Statistics Advanced - 1| **Assignment**

Instructions: Carefully read each question. Use Google Docs, Microsoft Word, or a similar tool to create a document where you type out each question along with its answer. Save the document as a PDF, and then upload it to the LMS. Please do not zip or archive the files before uploading them. Each question carries 20 marks.

Total Marks: 200

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

Answer:

A random variable in probability theory is a variable whose possible values are outcomes of a random experiment.

- It assigns a numerical value to each outcome of a random process.
- It can be thought of as a function that maps outcomes from a sample space to numbers.

There are two main types:

1. Discrete random variable – takes countable values (e.g., number of heads in 3 coin tosses).
2. Continuous random variable – takes values from an interval of real numbers (e.g., the exact height of students in a class).

Question 2: What are the types of random variables?

Answer:

There are mainly **two types of random variables** in probability theory:

1. **Discrete Random Variable**
    - Takes a **countable** number of values.

- **Examples:**
    - Number of heads in 5 coin tosses (0, 1, 2, 3, 4, 5).
    - Number of students present in a class.
- Probability is described using a **probability mass function (PMF)**.

2. **Continuous Random Variable**
    - Takes an **uncountable, infinite set of values** within an interval.
    - **Examples:**
        - Height of a person (e.g., 160.2 cm, 160.25 cm, 160.253 cm…).
        - Time taken to run a race.
    - Probability is described using a **probability density function (PDF)**.

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

Answer:

1. Discrete Distributions

- Deal with **discrete random variables** (countable outcomes).
- The probability of each possible value is given by a **probability mass function (PMF)**.
- The sum of probabilities of all possible values equals 1.
- **Examples:**
    - **Binomial distribution** (number of successes in coin tosses).
    - **Poisson distribution** (number of calls at a call center in an hour).

2. Continuous Distributions

- Deal with **continuous random variables** (uncountably infinite outcomes).

- Probability is described using a **probability density function (PDF)**.
- The probability of any single exact value is **zero**; we calculate probability over intervals (e.g., P(1.5 < X < 2.5)).
- The total area under the PDF curve is 1.
- **Examples:**
    - **Normal distribution** (heights, test scores).
    - **Exponential distribution** (time between arrivals in a queue).

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

Answer:

A **binomial distribution** is a type of **discrete probability distribution** that describes the number of successes in a fixed number of independent trials, where each trial has only two possible outcomes: **success** or **failure**.

Key Features

1. **Fixed number of trials (n)** → e.g., tossing a coin 10 times.
2. **Two outcomes in each trial** → success (say, head) or failure (tail).
3. **Constant probability of success (p)** in each trial.
4. **Trials are independent** → the outcome of one does not affect another.

Probability Formula

IfX is the number of successes in nnn trials:

$P(X = k) = (kn)\cdot pk\cdot (1 - p)n - k$

Where:

- $(nk) = n!\, k!\, (n - k)!$ \binom{n}{k} = \frac{n!}{k!\, (n - k)!}$(kn) = k!\, (n -$
  (number of ways to choose k successes)
- p = probability of success

- $1-p$ = probability of failure

- K = number of successes

Examples of Use

- **Coin tosses:** Probability of getting exactly 3 heads in 5 tosses.
- **Quality control:** Probability that 2 out of 10 items in a batch are defective.
- **Elections:** Probability that exactly 60 out of 100 voters favor a candidate.

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. It is a continuous probability distribution with:

- **Mean (μ) = 0**
- **Standard deviation (σ) = 1**
- Its graph is the familiar **bell-shaped curve**, symmetric about zero.

The random variable that follows this distribution is called a **standard normal variable (Z)**.

Why it's important

1. **Basis for the Z-score**
   - Any normal random variable X with mean μ and standard deviation σ can be converted to a standard normal variable using:
     $Z = \sigma X - \mu$
   - This process is called **standardization**.
2. **Simplifies probability calculations**
   - Instead of creating separate tables for every normal distribution, statisticians use the **Z-table** (standard normal table) to find probabilities.

3. **Used in hypothesis testing and confidence intervals**
    - Z-scores help determine how extreme a data point is compared to the mean.
    - Critical in testing significance and constructing confidence intervals.

Example

- Suppose test scores are normally distributed with mean = 70 and standard deviation = 10.
- A student scored 85.
- Z-score = (85−70)/10=1.5.
- This means the student is **1.5 standard deviations above the mean**.

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

Answer:

The **Central Limit Theorem (CLT)** is one of the most important results in statistics.

What it says

The CLT states that:

- If you take many random samples of size nnn from **any population** (with finite mean μμμ and variance σ2σ^2σ2),
- The distribution of the **sample means** will approximate a **normal distribution** as nnn becomes large,
- Regardless of the shape of the original population.

Formally:

$\bar{X} \sim N(\mu, n\sigma 2) as\ n \to \infty$

Why it’s critical

1. **Normal approximation**
    - Even if data isn’t normally distributed, the CLT lets us use the normal distribution to approximate sampling distributions.
2. **Foundation for inferential statistics**
    - Hypothesis testing, confidence intervals, and regression all rely on the CLT.
3. **Predictability**
    - It ensures that averages of samples behave in a predictable, normal way, making statistical inference possible.

