# RANDOM VARIABLES

**Definition** A **probability space** $(S, \mathcal{E}, P)$ is composed of a sample space $S$, the algebra $\mathcal{E}$ (see notes), and a probability function $P : \mathcal{E} \to [0, 1]$ that satisfies Kolmogorov's axioms.

**Definition** A real-valued function $X$ that maps one probability space $(S, \mathcal{E})$ to another probability space $(\Omega, \mathcal{F})$ is called a **random variable** (r.v.) if

$$X^{-1}(E) \in \mathcal{E} \text{ for all } E \in \mathcal{F}$$

That is, each event in the "new" algebra corresponds to (measureable) events in the original space. This ensures that $X$ induces a consistent probability measure on the new space.

**Definition** Suppose r.v. $X$ maps $(S, \mathcal{E}, P) \to (\Omega, \mathcal{F}, P_X)$. The probability function (measure) $P_X$ is called the **probability distribution of** $X$ and is given by

$$P_X(A) = P(\{s \in S : X(s) \in A\}) \text{ for all } A \in \mathcal{F}.$$

**NOTE:** By $X$ being real-valued, we mean that $\Omega \subseteq \mathbb{R}$ or $\Omega \subseteq \mathbb{R}^n$. In the latter case, we call $X$ a **random vector**.

**Theorem:** The distribution of X is uniquely determined by the **cumulative distribution function** (cdf) $F_X$ of X:

$$F_X(x) = P(X \leq x) = P((-\infty, x]).$$

### Properties of cdf

1. F is nondecreasing: If $x_1 \leq x_2$, then $F(x_1) \leq F(x_2)$;
2. F is right - continuous: for any x, $\lim_{y \to x^+} F(y) = F(x)$;
3. $\lim_{y \to \infty} F_Y(y) = 1$;
4. $\lim_{y \to -\infty} F_Y(y) = 0$.

**NOTE:** Here are two useful rules for computing probabilities:

1. For a sequence of *increasing* sets $A_1 \subset A_2 \subset \ldots$ the probability of their union is the limit of their probabilities, that is: $P(\bigcup_{i=1}^{\infty} A_i) = \lim_{i \to \infty} P(A_i)$.
2. For a sequence of *decreasing* sets $A_1 \supset A_2 \supset \ldots$ the probability of their intersection is the limit of their probabilities, that is: $P(\bigcap_{i=1}^{\infty} A_i) = \lim_{i \to \infty} P(A_i)$.

**Types of distributions**: There are three main types of distributions/random variables:

1. Discrete r.v.: CDF is a step function, has at most countable number of values.
2. Continuous r.v.: CDF is a continuous function, has intervals in the set of values.
3. Mixed r.v.: CDF is neither continuous nor step function.

## DISCRETE RANDOM VARIABLES (RVs)

**Definition.** Suppose a sample space S has finite or countable number of simple outcomes. Let p be a real valued function on S such that

1. $0 \le p(s) \le 1$ for every element $s$ of S;
2. $\sum_{s \in S} p(s) = 1$,

Then p is said to be a **discrete probability function**.

**NOTE:** For any event A defined on S: $P(A) = \sum_{s \in A} p(s)$.

**Definition.** A real valued function $X : S \to \mathbb{R}$ is called a **random variable**.

**Definition.** A random variable with finite or countably many values is called a **discrete random variable**.

**Definition.** Any discrete random variable X is described by its **probability density function** (or probability mass function), denoted $p_X(k)$, which provides probabilities of all values of X as follows:

$$p_X(k) = P(s \in S : X(s) = k). \tag{1}$$

**NOTE:** For any k not in the range (set of values) of X: $p_X(k) = 0$.

**NOTE:** For any $t \le s$, $P(t \le X \le s) = \sum_{k=t}^{s} P(X = k)$.

**NOTATION:** For simplicity, we denote $p_X(k) = P(X = k)$ thus suppressing the dependence on the sample space.

**Examples:**

1. Binomial random variable $X$ with n trials and probability of success equal to $p$, i.e., $X \sim \text{binom}(n, p)$.

$$p_X(k) = P(\text{k successes in n trials}) = \binom{n}{k} p^k (1-p)^{n-k}, \text{ for } k = 0, 1, 2, \ldots, n. \tag{2}$$

2. Hypergeometric random variable $X$.

**Definition.** Let X a discrete random variable. For any real number t, the cumulative distribution function F of X at t is given by

$$F_X(t) = P(X \leq t) = P(s \in S : X(s) \leq t). \tag{3}$$

**Linear transformation:** Let X be a discrete random variable (rv). let Y=aX+b, where a and b are real constants. Then $p_Y(y) = p_X(\frac{y-b}{a})$.

