Assignment: Statistics Advanced – 1

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

Answer:

In probability theory, a random variable is a function that assigns a numerical value to each outcome of a random experiment.

Random outcome → number

Why we should use random variables - Outcomes themselves can be messy or non-numeric (like “Heads” or “Rain”). Random variables let us:

- Do calculations
- Find probabilities
- Compute averages, variance, etc.

Example 1: Coin toss

Experiment: Toss a coin
Outcomes: `{Heads, Tails}`

Define a random variable **X**:

- X = 1 if Heads
- X = 0 if Tails

Now instead of talking about “Heads” or “Tails”, we work with numbers.

Example 2: Rolling a die

Experiment: Roll a die
Random variable **X** = number shown on the die

Possible values:

$X \in \{1, 2, 3, 4, 5, 6\}$

Each value has a probability.

Question 2: What are the types of random variables?

Answer:

In probability theory, **random variables are mainly of two types** are given below.

1. Discrete Random Variable

A discrete random variable takes countable values (finite or countably infinite).

Characteristics

- ✓ Values can be listed
- ✓ Probabilities are assigned to each value
- ✓ Uses a Probability Mass Function (PMF)

Examples

- ✓ Number of heads in 5 coin tosses → `{0,1,2,3,4,5}`
- ✓ Number of students present in a class
- ✓ Outcome of a dice roll → `{1,2,3,4,5,6}`

2. Continuous Random Variable

A continuous random variable takes uncountable values within a range.

Characteristics

- ✓ Values are measured, not counted
- ✓ Probability at an exact value is zero
- ✓ Uses a Probability Density Function (PDF)

Examples

- ✓ Height of a person
- ✓ Time taken to complete a race
- ✓ Temperature at a place

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

Answer:

1. Discrete Distribution

A discrete distribution describes a discrete random variable, which takes countable values.

Features are

- ✓ Values are finite or countably infinite
- ✓ Probability is assigned to each exact value
- ✓ Uses a Probability Mass Function (PMF)
- ✓ P(X=x)P(X = x)P(X=x) can be greater than 0

Examples

- ✓ Number of heads in coin tosses
- ✓ Number of students in a class
- ✓ Dice outcomes (1–6)

2. Continuous Distribution

A continuous distribution describes a continuous random variable, which takes any value in a range.

Features are

- Values are uncountable
- Probability is assigned over an interval
- Uses a Probability Density Function (PDF)
- P(X=x)=0P(X = x) = 0P(X=x)=0 for any exact value

Examples

- Height of a person
- Time taken to finish a race
- Temperature

Discrete Distribution Vs. Continuous Distribution

Feature	Discrete Distribution	Continuous Distribution
Values	Countable	Uncountable
Function	PMF	PDF
Exact value probability	>0>0>0 possible	Always 0
Probability calculation	Sum of probabilities	Area under curve
Graph	Bar chart	Smooth curve

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

Answer:

A 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, and the probability of success remains constant.

Conditions for a Binomial Distribution

A random experiment follows a binomial distribution if:

- ✓ The number of trials nnn is fixed
- ✓ Each trial is independent
- ✓ Each trial has two outcomes (success/failure)
- ✓ Probability of success ppp is the same for every trial

Probability Formula

If XXX is the number of successes in nnn trials, then:

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

- ✓ nnn = number of trials
- ✓ kkk = number of successes
- ✓ ppp = probability of success

Example

If a coin is tossed 5 times, what is the probability of getting exactly 3 heads?

Here:

- ✓ $n=5n = 5n=5$
- ✓ $k=3k = 3k=3$
- ✓ $p=0.5p = 0.5p=0.5$

P(X=3)=(53)(0.5)3(0.5)2P(X = 3) = \binom{5}{3}(0.5)^3(0.5)^2P(X=3)=(35)(0.5)3(0.5)2

Uses of Binomial Distribution

- ✓ Modeling pass/fail results in exams
- ✓ Quality control (defective vs non-defective items)
- ✓ Medical trials (success or failure of treatment)
- ✓ Coin toss and survey analysis

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

Answer:

The standard normal distribution is a special type of normal distribution with a mean of 0 and a standard deviation of 1. It is denoted by the random variable $Z \sim N(0, 1)$.

Key Features

- ✓ Bell-shaped and symmetric about 0
- ✓ Mean $\mu=0\mu = 0\mu=0$
- ✓ Standard deviation $\sigma=1\sigma = 1\sigma=1$
- ✓ Total area under the curve = 1

Why It Is Important

1. Standardization (Z-scores)
   Any normal random variable can be converted into the standard normal variable using:

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

   This allows comparison between different datasets.

