Statistics Advanced - 1| Assignment

Question 1: What is a random variable in probability theory?
Answer : A **random variable** assigns a **number** to each possible outcome of a random process or experiment.

A **random variable** is a function that maps outcomes from a **sample space (S)** to **real numbers (ℝ)**:

X:S→RX: S \rightarrow \mathbb{R}X:S→R

That means for every possible outcome sss in the sample space SSS, X(s)X(s)X(s) gives a real number.

Question 2: What are the types of random variables?
Answer : In **probability theory**, **random variables** are mainly divided into two major types — **Discrete** and **Continuous** — based on the kind of values they can take.

1. Discrete Random Variable : A **discrete random variable** takes **countable distinct values** — either **finite** or **countably infinite**.

Examples:

- Number on a die → {1, 2, 3, 4, 5, 6}
- Number of heads in 3 coin tosses → {0, 1, 2, 3}
- Number of students present in class

Key Features:

- Possible values are **separate and countable**.
- Represented by a **Probability Mass Function (PMF)**.
- The sum of all probabilities is **1**:
  ∑P(X=xi)=1\sum P(X = x_i) = 1∑P(X=xi)=1

2. Continuous Random Variable

A **continuous random variable** takes **infinitely many values** within a range or interval.

Examples:

- Height or weight of a person
- Time taken to finish a race
- Temperature in a city

Key Features:

- Values are **uncountable and continuous**.
- Represented by a **Probability Density Function (PDF)**.
- Probability at a single point is **0**:
  P(X=a)=0P(X = a) = 0P(X=a)=0
- Probability is defined over an interval:
  P(a<X<b)=∫abf(x) dxP(a < X < b) = \int_a^b f(x)\,dxP(a<X<b)=∫abf(x)dx

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

Answer:

🔹 Difference between Discrete and Continuous Distributions

Feature	Discrete Distribution	Continuous Distribution
Definition	A distribution that shows probabilities of a **discrete random variable** (which takes countable values).	A distribution that shows probabilities of a **continuous random variable** (which takes values over an interval).
Possible Values	Countable or finite values (e.g. 0, 1, 2, 3...).	Uncountably infinite values within a range (e.g. any value between 0 and 1).
Probability Function	Represented by a **Probability Mass Function (PMF)**.	Represented by a **Probability Density Function (PDF)**.
Probability of a Single Value	P(X=x)P(X = x)P(X=x) can be **non-zero**.	P(X=x)=0P(X = x) = 0P(X=x)=0; probability is only meaningful over a range.
Total Probability	Sum of all probabilities = 1	

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

Answer: # Binomial Distribution – Definition

A **Binomial Distribution** is a **discrete probability distribution** that describes the **number of successes** in a fixed number of **independent trials**, each having only **two possible outcomes** — success or failure.

A random variable XXX follows a binomial distribution if:

1. The experiment consists of **n independent trials**.
2. Each trial has **two outcomes** — Success (S) or Failure (F).
3. The **probability of success (p)** remains **constant** in all trials.
4. The random variable XXX counts the **number of successes** in those nnn trials.

Probability Formula

If XXX = number of successes in nnn trials,
then the probability of getting exactly xxx successes is:

P(X=x)=(nx)px(1−p)n−xP(X = x) = \binom{n}{x} p^x (1 - p)^{n - x}P(X=x)=(xn)px(1−p)n−x

where:

- nnn = total number of trials
- xxx = number of successes
- ppp = probability of success in one trial
- (1−p)(1-p)(1−p) = probability of failure
- (nx)=n!x!(n−x)!\binom{n}{x} = \dfrac{n!}{x!(n-x)!}(xn)=x!(n−x)!n!

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

Answer : # Standard Normal Distribution – Definition

The **Standard Normal Distribution** is a special type of **normal (Gaussian) distribution** that has:

Mean (μ)=0andStandard Deviation (σ)=1\text{Mean (μ)} = 0 \quad \text{and} \quad \text{Standard Deviation (σ)} = 1Mean (μ)=0andStandard Deviation (σ)=1

The random variable that follows this distribution is called a **standard normal variable**, usually denoted by **Z**.
Hence, it is also called the **Z-distribution**.