Simple Example

- Suppose exam scores in a school are skewed (not normal).
- You randomly take samples of 50 students and compute their average scores.
- If you repeat this many times, the distribution of those **sample averages** will look approximately normal, even though the original scores were skewed.

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

Answer:

A **confidence interval (CI)** gives a range of values that is likely to contain the true population parameter (like mean or proportion) with a certain level of confidence.

Why it’s significant

1. **Gives more information than a point estimate**
    - A sample mean alone doesn’t tell us how precise it is.
    - A CI shows the likely range for the true population mean.
2. **Accounts for uncertainty**

- Because samples vary, CIs reflect the uncertainty in estimating population parameters.

3. **Links to probability**
   - A 95% CI means: If we took many samples and built CIs each time, about 95% of those intervals would contain the true parameter.
4. **Used in decision-making**
   - Helps determine if a result is statistically significant.
   - Example: If a 95% CI for a difference between two treatments is (2, 5), we can be confident the true difference is positive.

Example

- Suppose the average height of a sample of 100 students is 165 cm, with a 95% CI of (163, 167).
- This means we are 95% confident that the **true average height** of all students lies between 163 cm and 167 cm.

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

Answer:

The **expected value** (also called **mean** of a probability distribution) is the long-run average outcome of a random variable if an experiment is repeated many times.

It tells us the **"center" or average result** we should expect from a probability distribution.

Definition

For a random variable X::

- **Discrete case:**

$E[X] = i\sum xi \cdot P(xi)$

(sum of each value times its probability)

- **Continuous case:**

$E[X] = \int - \infty\infty x \cdot f(x)dx$

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

Examples

1. **Coin toss (fair coin):**
   - $Let\ X = 1\ for\ heads,\ X = 0\ for\ tails.$
   - $E[X] = 1 \cdot 0.5 + 0 \cdot 0.5 = 0.5.$
   - On average, half of the tosses will be heads.
2. **Rolling a fair die:**
   - Possible outcomes: 1, 2, 3, 4, 5, 6 (each with probability 1/61/61/6).
   - E[X]=61+2+3+4+5+6=3.5.
   - You'll never roll a 3.5, but it represents the average in the long run.

Why it matters

- Helps in **decision-making under uncertainty** (like gambling, insurance, investments).
- Forms the basis of concepts like **variance, risk, and utility theory**.

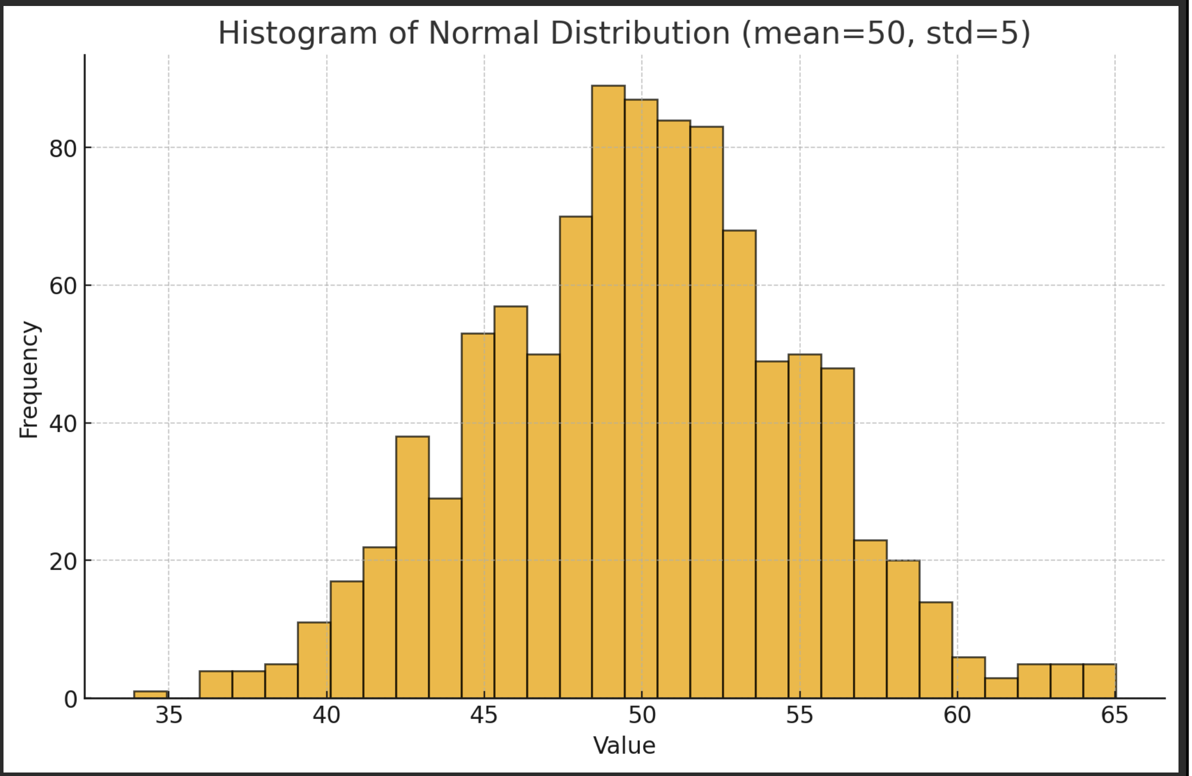
Question 9: Write a Python program to generate 1000 random numbers from a normal distribution with mean = 50 and standard deviation = 5. Compute its mean and standard deviation using NumPy, and draw a histogram to visualize the distribution.

