Statistics Advanced - 1

Question 1: What is a random variable in probability theory?
Ans. A **random variable** in probability theory is a variable that assigns a numerical value to each possible outcome of a random experiment.

- It is called *random* because its value depends on the outcome of a chance process.
- Formally, it is a **function** that maps outcomes from the sample space (all possible outcomes of an experiment) to real numbers.

Types of Random Variables:

1. **Discrete Random Variable**
   - Takes on a countable number of possible values.
   - Example: Number of heads in 3 coin tosses → {0, 1, 2, 3}.
2. **Continuous Random Variable**
   - Takes on values from an interval or continuum of numbers.
   - Example: The exact height of students in a class.

✅ **Example:**
 If we roll a fair six-sided die, define XXX as the random variable representing the number that appears:

- Sample space S={1,2,3,4,5,6}S = \{1, 2, 3, 4, 5, 6\}S={1,2,3,4,5,6}.

- Random variable XXX maps each outcome to a number between 1 and 6.

Question 2: What are the types of random variables?
Ans. In probability theory, **random variables** are mainly classified into two broad types, based on the kind of values they can take:

---

1. Discrete Random Variable

- Takes a **finite or countably infinite** set of values.
- Usually represents outcomes we can count.
- Probability distribution is described by a **Probability Mass Function (PMF)**.

✅ Examples:

- Number of heads in 3 coin tosses → {0,1,2,3}
- Number of customers arriving at a shop in an hour.
- Rolling a die → {1,2,3,4,5,6}

---

2. Continuous Random Variable

- Takes **uncountably infinite values** over an interval.
- Usually represents measurements.

- Probability distribution is described by a **Probability Density Function (PDF)**.
- Probability of a single exact value = 0 (we consider ranges/intervals instead).

✅ **Examples:**

- Height or weight of a student.
- Time taken to finish a task.
- Temperature in a city.

(Sometimes also mentioned)

3. Mixed Random Variable

- Has both discrete and continuous components.
- Less common, but useful in advanced probability.

✅ **Example:**

- Insurance claims: with probability ppp, claim = 0 (discrete), otherwise claim amount is continuous over positive values.

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

Ans. **Difference between Discrete and Continuous Distributions**

Feature	Discrete Distribution	Continuous Distribution
Definition	Probability distribution of a **discrete random variable** (countable outcomes).	Probability distribution of a **continuous random variable** (uncountably infinite outcomes).
Possible Values	Finite or countably infinite values.	Any value in an interval (infinite and uncountable).
Probability Function	Defined by **Probability Mass Function (PMF)**.	Defined by **Probability Density Function (PDF)**.
Probability at a Point	P(X=x)P(X = x)P(X=x) can be > 0.	P(X=x)=0P(X = x) = 0P(X=x)=0, but P(a≤X≤b)>0P(a \leq X \leq b) > 0P(a≤X≤b)>0.
Summation / Integration	Probabilities are calculated by **summing** over values.	Probabilities are calculated by **integrating** the PDF over an interval.
Examples	- Tossing a coin (X = number of heads)	

-
  Rolling a die (X = 1–6)
- Number of students in a class | - Height of a student
- Time taken to run 100m
- Temperature of a city |

✅ **Example (Discrete):**
Rolling a die: P(X=4)=16P(X = 4) = \frac{1}{6}P(X=4)=61.

✅ **Example (Continuous):**
If XXX = time to complete a task, then P(X=10)=0P(X = 10) = 0P(X=10)=0, but we can compute P(9.5≤X≤10.5)P(9.5 \leq X \leq 10.5)P(9.5≤X≤10.5) using integration.

Question 4: What is a binomial distribution, and how is it used in probability?
Ans. A **Binomial Distribution** is one of the most important discrete probability distributions in statistics and probability theory.

Definition

A random variable XXX follows a **binomial distribution** if it counts the number of successes in a fixed number of independent trials, each with the same probability of success.

