function (PMF)

Bernoulli PMF:

$$P(X=x) = p^x(1-p)^{1-x}, \quad x \in \{0,1\}$$

where

→ X is Bernoulli random variable.

→ x=1 indicates success (with probability p),

→ x=0 indicates failure (with probability 1-p)

Binomial PMF:

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in \{0,1,2,\ldots,n\}$$

where

→ X is a binomial random variable representing the number of successes,

→ k is the number of successes in n trials.

→ $\binom{n}{k}$ is the binomial coefficient which counts how many ways you can choose k successes from n trials.

5. Mean and variance

Bernoulli distribution

→ mean $\mu = p$

→ variance $\sigma^2 = p(1-p)$

Binomial distribution

→ Mean $\mu = np$

→ variance $\sigma^2 = np(1-p)$

6. Relationship between the two

→ The Bernoulli distribution is a special case of the binomial distribution where n=1, in other words a binomial distribution with n=1 trials is equally a Bernoulli distribution.

→ Bernoulli: 1 trials with Probability p.
Binomial:- n independent Bernoulli trials, each with probabilities p.

7. Example
Bernoulli Example: If a coin has a 50% chance of one landing heads (P=0.5), a Bernoulli trial is just one flip of the coin. The outcome will either heads (success, X=1) or tails (failure, X=0).
Binomial Example: If you flip the same coin 10 times, the number of heads (successes) will follow a binomial distribution with n=10 and P=0.5. The random variable X can have values from 0 (no heads) to 10 (all heads)

8. Use cases
Bernoulli distribution: Used when modeling the outcome of a single binary trial, such as:
→ Flipping a coin once
→ Whether a student passes or fails an exam.
→ Whether a customer makes a purchase (yes/no).
Binomial Distribution: Used when modeling the number of successes in a fixed number of independent binary trials, such as:
→ The number of heads in 10 coin flips.

→ The number of of students passing can exam out of 50 students.

→ The number of customers making ce purchase out of 100 who visited store.

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

⇒ The binomial distribution is used under specific conditions when you are dealing with a series of independent trails or experiments, each of which has two possible outcomes: success or failure. it arises in situations where you are interested in the number of success in a fixed number of trials.

Conditions for using the binomial distribution:

1. fixed number of trials (n): The experiment or process must consist of a fixed number of trials.
2. Binary outcomes: Each trail must have only two possible outcomes, usually termed "success" or "failure".
3. Constant Probability of success (P): The probability of success on each trial must be the same for every trial.
4. independence: Each trial must be independent meaning the outcome of one trial does not affect the outcome of any other trial.

Formula for the Binomial distribution

The probability of getting exactly k successes in n independent trails is given

by the binomial Probability mass function (PMF):

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

where

→ n is the number of trials,

→ k is the number of successes,

→ p is the probability of success in a single trial and

→ $\binom{n}{k}$ is the binomial coefficient, which represents the number of ways to choose k successes from n trials.

✱ Relationship to the Bernoulli distribution:

⇒ The Bernoulli distribution is a special case of the binomial distribution. It represents the distribution of a single trial (i.e. $n=1$) with two outcomes: success (with probability p) or failure (with probability $1-p$).

→ Bernoulli distribution: Describe a single trial with two possible outcomes.

→ Binomial distribution: Describe the number of successes in a fixed number of independent Bernoulli trials.

In other words, the binomial distribution can be thought of as the sum of several independent Bernoulli trials.

Example

Bernoulli trial: Flip a coin once (success = heads, failure = tails).

Binomial Experiment: flip the coin 10 times and count how many heads appear.

→ In this case, each coin flip is a Bernoulli trial, and the binomial distribution would describe the probability of getting a specific number of heads (successes) in the 10 flips (trials).

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

⇒ The Poisson distribution is a probability used to model the number of events that occur in a fixed interval of time, space, or another domain, under specific conditions. 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 events (i.e., events that can only take on non-negative integer values: 0, 1, 2, ...).
2. Rate of occurrence (λ): The parameter λ represents both the average number of events in the given interval and the rate at which the events occur. It must be positive.

