MATH 5010 –Foundations of Statistical Theory and Probability

❖ Random Variables and Distributions

Instructor: He Wang
Department of Mathematics
Northeastern University

❖ **Outline:**

1. **Random Variables**
2. **Distribution Functions**

❖ Random Variables

We are mainly interested in some functions of the outcome as opposed to the outcome itself.

Example. Flipping an **unfair** coin **twice**.

$$S = \{\text{FF, FT, TF, TT}\}$$

Suppose we want to calculate the probability of number of heads.

$$X = \text{number of heads}$$

Random Variables (formal)

Let S be a sample space. We want to relate events to numbers.

A **Random Variable** is a function from the **sample space** to real numbers

$$X: S \to \mathbb{R}$$

Example: Toss two dice

$$X = sum\ of\ the\ numbers.$$

Example: Apply different amounts of fertilizer to corn plants

$$X = yield/acre$$

Example. Suppose that our experiment consists of seeing how long a battery can operate before wearing down. Suppose also that we are not primarily interested in the actual lifetime of the battery but are concerned only about whether or not the battery lasts at least two years. In this case, we may define the random variable $\mathbb{I}$ by

$$\mathbb{I} = \begin{cases} 1 & \text{if the lifetime of battery is two or more years} \\ 0 & \text{otherwise} \end{cases}$$

If E denotes the event that the battery lasts two or more years, then the random variable $\mathbb{I}$ is known as the **indicator random variable** for event E.

❖ Probability distributions

Two important things of random variable:

1. the set of possible values for X (**range** $Ran(X)$)
2. the **probabilities** for those values (**cdf/pmf/pdf**)

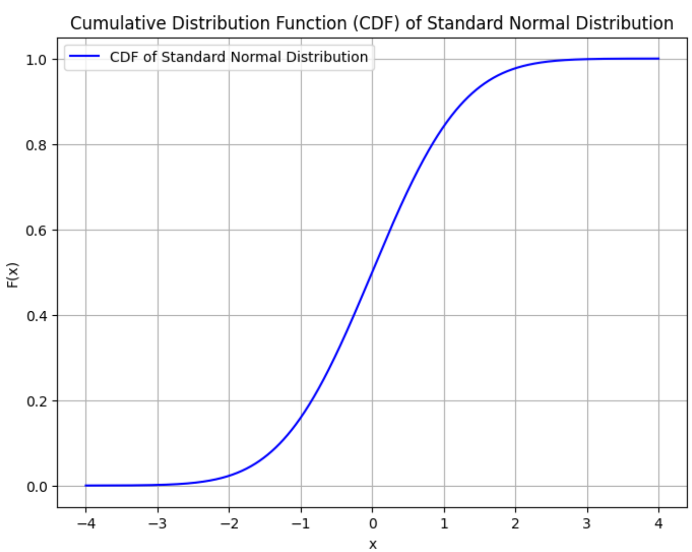
A **probability distribution** is a function that gives the probabilities of the occurrence of all possible events for an experiment.

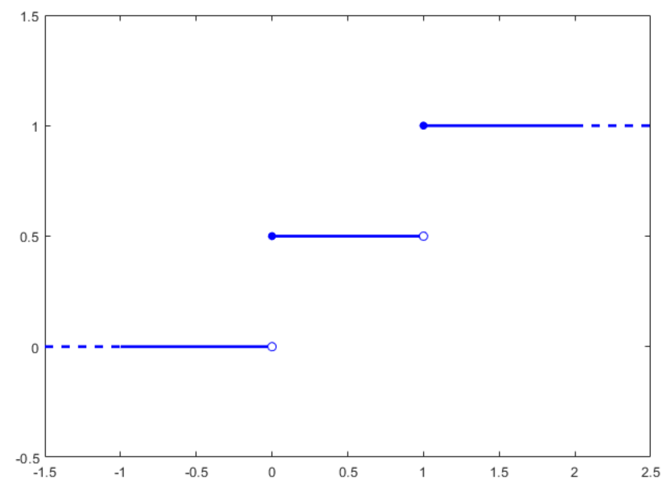
➢ Cumulative Distribution Function (CDF)

Definition: Cumulative Distribution Function (CDF) $F(x)$ is a very important characteristic of a random variable:

$$F_X(x) := P(X \le x) = P(\{s \in S : X(s) \le x\}).$$

If there is no confusion, we also denote $F_X(x) =: F(x)$.





Properties of CDF $\boldsymbol{F(x)}$:

1. $\lim_{x \to -\infty} F(x) = 0$
2. $\lim_{x \to +\infty} F(x) = 1$
3. $F(x)$ is a **nondecreasing** function.
4. $F(x)$ is right-continuous. That is $\lim_{x \downarrow x_0} F(x) = F(x_0)$

Definitions of continuous and discrete random variables:

- A random variable X is **continuous** if $F(x)$ is a continuous function of x.
- A random variable X is **discrete** if $F(x)$ is a step function of x. (Equivalently, range of X is a set of ether finite or countable number of values.)

Identical Distributions

The distribution of X is completely determined by the CDF.

