भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Mathematical Foundations for Data Science (Probability)

Cumulative Distribution Function, Discrete Random Variables, Continuous Random Variables, Multiple Random Variables, Joint CDF and its Properties, Independence of Random Variables

Karthik P. N.
Assistant Professor, Department of AI
Email: pnkarthik@ai.iith.ac.in

24 August 2024

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Cumulative Distribution Function

Cumulative Distribution Function (CDF)

Definition (Cumulative Distribution Function)

Fix a probability space $(\Omega, \mathscr{F}, \mathbb{P})$.
Given a random variable $X : \Omega \to \mathbb{R}$ with respect to $\mathscr{F}$, its cumulative distribution function (CDF) $F_X : \mathbb{R} \to [0, 1]$ is defined as

$$F_X(x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\}) = \mathbb{P}(\{X \leq x\}), \qquad x \in \mathbb{R}.$$

Remarks on notation:

- $\{\omega \in \Omega : X(\omega) \leq x\} = \{X \leq x\}$
- $\mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\}) = \mathbb{P}(X \leq x)$

Properties of CDF

Fix a probability space $(\Omega, \mathscr{F}, \mathbb{P})$
Let $X : \Omega \to \mathbb{R}$ be a random variable with respect to $\mathscr{F}$ with CDF F_X

- $\lim_{x \to -\infty} F_X(x) = 0, \qquad \lim_{x \to +\infty} F_X(x) = 0$

- (Monotonicity) If $x \leq y$, then $F_X(x) \leq F_X(y)$

- (Right-Continuity) F_X is right-continuous, i.e., for all $x \in \mathbb{R}$,

$$\lim_{\varepsilon \downarrow 0} F_X(x + \varepsilon) = F_X(x).$$

Properties of CDF

Fix a probability space $(\Omega, \mathscr{F}, \mathbb{P})$
Let $X : \Omega \to \mathbb{R}$ be a random variable with respect to $\mathscr{F}$ with CDF F_X

- For any $x \in \mathbb{R}$,
$$\lim_{\varepsilon \downarrow 0} F_X(x - \varepsilon) = \mathbb{P}(\{X < x\}).$$
- F_X is continuous at a point $x \in \mathbb{R}$ if and only if $\mathbb{P}(\{X = x\}) = 0$

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Discrete Random Variables

Discrete Random Variables

Fix a probability space $(\Omega, \mathscr{F}, \mathbb{P})$.

Definition (Discrete Random Variable)

A random variable $X : \Omega \to \mathbb{R}$ is said to be discrete if there exists a countable set $E \subset \mathbb{R}$, say $E = \{e_1, e_2, \ldots\}$, such that $\sum_{i=1}^{\infty} \mathbb{P}(\{X = e_i\}) = 1$.

Discrete Random Variables

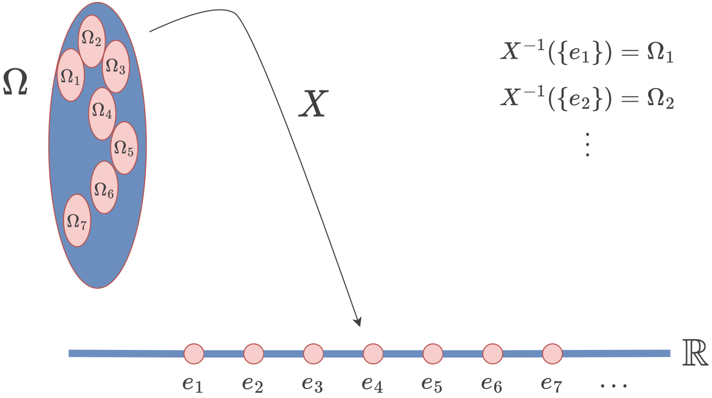
Fix a probability space $(\Omega, \mathscr{F}, \mathbb{P})$.

Definition (Discrete Random Variable)

A random variable $X : \Omega \to \mathbb{R}$ is said to be discrete if there exists a countable set $E \subset \mathbb{R}$, say $E = \{e_1, e_2, \ldots\}$, such that $\sum_{i=1}^{\infty} \mathbb{P}(\{X = e_i\}) = 1$.

$$\mathbb{P}(\{X \in E\}) = \sum_{i=1}^{\infty} \mathbb{P}(\{X = e_i\}) = 1, \qquad \mathbb{P}(\{X \in B\}) = \sum_{i:\ e_i \in B} \mathbb{P}(\{X = e_i\}), \quad B \subseteq \mathbb{R}.$$

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad



$$X^{-1}(\{e_1\}) = \Omega_1$$
$$X^{-1}(\{e_2\}) = \Omega_2$$
$$\vdots$$

$$E = \{e_1, e_2, \ldots\}$$

Probability Mass Function (PMF)

Fix a probability space $(\Omega, \mathscr{F}, \mathbb{P})$.

