LECTURE 1: PROBABILITY THEORY

MECO 7312.
INSTRUCTOR: DR. KHAI CHIONG

Probability Theory is the foundation of modern statistics. We start by introducing the mathematical concept of a *probability space*, which has three components $(\Omega, \mathcal{B}, P)$, respectively, the *sample space*, *event space*, and *probability function*.

1. Probability Space

1.1. Sample Space

Ω denotes the sample space. It is the set of possible outcomes of a particular experiment.

Examples:

(i) The experiment consists of tossing a coin. $\Omega = \{H, T\}$.

(ii) Tossing two coins. $\Omega = \{\text{“}HH\text{”}, \text{“}HT\text{”}, \text{“}TT\text{”}, \text{“}TH\text{”}\}$.

(iii) Tossing a single die. $\Omega = \{1, 2, 3, 4, 5, 6\}$.

(iv) In an A/B testing marketing experiment, the number of active users on a given day. $\Omega = \{0, 1, 2, \dots, \} = \{x \in \mathbb{Z} : x \geq 0\} = \mathbb{Z}_0^+$.

(v) In an A/B testing marketing experiment, the dollar revenue on a given day. $\Omega = \{x \in \mathbb{R} : x \geq 0\} = \mathbb{R}_0^+$.

1.2. Event

An event is a subset of Ω. We denote an event as A, as such, $A \subseteq \Omega$.

Example of an event:

(i) In the experiment of tossing two coins, $\Omega = \{HH, HT, TT, TH\}$, the event that at least one head is obtained, $A = \{HH, TH, HT\}$.

(ii) In the experiment of tossing a die where $\Omega = \{1, 2, 3, 4, 5, 6\}$, the event that a number greater than 4 is obtained, $A = \{5, 6\}$.

Date: August 21, 2024.

(iii) Obtains a number greater than 6 when rolling a die, $A = \emptyset$.[1]

(iv) Obtains a number fewer than 10 when rolling a die, $A = \{1, 2, 3, 4, 5, 6\} = \Omega$.

(v) In the experiment of observing revenue on a given day, $\Omega = \mathbb{R}_0^+$, $A = [100, 200] \subset \Omega$ is the event that the revenue is between \$100 and \$200.

If A and B are two events, then $A \cup B$ is the event that either A **or** B happens. That is, $A \cup B = \{x : x \in A \text{ or } x \in B\}$. For example, in the die rolling experiment, suppose $A = \{2, 4, 6\}$ is the event an even number is rolled, and $B = \{1, 3, 5\}$ is the event an odd number is rolled, then the event that either an even or an odd number is rolled is $A \cup B = \{1, 2, 3, 4, 5, 6\} = \Omega$.

If A and B are two events, then $A \cap B$ is the event that both events A **and** B happens. That is, $A \cap B = \{x : x \in A \text{ and } x \in B\}$. For example, suppose $A = \{2, 4, 6\}$ is the event an even number is rolled, and $B = \{5, 6\}$ is the event a number greater than 4 is rolled, then the event that the number is even and it is greater than 4 is $A \cap B = \{6\}$. Note that in the previous example, if $A = \{2, 4, 6\}$ and $B = \{1, 3, 5\}$, then $A \cap B = \emptyset$. The event that an even and an odd number is rolled is the empty set.

The complement of an event A, written A^c, is the event that A does **not** happen. It is a set that consists of all possible outcomes that are **not** in A. That is, $A^c = \{x \in \Omega : x \notin A\}$. For example, suppose $A = \{2, 4, 6\}$ is the event an even number is rolled. Then $A^c = \{1, 3, 5\}$, which is the event that an odd number is rolled.

Because events are sets, events inherit useful properties of set operations. Let Ω be the sample space, and let A, B, C be three events defined on Ω.

(i) Commutativity:
$A \cup B = B \cup A$
$A \cap B = B \cap A$.

(ii) Associativity:
$A \cup (B \cup C) = (A \cup B) \cup C$
$A \cap (B \cap C) = (A \cap B) \cap C$

(iii) Distributive Laws
$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$,
$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

[1] The empty set $\emptyset$ is a subset of any set

(iv) DeMorgan's Law.
$(A \cup B)^c = A^c \cap B^c$
$(A \cap B)^c = A^c \cup B^c$

1.3. Event Space

The event space $\mathcal{B}$ is a collection of events, or a set of subsets of Ω.

