STATISTICS ASSIGNMENT

Question 1: What is a random variable in probability theory?

Answer

A random variable is a way of turning the outcomes of a random experiment into numbers. It is really just a function or a map that takes each outcome from a sample space and assigns it a number. We normally use capital letters like X to represent a random variable.

- **A Random Experiment: Rolling of Dice**

  Let us take a basic random experiment: rolling two dice. The outcomes of this experiment include all possible pairs of numbers that can show up on the dice. So, the sample space (which we usually we can call Ω) includes combinations like (1, 1), (2, 3), and (6, 5). These pairs represent every possible result from rolling the two dice.

  Now, consider we only care about the sum of the two numbers on the dice. Here a random variable will be useful. Instead of caring about individual dice rolls, we define a random variable X such that, that takes each pair of dice rolls and returns their sum. For example −

  X(1, 1) = 2

  X(2, 3) = 5

  X(6, 5) = 11

  In this way, we can summarize all the possible outcomes of the dice roll with the sums. It extracts the specific piece of information we care about from the experiment.

- **Formal Definition of a Random Variable**

  Let us define the random variables formally. A random variable is a measurable function from a sample space Ω to the real number line R. It implies the following points −

  Sample Space − This is the set of all possible outcomes of a random experiment. For the dice example, it is the set of all pairs of dice rolls, like (1, 1), (2, 3), ...,(6, 6).

  Real Numbers − This is just the set of numbers we use every day. They could be positive, negative, fractions, decimals, and so on. In most cases, our random variables will map outcomes to real numbers, like the sum of the dice.

  Measurable Function − This means that the random variable behaves nicely with the probabilities we assign to the outcomes. If we think of the sample space as all the

possible outcomes, then the random variable maps those outcomes into a smaller number (like the sum in our dice example).

Types of Random Variables

There are two main types of random variables: the discrete and continuous. Depending on what type of experiment we are conducting, the random variable can either take on a set of distinct values (discrete) or any value within a range (continuous).

Discrete Random Variables

A discrete random variable is one that can take on only specific values. Like the integer or whole numbers. For example, in our dice rolling example, the random variable can only take on the values 2 through 12. These are the possible sums of the two dice.

Discrete random variables are used in situations where we can count the outcomes. The number of heads in a series of coin flips or the number of customers arriving at a store in an hour.

Continuous Random Variables

A continuous random variable can take on any value within a certain range. For example, if we are measuring the time it takes for a car to complete a race. It could take any value like 1.23 seconds, 3.5 seconds, 4.98 seconds, and so on. Continuous random variables are used in situations where the outcome can vary smoothly over a range of values. Random Variables and Probability Random variables are used to get probability in a very direct way. Once we define a random variable, we can use it to calculate probabilities of different events. For example, we might want to find probability that the sum of two dice is greater than 8. Or the probability that a randomly chosen person is taller than 6 feet.

Example: Sum of Two Dice

Let us see the dice example. Now that we have our random variable X to be the sum of the two dice, we can calculate probabilities related to that sum.

Consider we find the probability that the sum of the two dice is greater than 8. The possible outcomes where the sum is 9, 10, 11, or 12. Those outcomes are –

(3, 6), (4, 5), (5, 4), (6, 3) (for 9)

(4, 6), (5, 5), (6, 4) (for 10)

(5, 6), (6, 5) (for 11)

(6, 6) (for 12)

There are 10 outcomes where the sum is greater than 8. There are 36 possible outcomes when rolling two dice, the probability is −

P(X>8)=1036=518

So, the chance that the sum of the dice is greater than 8 is about 27.8%.

Properties of Random Variables

Random variables have several properties as shown below −

Expected Value −

This is the average value we expect the random variable to take on over many trials of the experiment. For a discrete random variable, the expected value is calculated by multiplying each outcome by its probability and summing them up. It is like calculating a weighted average of all the possible outcomes.

Variance −

Variance shows us how much the values of the random variable may vary from the expected value. A low variance means that the outcomes are close to the expected value, while a high variance means that the outcomes are more spread out.