Definition: The random variables X and Y on sample space S are called **identically distributed**, if for every event $A \subset S$, $P(X \in A) = P(Y \in A)$.

Equivalently, X and Y are identically distributed if and only if $F_X(x) = F_Y(x)$ for all x.

Example: Flipping a **fair** coin 2 times .

$X =$ number of heads

$Y =$ number of tails

$$F_X(x) = \begin{cases} 0 & -\infty < x < 0 \\ \frac{1}{4} & 0 \le x < 1 \\ \frac{3}{4} & 1 \le x < 2 \\ 1 & 2 \le x < \infty \end{cases}$$

Notation: The random variables X and Y has the same distribution: $X \sim Y$.

"X has a distribution given by $F_X(x)$" is denoted by $X \sim F_X(x)$.

- **Probability Density (Mass) Function**

Definition: For every discrete random variable X, the **probability mass function (pmf)** is defined as

$$p_X(x) := P(X = x) = P(\{s \in S : X(s) = x\})$$

where the index x runs over all possible values of X.

Example: Flipping a **fair** coin 2 times .

$$p_X(x) = \begin{cases} 1/4 & x = 0 \\ 2/4 & x = 1 \\ 1/4 & x = 2 \\ 0 & \text{otherwise} \end{cases}$$

- If the range of X is small then this list can be written down explicitly.
- If the range of X is large or infinite then the pdf is given by a formula.

Theorem: The probability mass function (pmf) of a discrete random variable X satisfies

1. $p_X(x) \geq 0$
2. $$\sum_{x \in \mathrm{Range}(X)} p_X(x) = 1$$

Relation with CDF:

$$p(k) = F(k) - F(k^-),$$

where $F(k^-) \coloneqq \lim_{\epsilon \to 0} F(x - \epsilon)$

Example (Bernoulli)

Random Variable X $\backsim$ $\mathbf{Bernouli}\ (\phi)$

A Bernoulli random variable takes only two values, so it is determined by a single probability.

pmf function $p_X(k) = \phi^k(1-\phi)^{1-k} = \begin{cases} \phi & if\ k = 1 \\ 1-\phi & if\ k = 0 \end{cases}$

Example 1. Flipping an un**fair** coin once.

Example: (Categorical)

Assume $Y \backsim \textbf{Categorical}(\phi_1, \dots, \phi_K)$ such that $\phi_1 + \cdots + \phi_K = 1$

Pdf function $p_Y(y) = \phi_1^{\mathbb{I}(y=1)} \phi_2^{\mathbb{I}(y=2)} \dots \phi_K^{\mathbb{I}(y=K)} = \begin{cases} \phi_1 & if\ y = 1 \\ \phi_2 & if\ y = 2 \\ \vdots & \vdots \\ \phi_K & if\ y = K \end{cases}$

Example. Rolling an unfair 6-sided die once.

Example (Categorical). Rolling an K-sided die once.

➢ Probability Density Function

Definition: In case of continuous random variable X, we may describe its distribution using the **probability density function (pdf):**

$$p(x) = F'(x) = \frac{d}{dx}\big(F(x)\big)$$

The cdf can be written as

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} p(u)du$$

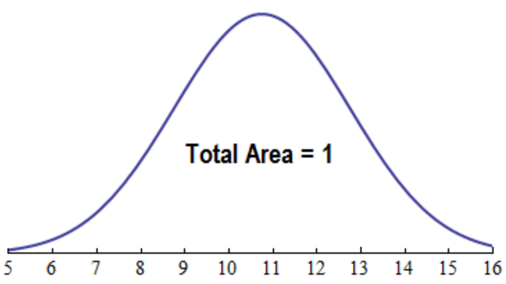
Theorem: The **Probability Density Function (pdf)** of a continuous random variable X is a piece-wise continuous function $f_X(x)$ satisfying

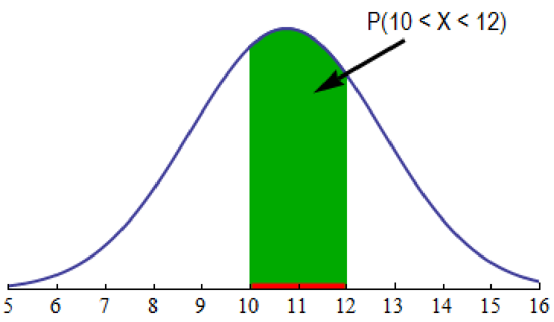
1. $f_X(x) \geq 0$

2. $\int_{-\infty}^{\infty} f(x)dx = 1$

The **probability** that X is in an interval $[a, b]$ is

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$





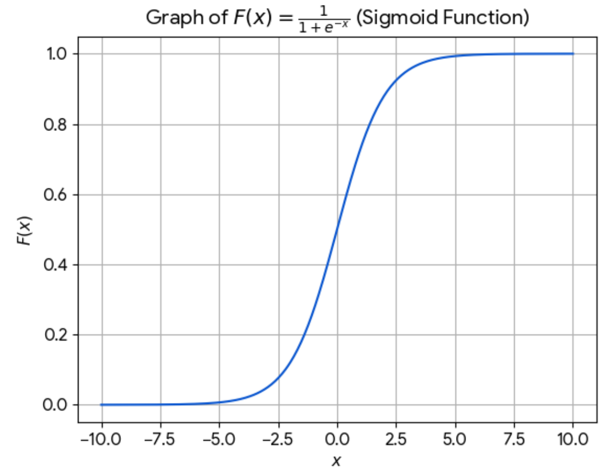
Example (logistics distribution)