3. Independence: The occurrence of one event does not affect the occurrence of another event. Each event happens independently.
4. Events occur Randomly: Events are distributed randomly over time (Rule 1).
5. Non-overlapping intervals: The number of events occurring in one interval is independent of the number of events occurring in any other non-overlapping interval.
6. No upper limit: Theoretically, the number of events can range from zero to infinity, though the probability of very high counts decreases as the count increases.
7. Mean and Variance: The mean (Expected Value) and variance of a Poisson distribution are both equal to λ.
- Mean: $E(X) = \lambda$
- Variance: $Var(X) = \lambda$

Poisson distribution formula

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

where
→ x is the random variable representing the number of events
→ k is the number of events (calls, ...)
→ λ is the average rate (mean number of events) in the given interval,

→ e is the base of the natural logarithm (approximately 2.7182)

☆ When to use the Poisson distribution

1. Modeling the number of events in a fixed interval of time, space, or volume e.g., number of phone call received in a hour, or number of customer arrival at a store, or number of earthquakes in a year.
2. The events are rare and occur independently relative to the length of the interval.
3. The events occur independently of each other.
4. The average rate of occurrence is constant over time or space. (i.e. λ remains constant).
5. The probability of more than one event occuring in a very short interval is negligible.

Use Cases

Call center, Accidents, Arrivals, Natural Phenomena.

Q.6. Define the terms "Probability distribution" and "Probability density function" (PDF). How does a PDF differ a Probability mass function (PMF)?

⇒ probability distribution

A probability distribution is a mathematical description of the likelihood of different outcomes in an experiment or random

Process. It assigns a probability to each possible outcome of a random variable. Depending on whether the random variable is discrete or continuous.

- For discrete random variables, the distribution is described by a probability mass function.
- For continuous random variables, the distribution is described by a probability density function (PDF).

A probability distribution provides a complete description of a random variable's behaviour including the possible values it can take and the likelihood or density of these values.

★ Probability density function (PDF)

A probability density function (PDF) describes the likelihood of a continuous random variable taking a particular value. Unlike a discrete probability distribution, where probabilities are assigned to specific outcomes, the PDF gives the density of probability at any given value of the random variable.

→ A PDF, $f(x)$, for a continuous random variable X, satisfies the following properties:

1. $f(x) \geq 0$ for all x.
2. The total area under the curve of

the PDF equals 1 (i.e., $\int_{-\infty}^{\infty} f(x)\,dx = 1$).

3. The probability that the variable X falls within a particular range [a, b] is given by the area under the curve between a and b:

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

In Continuous distribution, the probability of the random variable taking on an exact value (e.g. $P(X=x)$) is zero, but probabilities for intervals (e.g., $P(a \leq X \leq b)$) can be calculated.

★ Probability mass function (PMF)

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

→ A PMF, $P(X=x)$, satisfies:

1. $P(X=x) > 0$ for all x.
2. The sum of probabilities across all possible values of X equals 1:

$$\sum_x P(X=x) = 1$$

In a discrete distribution, the PMF assigns a specific probability to each possible outcome.

Page No.
Date

Differences between PDF and PMF

☆ Type of random variable:

→ PMF is used for discrete random variables where outcomes take distinct, countable values (e.g. 0, 1, 2, ...).

→ PDF is used for continuous random variables, where outcomes can take any value within an interval (e.g. any real number).

☆ Interpretation of values:

→ For a PMF, the function directly gives the probability of the random variable taking on a specific value (e.g. P(2) = 0.3).

→ For a PDF, the function gives the density of the random variable at a point, not the actual probability. Probabilities are obtained by integrating the PDF over an interval. (e.g., P(a < X ≤ b)).

☆ Summation vs. Integration

→ For PMFs, probabilities are summed over possible values.

→ For PDFs, probabilities are calculated by integrating over an interval of values.

☆ Example:

PMF (Discrete case): Consider a fair six-sided die. The PMF is:

$$P(X = k) = \frac{1}{6} \quad \text{for } k = 1, 2, 3, 4, 5, 6$$

Each outcome has an equal probability of 1/6.

Page No.
Date

2. PDF (Continuous case): for a random variable that is normally distributed with mean 0 and variance 1, the PDF is:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

Q.7. Explain the central limit theorem (CLT) with example.

⇒ The central limit theorem (CLT) is one of the most important results in statistics. It states that, regardless of the shape of the original population distribution, the distribution of the sample mean (or the sum) of a sufficiently large number of independent and identically distributed random variables will tend to follow a normal distribution (a bell curve), provided the sample size is large enough.

☆ Key Points of the Central Limit Theorem

1. Independence: The random variable being averaged or summed must be independent.
2. Identically distributed: The random variables should come from the same probability distribution.
3. Sample size: The sample size (n) must be sufficiently large ($n \geq 30$) is considered large enough).
4. Mean and Variance: The mean of the sample mean distribution is equal to the mean of the original distribution, and the variance of the sample mean distribution is equal to the

Page No.
Date

(variance of the original distribution divided by n (the sample size).