Probability Distribution −

This is a function that gives the probability of each possible value of the random variable. For example, the probability distribution of the sum of two dice tells us how likely each sum (2 through 12) is to occur.

Question 2: What are the types of random variables?

Answer

Random variables can be of two different types:

- Discrete random variables
- Continuous random variables

Discrete random variables:

These types of random variables can take on only a countable number of values be it a small or large number. Some examples of discrete random variables are:

- The number of students in a classroom
- The number of cars sold by a company in a day
- The number of people visiting a website in an hour

The probability distribution of discrete random variables can be calculated using what is called the probability mass function (PMF). PMF is a function that assigns probabilities to discrete outcomes. Probability mass functions can be applied to many discrete random variables at the same time to determine the probability distribution which is called a joint probability distribution. P(X=x, Y=y) denotes the probability that X is equal to x and Y is equal to y simultaneously.

Continuous random variables:

These types of random variables can take on any real value. Some examples of continuous random variables are:

- The time taken to complete a task
- The height of a person
- The weight of a person
- The income of a person

The probability distribution of a continuous random variable can be calculated using what is called the probability density function (PDF). PDF is a function that assigns probabilities to continuous outcomes. The probability of a continuous random variable taking on any particular value is 0 but the probability of it taking values in a range is non-zero. To calculate the probability, we need to integrate the PDF over that range.

Question 3: Explain the difference between discrete and continuous distributions.

Answer

In probability and statistics, discrete and continuous distributions are two fundamental types of probability distributions that describe how values of a random variable are distributed.

Discrete Distribution

A discrete distribution applies to scenarios where the random variable can take on a finite or countable set of distinct values. Each outcome has a specific probability, and the sum of all

probabilities equals 1. The probabilities are represented using a Probability Mass Function (PMF).

Examples:

Rolling a fair die (outcomes: 1, 2, 3, 4, 5, 6).

Flipping a coin (outcomes: heads or tails).

Drawing a card from a deck.

Key Properties:

Finite Outcomes: The set of possible outcomes is limited and countable.

Equal Probability: In uniform discrete distributions, all outcomes have the same probability.

PMF Formula: (P(X = x) = \frac{1}{n}), where (n) is the total number of outcomes

Continuous Distribution

A continuous distribution applies to scenarios where the random variable can take on any value within a continuous range. Probabilities are described using a Probability Density Function (PDF), and the total area under the curve of the PDF equals 1.

Examples:

Random time of day.

Selecting a random point on a line segment.

Measuring the height of individuals.

Key Properties:

Infinite Outcomes: The range of possible values is uncountable and continuous.

Equal Density: In uniform continuous distributions, the density is constant across the interval.

PDF Formula: (f(x) = \frac{1}{b-a}) for (a \leq x \leq b), where (a) and (b) are the lower and upper bounds.

Key Differences

Nature of Outcomes: Discrete distributions deal with finite, countable outcomes, while continuous distributions involve infinite, uncountable ranges.

Probability Representation: Discrete uses PMF, while continuous uses PDF.

Cumulative Distribution Function (CDF): Discrete CDF increases stepwise, while continuous CDF is a smooth, linear function within the interval.

Applications

Discrete distributions are ideal for scenarios like games of chance (e.g., dice rolls), while continuous distributions are suited for measurements like time, length, or weight. Understanding these distinctions is crucial for selecting the appropriate model for real-world problems..

Question 4: What is a binomial distribution, and how is it used in probability?

Answer

The binomial distribution is a discrete probability distribution that models the number of successes in a fixed number of independent trials, where each trial has only two possible outcomes: success or failure. It is widely used in probability theory, statistics, and real-world applications like quality control, survey analysis, and financial modeling.

Key Characteristics

- Fixed Number of Trials (n): The number of trials is predetermined and constant.
- Binary Outcomes: Each trial results in either success or failure.
- Constant Probability (p): The probability of success remains the same across all trials.
- Independence: The outcome of one trial does not affect the others.

