Data Science For Engineers
NPTEL PMRF Live Sessions

Teaching Assistant: Parvathy Neelakandan
PhD Student

12 - 08 - 2025

Discrete and continuous random phenomena: Depending on the type of outcome, random phenomena can be discrete or continuous. When the outcomes are finite, we call it as discrete random phenomena and when there are infinite possible outcomes, it is termed as continuous random phenomena.

Definition 1.1. Sample space is the set of all possible outcomes of a random phenomena.

Definition 1.2. Event is a subset of a sample space. Occurence of a head in first two toss of a coin.

Coin Toss Experiments

Sample space:

- Single coin toss: $S = \{H, T\}$
- Two coin tosses: $S = \{HH, HT, TH, TT\}$

For two coin tosses, examples of events include:

- First toss is heads: $= A = \{HH, HT\}$
- Exactly one head: $B = \{HT, TH\}$

Definition 1.4 (Conditional Probability). The conditional probability of event B given event A is:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

provided $P(A) > 0$.

Definition 1.5 (Bayes' Theorem). For events A and B with $P(B) > 0$:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Practice question 1

A diagnostic test for a rare disease has:

- Sensitivity (probability of positive test given disease): 0.95
- Specificity (probability of negative test given no disease): 0.98
- Disease prevalence: 0.001

If a person tests positive, what's the probability they have the disease?

Practice question 2

A fair six-sided die is rolled twice. Define events:

- A: sum of the two rolls is 6
- B: first roll is 3 or greater
- C: second roll is even
- Find $P(A)$, $P(B)$, and $P(C)$.
- Are A and B independent?
- Are B and C independent?
- Find $P(A|B)$.

3.3 Binomial Distribution

Definition 3.2 (Binomial Distribution)**.** A random variable X follows a binomial distribution with parameters n and p, denoted $X \sim \text{Binomial}(n, p)$, if:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n$$

Practice question

Items produced by a manufacturing process have a 5% defective rate. What is the probability that at least five of the thirty randomly chosen goods will be defective?

4.2.1 Normal Distribution

Definition 4.3 (Normal Distribution)**.** A random variable X follows a normal distribution with parameters μ and σ^2, denoted $X \sim N(\mu, \sigma^2)$, if:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

μ refers to the mean and σ^2 refers to variance.

Definition 4.4 (Standard Normal Distribution)**.** The standard normal distribution has $\mu = 0$ and $\sigma^2 = 1$, denoted $Z \sim N(0, 1)$.

Standardization: If $X \sim N(\mu, \sigma^2)$, then:

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

Practice question

Assume that IQ values fall within $N(100, 15^2)$. How often is it that someone has an IQ higher than 100?

Definition 7.1 (Mean). The mean is:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Definition 7.2 (Median). The median is the middle value when observations are arranged in order. For even n, it's the average of the two middle values.

Definition 7.3 (Mode). The mode is the most frequently occurring value in the dataset.

Practice question

Consider two datasets:

- Dataset A: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
- Dataset B: [1, 2, 3, 4, 5, 6, 7, 8, 9, 100]

Find the mean, median and mode of the two datasets?

8 Measures of spread

Definition 8.1 (Variance)**.** The variance is:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Definition 8.2 (Standard Deviation)**.** The standard deviation is:

$$s = \sqrt{s^2}$$

8.1 Range and Interquartile Range

Definition 8.3 (Range)**.** The range is:

$$\text{Range} = \max(x_i) - \min(x_i)$$

Definition 8.4 (Interquartile Range)**.** The interquartile range (IQR) is:

$$\text{IQR} = Q_3 - Q_1$$

where Q_1 and Q_3 are the first and third quartiles.

Quartiles:

- Q_1 (25th percentile): 25% of data falls below this value
- Q_2 (50th percentile): The median
- Q_3 (75th percentile): 75% of data falls below this value

Practice question

Consider two datasets:

- Dataset A: [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
- Dataset B: [1, 2, 3, 4, 5, 6, 7, 8, 9, 100]

Find the range, IQR, variance and standard deviation of the two datasets? Which one is more robust to outliers?

10 Statistical inference and hypothesis testing

We analyze or derive conclusions about populations based on sample data. Hypothesis testing provides a systematic framework for making decisions in the presence of uncertainty.

10.1 Types of Hypotheses

Definition 10.1 (Null Hypothesis)**.** The null hypothesis (H_0) represents the default assumption or the claim being tested.

Definition 10.2 (Alternative Hypothesis)**.** The alternative hypothesis (H_1 or H_a) represents what we conclude if we find sufficient evidence against H_0.

- Significance level:
- Test-statistic:
- Critical region:
- p-value:

Practice question

According to a pharmaceutical company, their new medication lowers blood pressure by an average of 5 mmHg. Twenty patients participated in a clinical trial and the mean reduction was 3.5 mmHg with a standard deviation of 2. Test the company's claim at $\alpha = 0.05$.

- State the null and alternative hypothesis
- Calculate the test statistic
- Find the p-value
- What is the conclusion?
- Will it change if $\alpha = 0.10$