* Mathematical statement
Let $X_1, X_2, \ldots, X_n$ be i.i.d random variables from a population with mean μ and variance σ^2. The CLT states that the distribution of the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ approaches a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$ as n becomes large.
In other words:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ as } n \to \infty$$

* Example of central limit theorem
Suppose you are rolling a die. The outcomes of each roll follows a uniform distribution (i.e. each face has an equal probability of $\frac{1}{6}$). The possible outcomes are 1, 2, 3, 4, 5, 6. The mean μ of this distribution is 3.5 and the variance σ^2 can be calculated for this uniform distribution.
Now, let's apply the CLT.

1. Single Roll: The distribution of outcomes of a single roll is not normal - it's uniform.
2. Average of multiple rolls: If you roll the die 30 times and compute the average outcome of those 30 rolls,

the distribution of the average will be approximately normal even though the individual roll outcomes are not normally distributed.

→ As the number of rolls increases, the distribution of the sample mean (average outcome) becomes closer and closer to a normal distribution with mean $\mu = 3.5$ and standard deviation $\frac{\sigma}{\sqrt{n}}$, where σ is the standard deviation of the uniform distribution.

* Visualization of the central limit theorem

1. If you roll a die 1 time, the distribution is flat (uniform).
2. If you roll a die 10 times and plot the average outcome from many experiments of 10 rolls, the distribution starts resembling a bell curve.
3. If you roll a die 50 times and plot the average outcome from many experiments of 50 rolls, the distribution of these averages is even closer to a normal distribution.

* importance of the Central Limit Theorem

→ Real World Application: The CLT allows us to use normal distribution techniques even when the underlying data is not normally distributed, as long as we are working with sample means or sums.

→ Statistical inference: The CLT justifies the use of normal distribution-based methods (confidence intervals and hypothesis tests) in many real-world situations, even when the population distribution is unknown or non-normal.

Q. 8. Compare Z-score and t-scores. When should you not use a Z-score, and when should a t-score be applied instead?

⇒ Z-score and t-scores (or t-values) are both used in statistics to measure how far a data point or a sample statistics (like a sample mean) is from the population mean, measured in terms of standard deviations or standard errors. However, they are used in different contexts depending on the available data and certain assumptions about the population.

Z-Score

A Z-score measures how many standard deviations an individual data point or sample statistic is away from the population mean. It assumes that the population standard deviation (σ) is known and that the population is normally distributed.

→ Formula for Z-score:

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

Where

→ x is the value or sample mean.

→ μ is the Population mean.

→ σ is the Population standard deviation.

★ When to use a Z-score

1. Large sample size: When the sample size n is large (typically $n \geq 30$).
2. Known Population Standard deviation (σ): When you know the population standard deviation
3. Normal distribution: Z-scores are often used when the population is normally distributed, but due to the Central limit theorem, they can also be applied when $n \geq 30$, even if the population distribution is not normal.

★ Examples

→ Finding probabilities using the standard normal distribution (Z-table) for standardized test scores.

→ Calculating Confidence intervals or Conducting hypothesis tests when the population standard deviation is known.

★ T-Score:

A t-score (or t-value) is used when the sample size is small and the population standard deviation σ is unknown, instead of using the population standard deviation, we use the sample standard deviation s which adds variability and uncertainty to the estimate. The t-distribution is

Similar to the normal distribution but has heavier tails, meaning there is more probability in the extremes due to the uncertainty in estimating the standard deviation

t-score formula:

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

where

→ X is the sample mean.
→ μ is the population mean.
→ s is the sample standard deviation.
→ n is the sample size

& when to use a t-score

1. Small sample size: when the sample size n is small (typically $n < 30$).
2. Unknown population standard deviation: when the population standard deviation σ is unknown and you are using the sample standard deviation s as an estimate.
3. Nearly normal distribution: The t-distribution assumes the sample comes from a normally or nearly normally distributed population, especially for small sample sizes. As n increases, the t-distribution approaches the normal distribution.

Page No.
Date

* Examples
→ Calculating confidence intervals or conducting hypothesis tests for small sizes where σ is unknown.
→ Students t-tests for comparing means of small samples sizes where σ is unknown.

* Example of when to use each
1. Z-score Example:
→ Suppose you want to calculate a confidence interval for the mean of a large sample of 200 test scores where the population standard deviation σ is known. You would use the Z-score to construct the interval.

2. T-Score example
→ Suppose you have a small sample of 15 students test scores, and you don't know the population standard deviation. You would the use the t-score to calculate the confidence interval accounting for the added uncertainty in estimating the population standard deviation.