It is denoted as:

X~Binomial(n,p)X \sim \text{Binomial}(n, p)X~Binomial(n,p)

where:

- nnn = number of trials
- ppp = probability of success in each trial
- q=1−pq = 1 - pq=1−p = probability of failure

Probability Formula (PMF)

The probability of getting exactly kkk successes in nnn trials is:

P(X=k)=(nk)pk(1−p)n−k,k=0,1,2,…,nP(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0,1,2,\dots,nP(X=k)=(kn)pk(1−p)n−k,k=0,1,2,…,n

where (nk)=n!k!(n−k)!\binom{n}{k} = \frac{n!}{k!(n-k)!}(kn)=k!(n−k)!n!.

Key Properties

- **Mean**: μ=np\mu = npμ=np
- **Variance**: σ2=np(1−p)\sigma^2 = np(1-p)σ2=np(1−p)

Uses in Probability

The binomial distribution is used when:

1. **Fixed number of trials** (nnn)
2. Each trial has **two possible outcomes** (success/failure, yes/no, pass/fail).
3. Trials are **independent**.
4. Probability of success (ppp) is the **same** for each trial.

✅ **Examples:**

1. Tossing a coin 10 times and finding the probability of getting exactly 6 heads.

2. Quality control: finding the probability that 3 out of 20 products are defective.

3. Exam: probability of passing exactly 7 out of 10 students when each has a 70% chance.

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

Standard Normal Distribution

A **standard normal distribution** is a special case of the normal (Gaussian) distribution.

- It is a **continuous probability distribution**.

- It has:

  - **Mean (μ\muμ) = 0**

  - **Standard deviation (σ\sigmaσ) = 1**

- Its probability density function (PDF) is:

f(z)=12πe−z22,−∞<z<∞f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad -\infty < z < \inftyf(z)=2π1e−2z2,−∞<z<∞

where zzz is called the **standard normal variable**.

Why is it Important?

1. **Basis for Other Normal Distributions**
   - Any normal distribution X~N(μ,σ2)X \sim N(\mu, \sigma^2)X~N(μ,σ2) can be **converted** into the standard normal using:
2. Z=X−μσZ = \frac{X - \mu}{\sigma}Z=σX−μ
   (This process is called **standardization**).
3. **Simplifies Probability Calculations**
   - Tables (Z-tables) and software give probabilities for the standard normal, so we convert general normal problems to this form.
4. **Widely Used in Statistics**
   - Hypothesis testing (z-tests).
   - Confidence intervals.
   - Control charts in quality management.
5. **Central Limit Theorem (CLT)**
   - Many sample means (even from non-normal populations) become approximately normal for large samples, and

standardization turns them into a standard normal.

✅ **Example:**
If students' test scores follow $N(70, 9^2)$, the probability a student scores less than 65 is found by:

$$Z = \frac{65 - 70}{9} = -0.56$$

Then use the **standard normal table** (or software) to find $P(Z < -0.56)$.

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

Ans. **Central Limit Theorem (CLT)**

The **Central Limit Theorem** states that:

> When we take sufficiently large random samples from any population with finite mean μ and finite variance σ^2, the **sampling distribution of the sample mean** $\bar{X}$ will be approximately **normal**, regardless of the shape of the original population distribution.

Formally, if $X_1, X_2, \dots, X_n$ are i.i.d. random variables with mean μ and variance σ^2:

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \;\; \xrightarrow{d}\;\; N(0,1) \quad \text{as } n \to \infty$$

Why is CLT Critical in Statistics?

1. **Normal Approximation**
    - Allows us to use the **normal distribution** to approximate probabilities for sample means, even if the population is not normal.
2. **Foundation of Inferential Statistics**
    - Enables **confidence intervals** and **hypothesis tests** based on the normal (or t-distribution for small samples).
3. **Practical Applications**
    - Quality control (sample averages).
    - Opinion polls (sample proportions).
    - Finance (portfolio returns).
4. **Simplifies Complex Problems**
    - Real-world data rarely follows an exact distribution, but CLT lets us model with the normal, which is mathematically convenient.