## CONTINUOUS RANDOM VARIABLES (RVs)

Suppose a sample space $\Omega$ is uncountable, e.g., $\Omega = [0, 1]$ or $\Omega = \mathbb{R}$. We can define a random variable $X : (\Omega, \mathcal{E}) \to (S, \mathcal{B})$ where the new sample space $S$ is a subset of $\mathbb{R}$ and the algebra $\mathcal{B}$ is the Borel sets (all unions, intersections and complements of the open and closed intervals in $S$). The probability structure on such a space can be described using a special function, $f$ called *probability density function* (pdf).

**Definition.** If sample space $S \subseteq \mathbb{R}$ then we say $P$ is a **continuous probability distribution** if there exists a function $f(t)$ such that for any closed interval $[a, b] \subset S$ we have that $P([a, b]) = \int_a^b f(t)dt$. It follows that $P(A) = \int_A f(t)dt$ for all events $A$.

For a function $f$ to be a pdf, it is necessary and sufficient that the following properties hold:

1. $f(t) \geq 0$ for every $t$;
2. $\int_{-\infty}^{\infty} f(t)dt = 1$.

**NOTE:** If $P(A) = \int_A f(t)dt$ for all A, then P satisfies all the Kolmogorov probability axioms.

**Definition:** Any function Y that maps S (a subset of real numbers) into the real numbers is called **a continuous random variable**. The pdf of Y is a function $f$ such that

$$P(a \leq Y \leq b) = \int_a^b f(t)dt.$$

For any event A defined on S: $P(A) = \int_A f(t)dt$.

**Theorem**: For any continuous random variable $P(X = a) = 0$ for any real number $a$.

**Definition.** The cdf of a continuous random variable Y (with pdf $f$) is $F_Y(t)$, given by

$$F_Y(y) = P(Y \leq y) = P(\{s \in S : Y(s) \leq y\}) = \int_{-\infty}^{y} f(t)dt \quad \text{for any real } y.$$

**Theorem.** If $F_Y(t)$ is a cdf and $f_Y(t)$ is a pdf of a continuous random variable $Y$, then

$$\frac{d}{dt}F_Y(t) = f_Y(t).$$

**Linear transformation:** Let X be a continuous random variable with pdf $f$. Let $Y = aX + b$, where a and b are real constants. Then the pdf of Y is: $g_Y(y) = \frac{1}{|a|}f_X(\frac{y-b}{a})$.

## EXPECTED VALUES OF RANDOM VARIABLES

To get an idea about the *central tendency* for a random variable, we compute its **expected value** (mean).

**Definition** Let X be a random variable.

1. If X is a discrete random variable with pdf $p_X(k)$, then the expected value of $X$ is given by
$$E(X) = \mu = \mu_X = \sum_{\text{all k}} k \cdot p_X(k) = \sum_{\text{all k}} k \cdot P(X = k)$$
2. If X is a continuous random variable with pdf $f$, then
$$EX = \mu = \mu_X = \int_{-\infty}^{\infty} xf(x)dx.$$
3. If X is a mixed random variable with cdf F, then the expected value of $X$ is given by
$$E(X) = \mu = \mu_X = \int_{-\infty}^{\infty} xF'(x)dx + \sum_{\text{all k}} k \cdot P(X = k),$$
where $F'$ is the derivative of $F$ where the derivative exists and $k$'s in the summation are the "discrete" values of X.

**NOTE:** For the expectation of a random variable to exist, we assume that all integrals and sums in the definition of the expectation above converge **absolutely**.

**Median** of a random variable - a value "dividing the distribution of X in halfs. If X is a discrete random variable, then its median $m$ is the point for which $P(X < m) = P(X > m)$. If there are two values $m$ and $m'$ such that $P(X \leq m) = 0.5$ and $P(X \geq m') = 0.5$, the median is the average of $m$ and $m'$, $(m + m')/2$.

If X is a continuous random variable with pdf $f$, the median is the solution of the equation:

$$\int_{-\infty}^{m} f(x)dx = 0.5.$$