  For example, flipping a coin 10 times to count the number of heads is a classic binomial scenario.

Formula

The probability of exactly r successes in n trials is given by:

P(X = r) = nCr * p^r * (1-p)^(n-r)

Where:

- nCr is the number of combinations.
- p is the probability of success.

- 1-p is the probability of failure.

Mean, Variance, and Standard Deviation

- Mean (μ): n * p
- Variance (σ^2): n * p * (1-p)
- Standard Deviation (σ): √(n * p * (1-p))

  For instance, if a coin is flipped 20 times (n=20) with a success probability of 0.5 (p=0.5), the mean number of heads is 10, and the standard deviation is approximately 2.23.

Practical Applications

- Quality Control: Estimating defective items in a batch.
- Survey Sampling: Modeling yes/no responses in surveys.
- Finance: Used in binomial option pricing models.

  Example

  If a fair coin is flipped 4 times, the probability of getting exactly 2 heads can be calculated as:

P(X = 2) = 4C2 * (0.5)^2 * (0.5)^2 = 6 * 0.25 * 0.25 = 0.375

- Important Considerations

  The binomial distribution is best suited for scenarios with fixed trials and constant probabilities. For large n or extreme probabilities, approximations like the normal distribution (via the Central Limit Theorem) or the Poisson distribution may be used

Question 5: What is the standard normal distribution, and why is it important?

Answer

The standard normal distribution is a special case of the normal distribution where:

- The mean (μ) is 0
- The standard deviation (σ) is 1

It is a bell-shaped, symmetric curve centered at zero. The variable used is typically z, representing the z-score, which measures how many standard deviations a value is from the mean.

The standard normal distribution is important because:

Universal Comparisons:

Any normal distribution can be converted to the standard normal using the z-score formula:

$z = x - \mu \sigma$ This allows comparison across different datasets and scales.

Probability Calculations:

It simplifies finding probabilities and percentiles using z-tables or software tools.

Statistical Inference:

Many statistical tests (e.g., z-tests, confidence intervals) rely on the standard normal distribution.

Central Limit Theorem:

It plays a key role in the CLT, which states that the sampling distribution of the sample mean approaches a normal distribution as sample size increases.

Real-World Modeling:

Many natural and social phenomena follow a normal distribution, making the standard normal a useful model.

Application of Standard Normal Distribution

Standard Normal Distribution has a wide range of applications and usage in several fields. Here are some important applications:

- Hypothesis Testing: Performing Z-tests and constructing confidence intervals.
- Probability Calculations: Determining standard normal distribution density probabilities and areas under the curve.
- Data Analysis: Standardizing scores (z-scores) and making suitable analysis.
- Quality Control: Monitoring processes using control charts.
- Risk Management: Calculating financial risks like Value at Risk (VaR).
- Machine Learning: Normalizing features to improve algorithm performance.
- Modeling Phenomena: Analyzing real-world data in natural and social sciences.
- Survey Analysis: Estimating and comparing population proportions.
- Psychometrics: Using standard scores in psychological assessments.

- Operations Research: Optimizing inventory management based on demand modeling.

Characteristics of Standard Normal Distribution

Standard normal distribution is defined by the following characteristics:

Mean:

The mean (average) is 0 that is symbolically represented as μ = 0. Standard Deviation: The standard deviation is 1 that is symbolically represented as σ = 1.

Symmetry:

It is symmetric around the mean (μ = 0). Bell-Shaped Curve: The graph is bell-shaped, that means most values cluster around the mean (μ = 0).

Total Area Under Curve:

The total area under the curve is 1, representing the total probability.

68-95-99.7 Rule:

Approximately 68% of data falls within 1 standard deviation of the mean, 95% within 2 standard deviations, and 99.7% within 3 standard deviations.

Asymptotic:

The tails of the distribution approach, but never touch, the horizontal axis.

Unimodal:

It has a single peak at the mean (μ = 0).

Standard Scores (Z-Scores):

Any normal distribution can be transformed into the standard normal distribution using z-scores where z = (x - μ)/σ.