2. Probability Calculation
   Probabilities for all normal distributions can be found using one standard normal table (Z-table).
3. Widely Used in Statistics

   - ✓ Hypothesis testing
   - ✓ Confidence intervals
   - ✓ Quality control
   - ✓ Data normalization

4. Simplifies Analysis
   Instead of separate tables for different means and standard deviations, one table works for all.

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

Answer:

The Central Limit Theorem (CLT) states that the distribution of the sample mean of a sufficiently large number of independent and identically distributed (i.i.d.) random variables will be approximately normal, regardless of the shape of the original population distribution, provided the population has a finite mean and variance.

Key features…..

- ✓ Applies to **sample means**, not individual observations
- ✓ Works for **any population distribution** (normal, skewed, uniform, etc.)
- ✓ Accuracy improves as **sample size increases** (commonly n≥30n \ge 30n≥30)
- ✓ Mean of sampling distribution = population mean μ\muμ
- ✓ Standard deviation = σn\frac{\sigma}{\sqrt{n}}nσ

Why CLT Is Critical in Statistics

1. Foundation of Inferential Statistics
   Enables hypothesis testing and confidence interval estimation.
2. Justifies Normal Approximation
   Allows use of normal distribution even when population is not normal.
3. Simplifies Real-World Analysis
   Makes complex distributions manageable in practice.
4. Supports Statistical Modeling
   Underlies many statistical methods used in data science and research.

Example

If individual incomes are skewed, the **average income** of large samples will still follow a normal distribution due to CLT.

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

Answer:

Confidence intervals (CIs) are used in statistical analysis to estimate an unknown population parameter (such as a mean or proportion) and to quantify the uncertainty associated with that estimate.

Significance of Confidence Intervals

1. Range of Plausible Values
   A confidence interval provides a range within which the true population parameter is likely to lie.
2. Measure of Uncertainty
   Instead of giving a single point estimate, CIs show how precise or reliable the estimate is.
3. Confidence Level Interpretation
   A 95% confidence interval means that if we repeatedly took samples, about 95% of the constructed intervals would contain the true parameter.
4. Decision Making
   Used in:

   - ✓ Hypothesis testing
   - ✓ Medical and scientific research
   - ✓ Business and policy decisions

5. Comparison Between Groups
   Overlapping or non-overlapping confidence intervals help assess whether differences between groups may be statistically meaningful.

Example

If the average test score is estimated as 70 with a 95% confidence interval of (66, 74), we are 95% confident that the true population mean lies between 66 and 74.

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

Answer:

The expected value (also called the mean or mathematical expectation) of a probability distribution represents the long-run average outcome of a random variable if an experiment were repeated many times.

Concept of Expected Value

It is a weighted average of all possible values of a random variable, where the weights are their probabilities.

Formula

For a discrete random variable:

E(X)=∑x P(X=x)E(X) = \sum x \, P(X = x)E(X)=∑xP(X=x)

For a continuous random variable:

E(X)=∫x f(x) dxE(X) = \int x \, f(x)\, dxE(X)=∫xf(x)dx

where f(x)f(x)f(x) is the probability density function.

Intuitive Meaning

Expected value does not necessarily represent a value that will occur in a single trial. Instead, it indicates the average result over a large number of trials.

Example

If a fair die is rolled:

E(X)=1+2+3+4+5+66=3.5E(X) = \frac{1+2+3+4+5+6}{6} = 3.5E(X)=61+2+3+4+5+6=3.5

You'll never roll a 3.5, but over many rolls, the average outcome approaches 3.5.

Importance of Expected Value

- ✓ Summarizes the center of a distribution
- ✓ Used in decision-making under uncertainty
- ✓ Fundamental in economics, finance, and risk analysis
- ✓ Basis for variance and standard deviation calculations

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.

(Include your Python code and output in the code box below.)