Probability Density Function (PDF)

f(z)=12πe−z22f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}f(z)=2π1e−2z2

where

- zzz = standard normal variable
- The total area under the curve = **1**.

Shape and Properties

- It is a **bell-shaped**, **symmetric curve** about z=0z = 0z=0.
- Mean = Median = Mode = 0
- Most values lie close to the mean.
- About **68%** of the area lies between z=−1z = -1z=−1 and z=+1z = +1z=+1.
- About **95%** between z=−2z = -2z=−2 and z=+2z = +2z=+2.
- About **99.7%** between z=−3z = -3z=−3 and z=+3z = +3z=+3.

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

Answer: # Central Limit Theorem (CLT) – Definition

The **Central Limit Theorem (CLT)** states that:

> When we take **a large number of independent random samples** from any population (with finite mean μμμ and finite standard deviation σσσ), the **sampling distribution of the sample mean** tends to become **approximately normal**, regardless of the original population's shape.

Mathematical Form

If X1,X2,X3,…,XnX_1, X_2, X_3, \dots, X_nX1,X2,X3,…,Xn are independent random variables with
mean μμμ and standard deviation σσσ, then the **sample mean**:

X¯=1n∑i=1nXi\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_iX¯=n1i=1∑nXi

has approximately a **normal distribution** when nnn is large:

X¯~N(μ,σn)\bar{X} \sim N\left(μ, \frac{σ}{\sqrt{n}} \right)X¯~N(μ,nσ)

Standardized Form (Z-score)

Z=X¯−μσ/nZ = \frac{\bar{X} - μ}{σ / \sqrt{n}}Z=σ/nX¯−μ

As n→∞n \to \inftyn→∞, ZZZ follows the **standard normal distribution**.

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

Answer : # Confidence Intervals – Definition

A **confidence interval (CI)** is a **range of values**, derived from sample data, that is likely to contain the **true population parameter** (such as the mean or proportion) with a certain level of confidence.

🔹 Meaning

If we say we have a **95% confidence interval** for the population mean μμμ:

> It means we are **95% confident** that the true value of μμμ lies within that interval.

In other words, if we repeated the same sampling many times, **95 out of 100 intervals** constructed in this way would contain the true mean.

🔹 General Formula

Confidence Interval=Sample Estimate ± Margin of Error\text{Confidence Interval} = \text{Sample Estimate} \; \pm \; \text{Margin of Error}Confidence Interval=Sample Estimate±Margin of Error

For the population mean:

CI=X¯±Zα/2(σn)CI = \bar{X} \pm Z_{\alpha/2} \left(\frac{σ}{\sqrt{n}} \right)CI=X¯±Zα/2(nσ)

where:

- X¯\bar{X}X¯ = sample mean
- σσσ = population standard deviation

- nnn = sample size
- Zα/2Z_{\alpha/2}Zα/2 = critical value from the **standard normal distribution** for the desired confidence level
  (e.g., 1.96 for 95%, 2.58 for 99%)

🔹 Example

Suppose the average weight of a sample of 100 people is 70 kg with a standard deviation of 8 kg.
A 95% CI for the population mean is:

70±1.96×8100=70±1.56870 \pm 1.96 \times \frac{8}{\sqrt{100}} = 70 \pm 1.56870±1.96×1008=70±1.568 ⇒(68.43, 71.57)\Rightarrow (68.43,\; 71.57)⇒(68.43,71.57)

So, we are **95% confident** that the true population mean lies between **68.43 kg and 71.57 kg**.

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

Answer : # Expected Value – Definition

The **expected value (EV)** of a random variable is the **long-run average** or **mean value** of its possible outcomes, **weighted by their probabilities**.

In simple terms:

> It is what you would **expect to happen on average** if you repeated an experiment many times.