Question 6: What is the Central Limit Theorem (CLT), and why is it critical in statistics?

The Central Limit Theorem in Statistics states that as the sample size increases and its variance is finite, then the distribution of the sample mean approaches the normal distribution, irrespective of the shape of the population distribution.

Assumptions of the Central Limit Theorem

The Central Limit Theorem is valid for the following conditions:

The drawing of the sample from the population should be random.

The drawing of the sample should be independent of each other.

The sample size should not exceed ten percent of the total population when sampling is done without replacement.

Sample Size should be adequately large.

CLT only holds for a population with finite variance.

Steps to Solve Problems on Central Limit Theorem

Problems of Central Limit Theorem that involves >, < or between can be solved by the following steps:

Step 1: First identify the >, < associated with sample size, population size, mean and variance in the problem. Also there can be 'betwee; associated with range of two numbers.

Step 2: Draw a Graph with Mean as Centre

Step 3: Find the Z-Score using the formula

Step 4: Refer to the Z table to find the value of Z obtained in the previous step.

Step 5: If the problem involves '>' subtract the Z score from 0.5; if the problem involves '<' add 0.5 to the Z score and if the problem involves 'between' then perform only step 3 and 4.

Step 6: The Z score value is found along $X^- X$

Step 7: Convert the decimal value obtained in all three cases to decimal.

Why Is It Critical in Statistics?

The CLT is essential because it allows us to:

- Use Normal Distribution for Inference: Even if the population is not normal, we can use normal-based methods (like z-tests and confidence intervals) for sample means.
- Make Predictions: It enables us to estimate probabilities and make predictions about population parameters using sample data.
- Build Confidence Intervals: CLT justifies the use of confidence intervals for means, which are crucial in research and decision-making.

- Perform Hypothesis Testing: Many statistical tests rely on the assumption of normality, which CLT helps satisfy for large samples.

Real-World Example

- Imagine you're measuring the average height of students in a school:
- The actual height distribution might be skewed.
- But if you take many samples of, say, 30 students each and plot the sample means, the resulting distribution will look normal.
- This lets you apply powerful statistical tools even if the original data isn't normally distributed.

Question 7: What is the significance of confidence intervals in statistical analysis?

Answer

What Is a Confidence Interval?

A confidence interval is a range of values, derived from sample statistics, that is likely to contain the true population parameter. It's usually expressed with a confidence level, such as 95% or 99%.

For example:

A 95% confidence interval for the average height of students might be [160 cm, 170 cm]. This means we are 95% confident that the true average height lies within that range.

Why Are Confidence Intervals Significant?

Confidence intervals are critical because they:

- Quantify Uncertainty: Instead of giving a single estimate (point estimate), CIs show how precise that estimate is.
- Support Decision-Making: They help researchers and policymakers assess risk and make informed choices based on data.
- Enable Comparisons: Overlapping or non-overlapping CIs can indicate whether differences between groups are statistically meaningful.
- Avoid Misleading Certainty: Unlike a point estimate, which might falsely suggest exactness, a CI acknowledges variability and sampling error.

Formula for a Confidence Interval (for a mean)

CI = $\bar{x} \pm z \cdot \sigma / n$ Where:

$\bar{x}$ = sample mean

z = z-score corresponding to the confidence level (e.g., 1.96 for 95%)

σ = population standard deviation

n = sample size

Real-World Example

Suppose a pharmaceutical company tests a new drug and finds the average recovery time is 5.2 days with a 95% CI of [4.8, 5.6]. This interval tells doctors and regulators that the true average recovery time is likely between 4.8 and 5.6 days—not just 5.2.

Would you like to explore how changing the confidence level or sample size affects the width of the interval? It's a fascinating aspect of statistical design.

Question 8: What is the concept of expected value in a probability distribution?

Answer

Properties of Expected Value of a Random Variable The expected value of a random variable possesses several important properties that enhance its usefulness in probability and statistics. Let's explore some key properties:

Linearity of Expectation: This property states that the expected value of the sum or difference of random variables is equal to the sum or difference of their individual expected values. Mathematically, for random variables X and Y and constants a and b , this property can be expressed as: E(aX+bY)=aE(X)+bE(Y) .