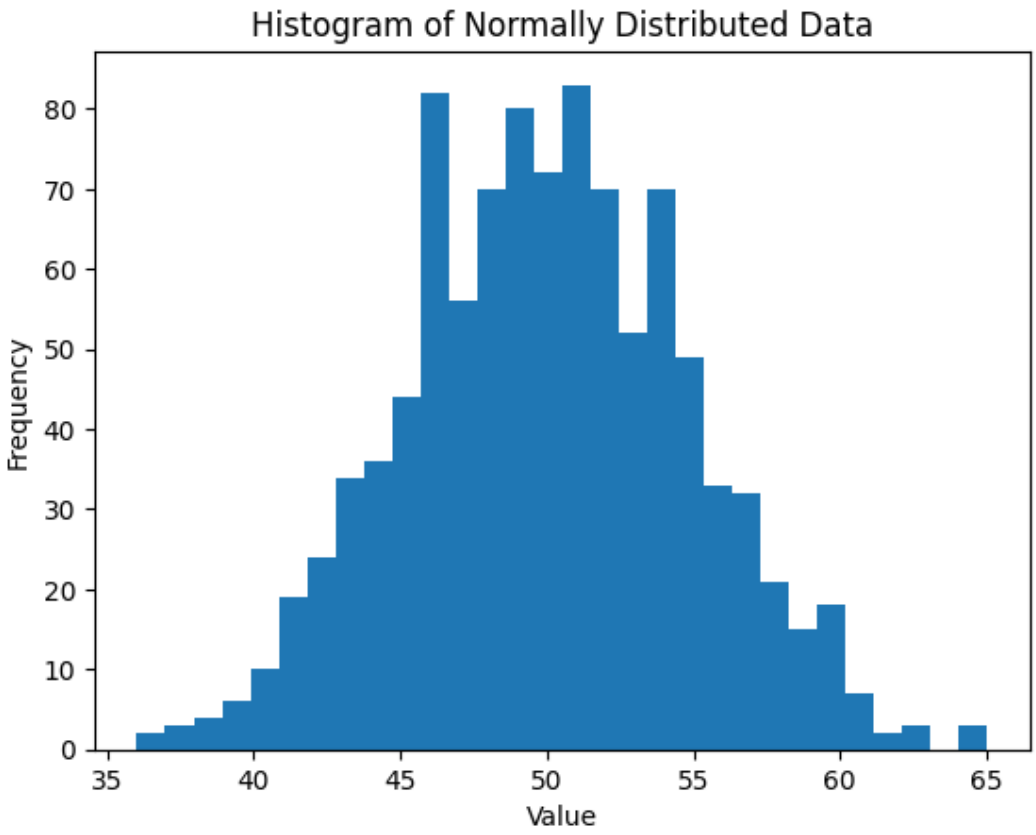
Answer:

```
import numpy as np

import matplotlib.pyplot as plt

# Generate 1000 random numbers from a normal distribution

data = np.random.normal(loc=50, scale=5, size=1000)

# Compute mean and standard deviation

mean_value = np.mean(data)

std_value = np.std(data)

print("Mean:", mean_value)

print("Standard Deviation:", std_value)

# Plot histogram

plt.hist(data, bins=30)

plt.title("Histogram of Normally Distributed Data")

plt.xlabel("Value")

plt.ylabel("Frequency")

plt.show()
```

Output:

Mean: 49.99438493073758

Standard Deviation: 4.86257792421607



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.

(Include your Python code and output in the code box below.)

Answer: Below the python program

```
import numpy as np

from scipy import stats

# Given 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)

# 95% confidence interval using CLT

z_score = 1.96

margin_of_error = z_score * (std_sales / np.sqrt(n))

lower_bound = mean_sales - margin_of_error

upper_bound = mean_sales + margin_of_error

print("Mean Daily Sales:", mean_sales)

print("95% Confidence Interval:", (lower_bound, upper_bound))
```

Output:

Mean Daily Sales: 248.2595% Confidence Interval:
np.float64(240.68312934041109,
np.float64(255.816687065958891

Explanation of the above program…..

Applying the Central Limit Theorem (CLT)

Even though you have daily sales data, individual daily sales may fluctuate and may not be perfectly normally distributed.
The Central Limit Theorem (CLT) helps us here.

How CLT applies:

- ✓ When the sample size is sufficiently large (typically n≥30n \ge 30n≥30; in practice, even moderate sizes work),
- ✓ The sampling distribution of the sample mean becomes approximately normal, regardless of the original data distribution.

Because of this:

- ✓ We can assume the sample mean follows a normal distribution
- ✓ We can construct a 95% confidence interval for the true average daily sales

95% Confidence Interval Formula (Using CLT)…

CI=x¯±zα/2×sn\text{CI} = \bar{x} \pm z_{\alpha/2} \times \frac{s}{\sqrt{n}}CI=x¯±zα/2×ns

Where:

- ✓ x¯\bar{x}x¯ = sample mean
- ✓ sss = sample standard deviation
- ✓ nnn = sample size
- ✓ zα/2=1.96z_{\alpha/2} = 1.96zα/2=1.96 for a 95% confidence level