Definition (Probability Mass Function)

For any random variable $X : \Omega \to \mathbb{R}$, the function $p_X : \mathbb{R} \to [0, 1]$ defined as

$$p_X(x) = \mathbb{P}(\{X = x\}), \qquad x \in \mathbb{R},$$

is called the probability mass function (PMF) of X.

Probability Mass Function (PMF)

Fix a probability space $(\Omega, \mathscr{F}, \mathbb{P})$.

Definition (Probability Mass Function)

For any random variable $X : \Omega \to \mathbb{R}$, the function $p_X : \mathbb{R} \to [0, 1]$ defined as

$$p_X(x) = \mathbb{P}(\{X = x\}), \qquad x \in \mathbb{R},$$

is called the probability mass function (PMF) of X.

Remark

For a discrete random variable X taking values in the countable set $E = \{e_1, e_2, \dots\}$,

$$\sum_{i=1}^{\infty} p_X(e_i) = 1.$$

CDF in Terms of PMF

Fix a probability space $(\Omega, \mathscr{F}, \mathbb{P})$.

Definition (Probability Mass Function)

Let $X : \Omega \to \mathbb{R}$ be a discrete random variable taking values in a countable set $E = \{e_1, e_2, \ldots\} \subset \mathbb{R}$. Then,

$$F_X(x) = \sum_{i:\ e_i \leq x} \mathbb{P}(\{X = e_i\}) = \sum_{i:\ e_i \leq x} p_X(e_i), \qquad x \in \mathbb{R}.$$

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Examples of Discrete Random Variables

Definition (Discrete Random Variable)

A random variable $X : \Omega \to \mathbb{R}$ is said to be discrete if there exists a countable set $E \subset \mathbb{R}$, say $E = \{e_1, e_2, \ldots\}$, such that $\sum_{i=1}^{\infty} \mathbb{P}(\{X = e_i\}) = 1$.

- $X \sim \text{Bernoulli}(p), \quad p \in [0, 1]$

$$E = \{0, 1\}, \qquad p_X(x) = \begin{cases} p, & x = 1, \\ 1 - p, & x = 0, \\ 0, & \text{otherwise.} \end{cases}$$

- $X \sim \text{unif}(\{1, \ldots, n\})$ for some fixed $n \in \mathbb{N}$

$$E = \{1, \ldots, n\}, \quad p_X(x) = \begin{cases} \frac{1}{n}, & x \in \{1, \ldots, n\}, \\ 0, & \text{otherwise.} \end{cases}$$

Examples of Discrete Random Variables

Definition (Discrete Random Variable)

A random variable $X : \Omega \to \mathbb{R}$ is said to be discrete if there exists a countable set $E \subset \mathbb{R}$, say $E = \{e_1, e_2, \ldots\}$, such that $\sum_{i=1}^{\infty} \mathbb{P}(\{X = e_i\}) = 1$.

- $X \sim \text{Geometric}(p), \quad p \in [0, 1]$

$$E = \mathbb{N}, \qquad p_X(x) = \begin{cases} p(1-p)^{x-1}, & x \in \mathbb{N}, \\ 0, & \text{otherwise.} \end{cases}$$

- $X \sim \text{Binomial}(n, p)$ for some fixed $n \in \mathbb{N} \cup \{0\}$ and $p \in [0, 1]$

$$E = \{0, \ldots, n\}, \quad p_X(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & x \in \{0, \ldots, n\}, \\ 0, & \text{otherwise.} \end{cases}$$

Examples of Discrete Random Variables

Definition (Discrete Random Variable)

A random variable $X : \Omega \to \mathbb{R}$ is said to be discrete if there exists a countable set $E \subset \mathbb{R}$, say $E = \{e_1, e_2, \ldots\}$, such that $\sum_{i=1}^{\infty} \mathbb{P}(\{X = e_i\}) = 1$.

- $X \sim \text{Poisson}(\lambda), \quad \lambda > 0$

  $$E = \{0, 1, 2, \ldots\}, \qquad p_X(x) = \begin{cases} e^{-\lambda} \frac{\lambda^x}{x!}, & x \in \{0, 1, 2, \ldots\}, \\ 0, & \text{otherwise.} \end{cases}$$

- $$E = \{1, 2, \ldots\}, \qquad p_X(x) = \begin{cases} \frac{6}{\pi^2} \frac{1}{x^2}, & x \in \{1, 2, \ldots\}, \\ 0, & \text{otherwise.} \end{cases}$$

Continuous Random Variables

Continuous Random Variables

Fix a probability space $(\Omega, \mathscr{F}, \mathbb{P})$.