Law of Large Numbers:

According to this law, as the sample size increases, the average of a sequence of independent and identically distributed random variables converges to the expected value.

Monotonicity:

If Xand Y are random variables such that X≤Y for all possible outcomes, then the expected value of X is less than or equal to the expected value of Y . In other words, the expected value

preserves the order of random variables.

Non-Negativity:

The expected value of a non-negative random variable is always non-negative. This property is intuitive since the expected value represents an average outcome, and negative values are not possible in this context.

Constant Random Variables:

For a constant random variable C , the expected value is simply equal to that constant. This property is straightforward since a constant value does not vary, and its expected value is the same as the constant itself.

Indicator Random Variables:

Indicator random variables are commonly used to represent events or conditions. If an event occurs with probability p , the expected value of the corresponding indicator random variable is equal to p .

Transformation Invariance:

The expected value is invariant under certain types of transformations. For instance, if g is a function, then the expected value of g(X) is the same as g applied to the expected value of X .

Example

Suppose you play a game where:

You win ₹100 with probability 0.2

You win ₹50 with probability 0.3

You lose ₹20 with probability 0.5

Then the expected value is:

$E(X) = (100 \times 0.2) + (50 \times 0.3) + (-20 \times 0.5) = 20 + 15 - 10 = ₹25$

So, on average, you'd expect to gain ₹25 per game over the long run.

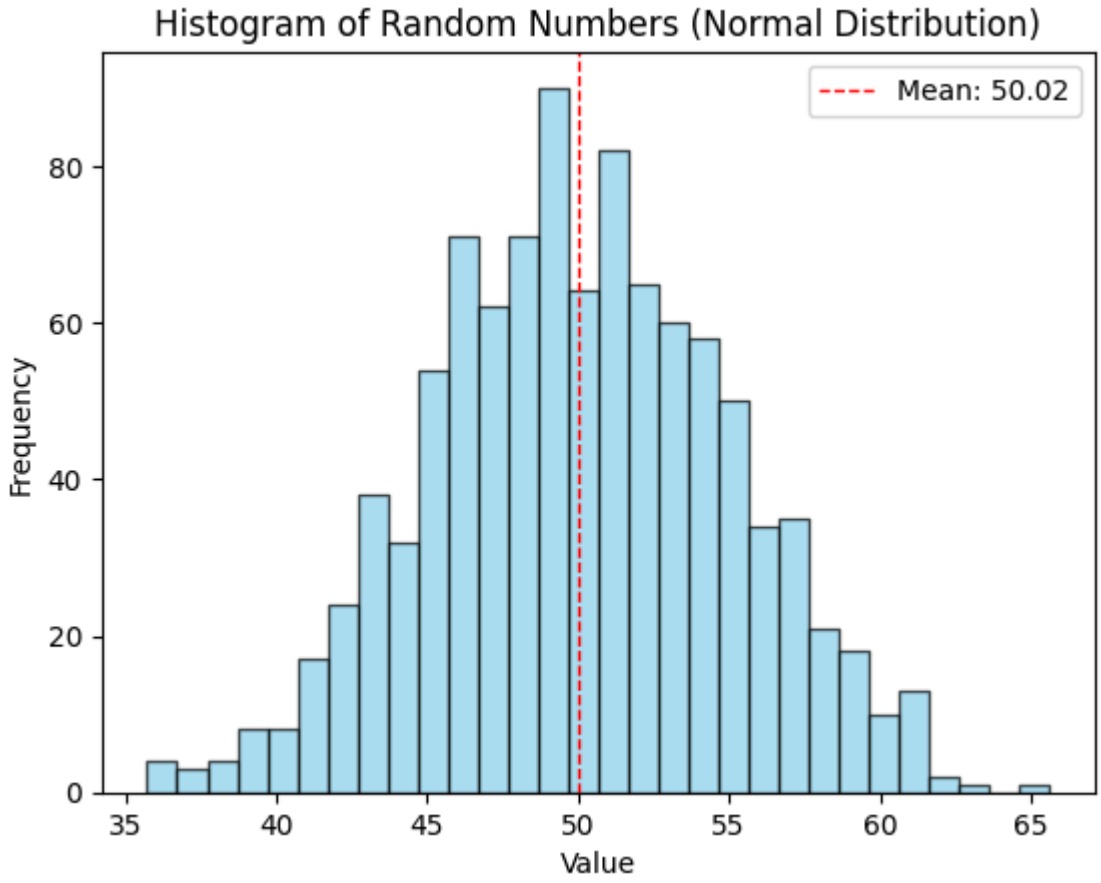
```
#Question 9: Write a Python program to generate 1000 random numbers from a normal
#distribution with mean = 50 and standard deviation = 5. Compute its mean and standard
#deviation using NumPy, and draw a histogram to visualize the distribution.
import numpy as np
import matplotlib.pyplot as plt
```

```
# Generate 1000 random numbers from a normal distribution
mean = 50
std_dev = 5
random_numbers = np.random.normal(loc=mean, scale=std_dev, size=1000)

# Compute the mean and standard deviation of the generated numbers
calculated_mean = np.mean(random_numbers)
calculated_std_dev = np.std(random_numbers)
# Print the results
print(f"Calculated Mean: {calculated_mean}")
print(f"Calculated Standard Deviation: {calculated_std_dev}")

# Plot a histogram to visualize the distribution
plt.hist(random_numbers, bins=30, color='skyblue', edgecolor='black', alpha=0.7)
plt.title('Histogram of Random Numbers (Normal Distribution)')
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.axvline(calculated_mean, color='red', linestyle='dashed', linewidth=1,
label=f'Mean: {calculated_mean:.2f}')
plt.legend()
plt.show()
```

```
Calculated Mean: 50.02476999872907
Calculated Standard Deviation: 4.972585751703926
```




```
#Question 10: You are working as a data analyst for a retail company. The company has
#collected daily sales data for 2 years and wants you to identify the overall sales trend
#daily_sales = [220, 245, 210, 265, 230, 250, 260, 275, 240, 255,
#235, 260, 245, 250, 225, 270, 265, 255, 250, 260]
#● Explain how you would apply the Central Limit Theorem to estimate the average sales
```

```
#with a 95% confidence interval.
#● Write the Python code to compute the mean sales and its confidence interval.
```