🔹 Mathematical Definition

For a random variable XXX:

- **If XXX is discrete:**
  E(X)=∑xi P(xi)E(X) = \sum x_i \, P(x_i)E(X)=∑xiP(xi)
- **If XXX is continuous:**
  E(X)=∫−∞∞x f(x) dxE(X) = \int_{-\infty}^{\infty} x \, f(x)\, dxE(X)=∫−∞∞xf(x)dx

where

- xix_ixi or xxx = possible values of the random variable

- P(xi)P(x_i)P(xi) or f(x)f(x)f(x) = probability (or probability density)

🔹 Interpretation

The expected value represents the **center** or **balance point** of a probability distribution.
It's not necessarily an actual observed value, but an **average outcome** over the long term.

🔹 Example (Discrete Case)

Suppose a fair die is rolled once.
Possible outcomes: 1, 2, 3, 4, 5, 6
Each has probability = 1/6.

E(X)=(1+2+3+4+5+6)×16=216=3.5E(X) = (1+2+3+4+5+6) \times \frac{1}{6} = \frac{21}{6} = 3.5E(X)=(1+2+3+4+5+6)×61=621=3.5

➡️ The expected value of the die roll is **3.5**, even though 3.5 is not an actual face of the die — it's the **average result** over many rolls.

🔹 Example (Continuous Case)

If XXX is a continuous random variable with PDF f(x)f(x)f(x),
then E(X)=∫xf(x)dxE(X) = \int x f(x) dxE(X)=∫xf(x)dx.
For example, the expected lifetime of a light bulb can be computed using this formula.

🔹 Properties of Expected Value

1. **Linearity:**
   E(aX+b)=aE(X)+bE(aX + b) = aE(X) + bE(aX+b)=aE(X)+b

2. **Sum of Variables:**
   E(X+Y)=E(X)+E(Y)E(X + Y) = E(X) + E(Y)E(X+Y)=E(X)+E(Y)

3. **Constant:**
   E(c)=cE(c) = cE(c)=c, where ccc is a constant.

🔹 Importance / Significance

- Helps **summarize** a probability distribution with a single representative value.
- Used to **predict average outcomes** in games, investments, insurance, and risk analysis.
- Basis for finding **variance** and **standard deviation** (measures of spread).
- Key tool in **decision theory** — choosing actions that maximize expected value.

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 (mean=50, std=5)
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("Computed Mean:", mean_value)
print("Computed Standard Deviation:", std_value)

# Draw histogram
plt.hist(data, bins=30, color='skyblue', edgecolor='black')
plt.title('Normal Distribution (mean=50, std=5)')
plt.xlabel('Value')
plt.ylabel('Frequency')

# Show the mean line
plt.axvline(mean_value, color='red', linestyle='dashed', linewidth=2, label=f'Mean =
{mean_value:.2f}')
plt.legend()
plt.show()
```

Output : Computed Mean: 49.92
Computed Standard Deviation: 5.08

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 : **Applying the Central Limit Theorem (CLT)**

The **Central Limit Theorem** states that the **sampling distribution of the sample mean** will be approximately normal if the sample size is sufficiently large, even if the original data is not perfectly normal.

How to apply CLT to estimate average sales:

1. Compute the **sample mean** ($\bar{X}$) and **sample standard deviation** (s) from the daily sales data.
2. Use CLT to assume the **sample mean** is approximately normally distributed.
3. Construct a **95% confidence interval** for the population mean using:

$$CI = \bar{X} \pm Z_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right)$$

where:

- n = sample size
- $Z_{\alpha/2} = 1.96$ for 95% confidence level

This gives a range where we are **95% confident** the true average daily sales lie.

Step 2: 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)  # sample standard deviation
n = len(sales)

# 95% confidence interval using Z-score (for large n, or normal
approx)
confidence_level = 0.95
z_score = stats.norm.ppf(0.975)  # two-tailed 95%

margin_of_error = z_score * (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: 249.25
95% Confidence Interval: (240.81, 257.69)