✅ **Example:**
Suppose IQ scores in a city are skewed but have mean 100100100 and standard deviation 151515.

- If we take a sample of n=50n=50n=50 people, the distribution of the **sample mean IQ** will be approximately normal with:
    - Mean = 100100100
    - Standard error = 1550≈2.12\frac{15}{\sqrt{50}} \approx 2.125015≈2.12.

This lets us make probability statements about the sample mean IQ.

Question 7: What is the significance of confidence intervals in statistical analysis?
Ans. **Confidence Intervals (CIs)**

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

For example:
If we calculate a 95% confidence interval for the mean weight of students as [58,62][58, 62][58,62], it means we are **95% confident** that the true population mean lies between 58 and 62.

Significance of Confidence Intervals in Statistical Analysis

1. **Estimation with Uncertainty**
    - Instead of giving just a point estimate (like sample mean = 60), CIs provide a **range**, showing the uncertainty around

the estimate.

2. **Link to Sampling Variability**
   - Reflects the natural variation due to random sampling. Wider intervals mean more uncertainty; narrower intervals mean more precision.
3. **Better Decision-Making**
   - Helps policymakers, scientists, and businesses judge reliability before acting on data.
4. **Alternative to Hypothesis Testing**
   - If a confidence interval for the mean difference does **not include 0**, it suggests a statistically significant difference.
5. **Adjustable Confidence Levels**
   - 90%, 95%, 99% confidence levels can be chosen depending on how much certainty vs. precision is needed.

✅ **Example:**

- Suppose a survey of 100 people finds an average height of 165 cm with a 95% CI = [162, 168].

- This means if we repeated the survey many times, about 95% of such intervals would capture the true average height.

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

Ans. **Expected Value (EV)**

The **expected value** of a random variable is the **long-run average outcome** we expect if we repeat a random experiment many times.

It is like the "center of gravity" of a probability distribution.

---

1. For a Discrete Random Variable

If XXX is a discrete random variable with values x1,x2,…,xnx_1, x_2, \dots, x_nx1,x2,…,xn and probabilities P(X=xi)=piP(X=x_i) = p_iP(X=xi)=pi, then:

E[X]=∑i=1nxi piE[X] = \sum_{i=1}^{n} x_i \, p_iE[X]=i=1∑nxipi

2. For a Continuous Random Variable

If XXX is a continuous random variable with probability density function f(x)f(x)f(x), then:

E[X]=∫−∞∞x f(x) dxE[X] = \int_{-\infty}^{\infty} x \, f(x) \, dxE[X]=∫−∞∞xf(x)dx

---

Why is Expected Value Important?

1. **Measure of Central Tendency**

- Represents the average or "mean" of the distribution.

2. **Decision-Making Under Uncertainty**
   - Used in economics, insurance, and gambling to evaluate choices.

3. **Foundation for Variance and Standard Deviation**
   - Variance is defined using the expected value:
     Var(X)=E[(X−E[X])2]\; Var(X) = E[(X - E[X])^2]Var(X)=E[(X−E[X])2].

✅ Examples

- **Discrete case (dice roll):**
  Roll a fair die (X=1,2,3,4,5,6X = 1,2,3,4,5,6X=1,2,3,4,5,6):
  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
  Meaning: the long-run average roll is 3.5.

- **Continuous case:**
  If X~U(0,1)X \sim U(0,1)X~U(0,1) (uniform distribution),
  E[X]=∫01x dx=12E[X] = \int_0^1 x \, dx = \frac{1}{2}E[X]=∫01xdx=21

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.
Ans. import numpy as np

```
import matplotlib.pyplot as plt

# Generate 1000 random numbers from Normal(μ=50, σ=5)
data = np.random.normal(loc=50, scale=5, size=1000)

# Compute sample mean and standard deviation
mean_val = np.mean(data)
std_val = np.std(data)

print("Sample Mean:", mean_val)
print("Sample Standard Deviation:", std_val)

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