Definition (Continuous Random Variable)

A random variable $X : \Omega \to \mathbb{R}$ is said to be continuous if there exists a function $f_X : \mathbb{R} \to [0, \infty)$ such that

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt, \qquad \forall x \in \mathbb{R}.$$

Continuous Random Variables

Definition (Continuous Random Variable)

A random variable $X : \Omega \to \mathbb{R}$ is said to be continuous if there exists a function $f_X : \mathbb{R} \to [0, \infty)$ such that

$$\mathbb{P}_X((-\infty, x]) = \int_{-\infty}^{x} f_X(t)\, dt, \qquad \forall x \in \mathbb{R}.$$

Remarks:

- If $X : \Omega \to \mathbb{R}$ is a continuous random variable, its CDF F_X is continuous
- The function f_X in the definition is called the probability density function (PDF) of the random variable X
- For a continuous random variable X, its PDF f_X provides the full probabilistic description of X

Examples

- $X \sim \text{Uniform}([0,1])$, $\quad f_X(x) = \begin{cases} 1, & x \in [0,1], \\ 0, & \text{otherwise.} \end{cases}$
- $X \sim \text{Exponential}(\lambda)$ for some fixed $\lambda > 0$, $\quad f_X(x) = \begin{cases} \lambda\, e^{-\lambda x}, & x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$
- $X \sim \text{Gaussian}(\mu, \sigma^2)$ for some fixed $\mu \in \mathbb{R}$, $\sigma > 0$,

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\,\sigma^2}\right), \quad x \in \mathbb{R}.$$

- $X \sim \text{Normal} = \text{Gaussian}(0,1)$

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \quad x \in \mathbb{R}.$$

PDF $\neq$ Probabilities

Note

A probability density function (PDF) does not have the interpretation of a probability. Only integrals of PDF have interpretation of probabilities.

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Multiple Random Variables

Joint CDF of Two Random Variables

Fix a probability space $(\Omega, \mathscr{F}, \mathbb{P})$.

Definition (Joint CDF)

Given random variables $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ with respect to $\mathscr{F}$, their joint CDF $F_{X,Y} : \mathbb{R}^2 \to [0, 1]$ is defined as

$$F_{X,Y}(x, y) = \mathbb{P}(\{X \leq x\} \cap \{Y \leq y\}), \qquad x, y \in \mathbb{R}.$$

Notation

- $\{X \leq x\} \cap \{Y \leq y\} = \{X \leq x,\ Y \leq y\}$
- $\mathbb{P}(\{X \leq x\} \cap \{Y \leq y\}) = \mathbb{P}(X \leq x, Y \leq y)$

Properties of Joint CDF

Fix a probability space $(\Omega, \mathscr{F}, \mathbb{P})$
Let $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ be random variables with respect to $\mathscr{F}$ with joint CDF $F_{X,Y}$

- $\lim_{x,y\to-\infty} F_{X,Y}(x, y) = 0, \qquad \lim_{x,y\to+\infty} F_{X,Y}(x, y) = 1$

- (Monotonicity) If $x_1 \leq x_2$ and $y_1 \leq y_2$, then $F_{X,Y}(x_1, y_1) \leq F_{X,Y}(x_2, y_2)$

- $F_{X,Y}$ is continuous from the right and top, i.e., for all $x, y \in \mathbb{R}$,

$$\lim_{u\downarrow 0,\ v\downarrow 0} F_{X,Y}(x+u,\ y+v) = F_{X,Y}(x, y).$$

- $\lim_{y\to\infty} F_{X,Y}(x, y) = F_X(x)$ for all $x \in \mathbb{R}$
  $\lim_{x\to\infty} F_{X,Y}(x, y) = F_Y(y)$ for all $y \in \mathbb{R}$

Joint CDF of More Than 2 Random Variables

Fix a probability space $(\Omega, \mathscr{F}, \mathbb{P})$.

Definition (Joint CDF of More Than 2 Random Variables)

Fix $n > 2$, and let $X_1, \ldots, X_n$ be random variables with respect to $\mathscr{F}$. The joint CDF of $X_1, \ldots, X_n$ is a function $F_{X_1,\ldots,X_n} : \mathbb{R}^n \to [0, 1]$ defined as

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathbb{P}\left(\bigcap_{i=1}^{n} \{X_1 \leq x_i\}\right), \qquad x_1, \ldots, x_n \in \mathbb{R}.$$

Independence of Random Variables

Independence of Random Variables

Fix a probability space $(\Omega, \mathscr{F}, \mathbb{P})$.

Definition (Independence of Random Variables)

1. Two random variables $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ defined with respect to $\mathscr{F}$ are said to be independent if

$$F_{X,Y}(x, y) = F_X(x) \cdot F_Y(y) \qquad \forall x, y \in \mathbb{R}.$$

2. A collection of random variables $X_1, \ldots, X_n$, all defined with respect to $\mathscr{F}$, are said to be independent if

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n), \qquad x_1, \ldots, x_n \in \mathbb{R}.$$

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Can a Random Variable be Independent of Itself?

Fix a probability space $(\Omega, \mathscr{F}, \mathbb{P})$.
Let $X : \Omega \to \mathbb{R}$ be a random variable defined with respect to $\mathscr{F}$.
Can X be independent of itself?