The event space contains all events of interest to the researcher. Consider the coin tossing experiment, $\Omega = \{H, T\}$. An event space of Ω is $\mathcal{B} = \{\{\emptyset\}, \{H\}, \{T\}, \{H, T\}\}$. Here, we see that $\mathcal{B}$ is the set of all possible subsets of $\Omega = \{H, T\}$.

Not all collection of events will be a valid event space. Consider the die rolling experiment $\Omega = \{1, 2, 3, 4, 5, 6\}$. Supose we define the event space as $\mathcal{B} = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}\}$. Although each element of $\mathcal{B}$ is a valid event, this event space does not contain the event "not 2", $A = \{1, 3, 4, 5, 6\}$. It also does not contain the event "either 1 or 3", $A = \{1, 3\}$. Therefore, $\mathcal{B}$ does not make sense as an event space.

What makes $\mathcal{B}$ a valid event space? An event space must satisfy three properties:

(1) $\emptyset \in \mathcal{B}$. The empty set is an element of $\mathcal{B}$.

(2) It must be closed under complementation. If $A \in \mathcal{B}$, then $A^c \in \mathcal{B}$.

(3) It must be closed under countable unions. Let $A_1, A_2, \ldots, A_n$ be events in $\mathcal{B}$, then $\cup_{i=1}^{n} A_i \in \mathcal{B}$. For instance, if A_1 and A_2 are two events in $\mathcal{B}$, then the event $A_1 \cup A_2$ must also be in $\mathcal{B}$.

A collection of events (a set of subsets of Ω) that satisfy all of the above properties is called a σ-algebra on Ω. A valid event space is a σ-algebra of Ω. Further note that properties (2) and (3) imply that if $A_1, A_2 \in \mathcal{B}$, then $A_1 \cap A_2 \in \mathcal{B}$. To see this: since $A_1, A_2 \in \mathcal{B}$, we have $A_1^c, A_2^c \in \mathcal{B}$ by property (2). By property (3), we have $(A_1^c \cup A_2^c) \in \mathcal{B}$. Finally, by property (2) again, $(A_1^c \cup A_2^c)^c \in \mathcal{B}$. From DeMorgan's Law, $(A_1^c \cup A_2^c)^c = A_1 \cap A_2$. Therefore, $A_1 \cap A_2 \in \mathcal{B}$. By induction, a σ-algebra is closed under countable intersection.

It is easy to construct a σ-algebra on Ω when Ω is countable and finite: simply define $\mathcal{B} = \{\text{all subsets of } \Omega, \text{ including } \Omega\}$. In terms of notation, $\mathcal{B} = \mathcal{P}(\Omega)$, where $\mathcal{P}$ denote the Power Set operation. This construction ensures that $\mathcal{B}$ is a valid event space (a σ-algebra), that is of interest ot us. The trivial σ-algebra $\{\emptyset, \Omega\}$ is not of interest. If $S = \{1, 2, 3\}$, then the event space (or power set) is a collection of $2^3 = 8$ sets. $\mathcal{B} = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$.

When Ω is large, it is not practical to explicitly write out the event space. Even for the die rolling experiment, the event space will be very large and has $2^6 = 64$ elements. When Ω is countably infinite, or uncountable, then we cannot enumerate the event space.[2]

1.4. Probability Function

Finally, a probability function P, assigns a number (a probability) to each event in the event space. It is a function mapping $\mathcal{B} \to [0, 1]$ satisfying:

(i) $P(A) \geq 0$, for all $A \in \mathcal{B}$.

(ii) $P(\Omega) = 1$

(iii) Countable additivity: If $A_1, A_2, \cdots \in \mathcal{B}$ are pairwise disjoint, then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Two events A and B are disjoint (or mutually exclusive) if $A \cap B = \emptyset$. The events $A_1, \ldots, A_n$ are pairwise disjoint if $A_i \cap A_j = \emptyset$ for all $i \neq j$. These are called the Axioms of Probability, or Kolmogorov Axioms.