Answer

The Central Limit Theorem (CLT) states that if we take many random samples from a population and compute their means, the distribution of those sample means will approximate a normal distribution, regardless of the population's original distribution—provided the sample size is large enough (typically $n \geq 30$).

In this case:

We have daily sales data for 20 days (a small sample).

To apply CLT more robustly, we'd ideally draw multiple samples from a larger dataset. But since we only have one sample, we'll assume it's representative and use the t-distribution to estimate the confidence interval.

```
#python code

import numpy as np
from scipy import stats

# Daily sales data
daily_sales = [220, 245, 210, 265, 230, 250, 260, 275, 240, 255,
               235, 260, 245, 250, 225, 270, 265, 255, 250, 260]

# Convert to NumPy array
sales = np.array(daily_sales)

# Sample statistics
mean_sales = np.mean(sales)
std_sales = np.std(sales, ddof=1)  # Sample standard deviation
n = len(sales)

# t-score for 95% confidence interval
t_score = stats.t.ppf(0.975, df=n-1)

# Margin of error
margin_error = t_score * (std_sales / np.sqrt(n))

# Confidence interval
ci_lower = mean_sales - margin_error
ci_upper = mean_sales + margin_error

# Output
print(f"Mean Sales: {mean_sales:.2f}")
print(f"95% Confidence Interval: ({ci_lower:.2f}, {ci_upper:.2f})")
```

```
Mean Sales: 248.25
95% Confidence Interval: (240.17, 256.33)
```

```
#This means we're 95% confident that the true average daily sales lies between ~240 and ~2
```