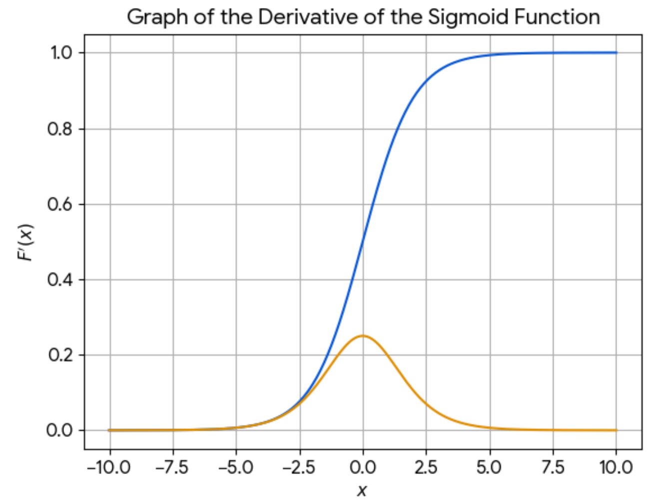
The CDF of a continuous random variable X

$$F_X(x) = \frac{1}{1 + e^{-x}}$$

The pdf of X is

$$p(x) = F'(x) = \frac{e^{-x}}{(1 + e^{-x})^2}$$



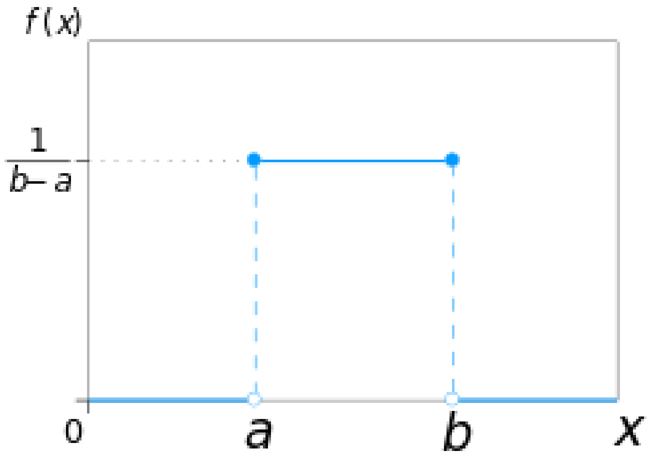


Example: **Uniform Distribution.**

The **uniform** distribution describes an experiment that choose a number randomly from the interval [a, b].

The *probability density function* of the uniform distribution is

$$f(x) = P(X = x) = \begin{cases} \dfrac{1}{b-a} & \text{for } a \le x \le b. \\ 0 & \text{for } x < a \text{ or } x > b \end{cases} \qquad \text{(Question?)}$$



Example(No pdf/pmf)

Let X be a random variable such that

- with a probability of 0.5, it always takes a fixed value 2
- with a probability of 0.5, it is from a uniform PDF over [0, 1].

We can **not** use pdf/pmf to describe this random variable.

We can use cdf to describe it:

$$F(x) = P(X \leq x) = \begin{cases} 0 & if\ x \leq 0. \\ \frac{x}{2} & if\ 0 < x \leq 1 \\ \frac{1}{2} & if\ 1 < x < 2 \\ 1 & if\ x \geq 2 \end{cases}$$

- **Continuous and Absolutely Continuous Random Variables**

Definition: A random variable X is **continuous** if $P(X = x) \; = \; 0$ for all x.

Definition: A random variable X is **absolutely continuous** if there exists a (**pdf**) function $f(x)$ such that

$$P(X \in A\,) \; = \int_A f(x)dx$$

for all Borel sets A.

- Absolutely continuous implies continuous.
- The converse is not true. (Cantor function).
- But most of the continuous random variables we met are also absolutely continuous. Our class does not distinguish between continuity and absolute continuity.

References:

- **Book 1. [CB] Statistical Inference**, by Casella, George, Berger, Roger L, 2nd edition
- **Book 2. [W]: All of Statistics: Larry Wasserman**
- **Book 3. Introduction to Probability**. C.M. Grinstead and J.L. Snell. American Mathematical Society, 2012
- **Book 4. Introduction to Probability Models**, S. Ross, 12th edition (published by Academic Press).

Online books:

https://www.probabilitycourse.com/

Extra Reading:

Baby Measure Theory: https://www.stat.umn.edu/geyer/8501/measure.pdf

YouTube video about coin flips by a famous statistician,
YouTube video about dice rolls (Part I and Part II).