NAME:- ASHISH MAHAPATRA
EMAIL ID : Mahapatraa665@gmail.com
ASSIGNMENT NAME : Statistics Advanced - 1

Github LINK :LINK

Drive Link :LINK

Assignment

Statistics Advanced - 1

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

ANS:- random variable in probability theory is a variable whose possible values are determined by the outcomes of a random experiment.

- Formally, it is a function that maps each outcome of a sample space to a numerical value.

- It allows us to work with random events using mathematics, by assigning numbers to outcomes.

Question 2: What are the types of random variables?

ANS:- Types:

1. Discrete Random Variable – takes a countable number of values.
   Example: Number of heads in 5 coin tosses → {0,1,2,3,4,5}.

2. Continuous Random Variable – takes infinitely many values within an interval.
   Example: The time taken for a computer to complete a task (could be any real number ≥ 0).

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

ANS:-1. Discrete Distribution

- Deals with **discrete random variables** (countable outcomes).

- Probability is assigned to **individual values**.

- Represented by **Probability Mass Function (PMF)**.

- Example: Binomial distribution (number of heads in 5 coin tosses).

---

2. Continuous Distribution

- Deals with **continuous random variables** (uncountable outcomes).
- Probability of any **single exact value is zero**; probabilities are assigned over **intervals**.
- Represented by **Probability Density Function (PDF)**.
- Example: Normal distribution (heights of people, exam scores).

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

ANS:- A **binomial distribution** is a type of 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.

Key Characteristics:

1. Fixed number of trials (**n**).
2. Each trial has only **two outcomes** (success/failure).
3. Probability of success (**p**) remains constant in every trial.
4. Trials are **independent** of each other.

Probability Formula:

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

N = number of trials

K = number of successes

P = probability of success

1−p1-p1−p = probability of failure

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

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

- The mean (μ) = 0
- The standard deviation (σ) = 1
- It is bell-shaped and symmetric about the mean.

The variable that follows this distribution is called the standard normal variable (Z), and values are often referred to as z-scores.

Why it is important?

Simplifies calculations – Any normal distribution can be converted into the standard normal using

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

Widely used in statistics – for hypothesis testing, confidence intervals, and control charts.

Probability tables (Z-tables) – provide cumulative probabilities for the standard normal, making it easy to find probabilities for any normal distribution.

Foundation for inferential statistics – many statistical methods assume normality.

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

ANS:- The CLT states that:

When we take many random samples of sufficiently large size nnn from any population (with mean μ\muμ and finite variance σ2\sigma^2σ2), the **sampling distribution of the sample mean** will be approximately **normal**, regardless of the original population's distribution.

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

Why it is Critical?

1. **Foundation of Inferential Statistics** – It allows us to use the normal distribution to make conclusions about population means, even if the population itself is not normal.

2. **Practical Use** – With large enough sample sizes (usually n≥30n \geq 30n≥30), the sample mean behaves like a normal variable.

3. **Supports Confidence Intervals & Hypothesis Testing** – Enables calculation of probabilities and critical values.

4. **Universal Applicability** – Works for many types of data (finance, manufacturing, surveys, etc.).

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

ANS:- A **confidence interval (CI)** is a range of values, calculated from sample data, that is likely to contain the true population parameter (like mean or proportion) with a certain level of confidence.

Significance of Confidence Intervals

1. **Provides an Estimate with Precision**
    - Instead of giving just a single point estimate (e.g., sample mean = 50), a CI gives a range (e.g., 47 to 53) showing the possible values of the population parameter.
2. **Accounts for Sampling Error**
    - Acknowledges uncertainty in sample data and adjusts for variability.
3. **Connects to Probability**
    - A 95% CI means: if we repeatedly take samples and build CIs, about 95% of them will capture the true population parameter.
4. **Supports Decision-Making**
    - Widely used in research, business, and quality control to judge reliability of results.

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

ANS:- The **expected value (EV)** of a probability distribution is the long-run average or mean value of a random variable if an experiment is repeated many times. It represents the “center of gravity” of the distribution.

For a random variable XXX:

- **Discrete case:**

$$E[X] = \sum_i x_i \cdot P(x_i)$$

- **Continuous case:**

$$E[X] = \int_{-\infty}^{\infty} x \cdot f(x)\, dx$$

Interpretation:

- It is the **weighted average** of all possible values of the random variable, with probabilities as weights.
- Example: If a dice is rolled,

$$E[X] = \frac{1+2+3+4+5+6}{6} = 3.5$$

Even though 3.5 is not a possible outcome, it's the average 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. (Include your Python code and output in the code box below.)

ANS:- **PYTHON CODE**

```
# Import necessary libraries

import numpy as np

import matplotlib.pyplot as plt
```

```
# Set parameters
mean = 50
std_dev = 5
num_samples = 1000

# Generate 1000 random numbers from normal distribution
data = np.random.normal(loc=mean, scale=std_dev, size=num_samples)

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

print(f"Calculated Mean: {calculated_mean:.2f}")
print(f"Calculated Standard Deviation: {calculated_std:.2f}")

# Plot histogram
plt.figure(figsize=(8,5))
plt.hist(data, bins=30, color='skyblue', edgecolor='black')
plt.title('Histogram of Normally Distributed Data')
plt.xlabel('Value')
plt.ylabel('Frequency')
plt.grid(True)
```

plt.show()

OUTPUT

Calculated Mean: 50.08

Calculated Standard Deviation: 5.02

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.

ANS:-1: Applying the Central Limit Theorem (CLT)

1. The daily sales data represents a sample of all possible daily sales.

2. According to the CLT, if we take repeated random samples of daily sales, the sampling distribution of the sample mean will be approximately normal, even if the underlying data isn't perfectly normal, provided the sample size is reasonably large.

3. This allows us to compute a confidence interval (CI) for the population mean (true average daily sales).

4. A 95% confidence interval gives a range in which we are 95% confident that the true average daily sales lies.

The formula for a 95% CI for the mean is:

$$\text{CI} = \bar{x} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

- 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 95% confidence

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 mean and standard deviation
mean_sales = np.mean(sales)
std_sales = np.std(sales, ddof=1)  # ddof=1 for sample std deviation
n = len(sales)

# 95% confidence interval using z-value
confidence_level = 0.95
```

```
z = stats.norm.ppf(0.975)  # two-tailed 95% CI


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

ci_lower = mean_sales - margin_of_error

ci_upper = mean_sales + margin_of_error


print(f"Mean Daily Sales: {mean_sales:.2f}")

print(f"95% Confidence Interval: ({ci_lower:.2f}, {ci_upper:.2f})")
```

OUTPUT

```
Mean Daily Sales: 250.25

95% Confidence Interval: (241.84, 258.66)
```