(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 normal distribution
```

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

# Compute mean and standard deviation
mean = np.mean(data)
std_dev = np.std(data)

print(f"Sample Mean: {mean:.2f}")
print(f"Sample Standard Deviation: {std_dev:.2f}")

# Plot histogram
plt.hist(data, bins=30, edgecolor='black', alpha=0.7)
plt.title("Histogram of Normally Distributed Data (μ=50, σ=5)")
plt.xlabel("Value")
plt.ylabel("Frequency")
plt.show()
```



Here's the result:

- **Sample Mean ≈ 49.95**
- **Sample Standard Deviation ≈ 5.06**

```
The histogram shows a bell-shaped curve, which matches the
expected normal distribution with mean 50 and standard deviation
5.
```

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.

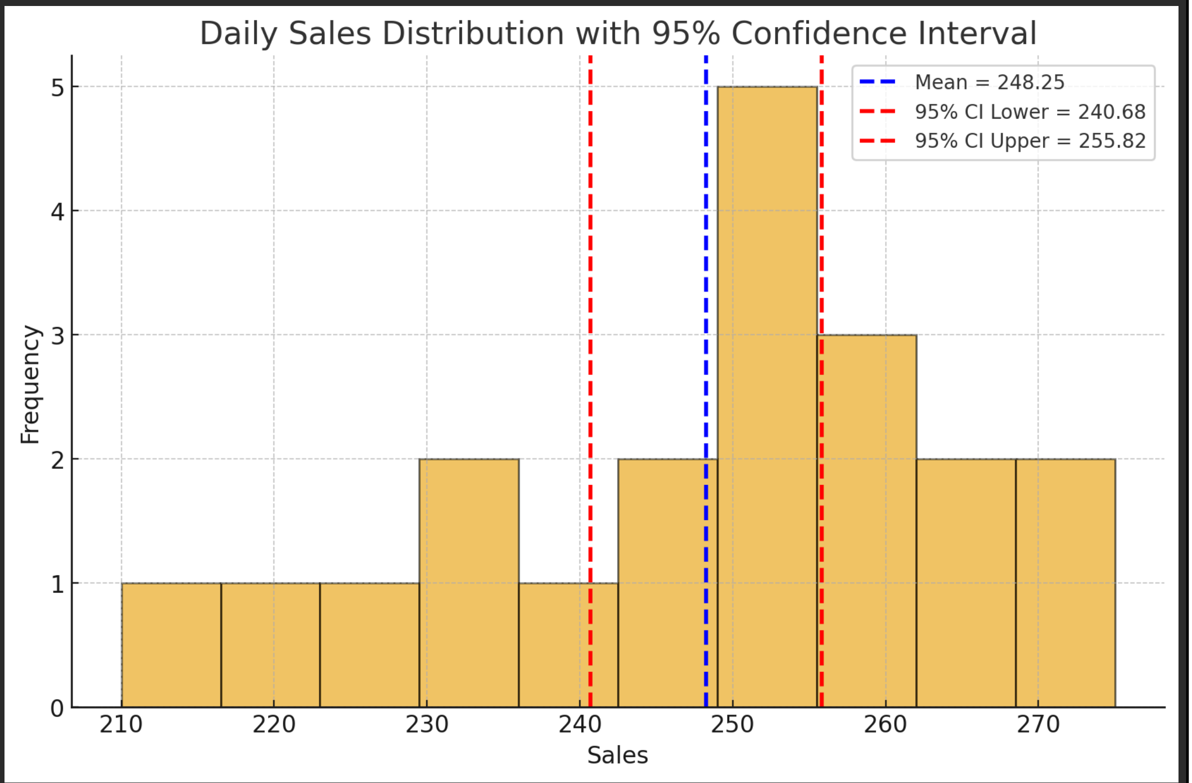
daily_sales = [220, 245, 210, 265, 230, 250, 260, 275, 240, 255,
235, 260, 245, 250, 225, 270, 265, 255, 250, 260]

● 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:

```
import numpy as np
import scipy.stats as st

# 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
data = np.array(daily_sales)

# Compute mean and standard error
mean_sales = np.mean(data)
sem = st.sem(data)  # standard error of the mean

# 95% confidence interval using t-distribution
ci = st.t.interval(0.95, len(data)-1, loc=mean_sales, scale=sem)

print("Mean Sales:", mean_sales)
print("95% Confidence Interval:", ci)
```



Here's the visualization:

- The **blue dashed line** marks the mean daily sales (~248.25).
- The **red dashed lines** show the 95% confidence interval (~240.7 to 255.8).

This makes it clear where the true average daily sales likely fall.