Example: Consider rolling a die. Assuming that the die is fair, then the probability function for each event in $\mathcal{B}$ looks like:

Event $A \in \mathcal{B}$	$P(A)$
{1}	1/6
{2}	1/6
⋮	⋮
{1,2}	1/3
{1,3,6}	1/2
{2,3,4,6}	2/3
{2,3,4,5,6}	5/6
⋮	⋮
$\emptyset$	0
{1,2,3,4,5,6}	1

[2] An event space for when Ω is an uncountably large set can be defined as follows. Suppose Ω is an interval of the real line, say $\Omega = [0, 1] \subset \mathbb{R}$. At the very least, $\mathcal{B}$ should contain all sets of the form $[a, b], (a, b), (a, b], [a, b)$ for all real numbers $0 \leq a \leq 1$ and $0 \leq b \leq 1$. Why? Because we need to make sense of an event such as: the outcome is between a and b, not inclusive of a. Further, to ensure that $\mathcal{B}$ is a σ-algebra, we need to include all (possibly infinite) unions and intersections of any intervals of the form $[a, b], (a, b), (a, b], [a, b)$.

1.5. Additional properties of the probability function

These properties can be derived from the Axioms of Probability. Venn diagrams are useful to visualize these properties.

(i) $P(\emptyset) = 0$.

(ii) $P(A) \leq 1$

(iii) $P(A^c) = 1 - P(A)$

(iv) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

(v) $P(A \cup B) \leq P(A) + P(B)$. More generally, we have the Boole's inequality: $P(\cup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i)$. Proof follows from (iv).

(vi) $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$. The Inclusion-Exclusion principle generalizes this the union of many events:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) + \cdots + (-1)^{n-1} \sum_{i<\ldots<n} \mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right)$$

(vii) Bonferroni's Inequality. $P(A \cap B) \geq P(A) + P(B) - 1$. Example: suppose that 80% of you like cats, and 90% of you like dogs, then it must be that at least 70% of you like *both* cats *and* dogs ($0.8 + 0.9 - 1 = 0.7$). Proof follows from (iv).

(viii) Frechet bounds:

For intersection of events:

$$\begin{aligned} \max(0, P(A_1) + P(A_1) + \cdots + P(A_n) - (n-1)) \leq \\ P(A_1 \cap A_2 \cap \ldots \cap A_n) \leq \\ \min(P(A_1), P(A_2), \ldots, P(A_n)) \end{aligned}$$

For union of events:

$$\begin{aligned} \max(P(A_1), P(A_2), \ldots, P(A_n)) \leq \\ P(A_1 \cup A_2 \cup \ldots \cup A_n) \leq \\ \min(1, P(A_1) + P(A_2) + \cdots + P(A_n)) \end{aligned}$$

Using these bounds, we can say that the percentage of students who like both *both* cats *and* dogs is between 70% and 80%. Further, the percentage of students who like either cats *or* dogs is between 90% and 100%.

1.6. Updating information: conditional probability

Consider two events $A, B \in \mathcal{B}$, the probability of event A given event B, denoted $P(A|B)$ is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1}$$

Examples:

(1) What is the probability of drawing two Kings consecutively from a well-shuffled deck of cards? Event A is drawing a King first, and Event B is drawing a King second. $P(A) = \frac{4}{52}$. $P(B|A) = \frac{3}{51}$. Hence, $P(A \cap B) = P(B|A) \times P(A) = \frac{1}{221}$. So the chance of getting 2 Kings is 1 in 221, or about 0.5%.

(2) 25% of the class belong to Marketing, and 15% uses Mac OS *and* belong to Marketing. Randomly pick someone in the class, if I told you that person is in Marketing, what is the probability that the person uses Mac OS? $P(\text{Mac}|\text{Marketing}) = \frac{P(\text{Mac and Marketing})}{P(\text{Marketing})} = 0.15/0.25 = 0.6$.

From the definition in Equation 1 above, we see that $P(A|A) = 1$. Further, if A and B are disjoint, then $P(A \cap B) = 0$ and $P(A|B) = P(B|A) = 0$.

From the definition of conditional probability, we can derive an important result called the Bayes' Rule. Let A and B be two events, then we can "flip" the conditional probabilities via the formula $P(A|B) = P(B|A)\frac{P(A)}{P(B)}$.

From the conditional probability of event B given A:

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

$$P(A \cap B) = P(B|A) \times P(A) \tag{2}$$

From the conditional probability of event A given B:

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \quad \text{now we plug in Equation 2,} \\ &= \frac{P(B|A) \times P(A)}{P(B)} \quad \text{which gives us the Bayes' Rule.} \end{aligned}$$

2. Independence

Two events $A, B \in \mathcal{B}$ are statistically independent if and only if

$$P(A \cap B) = P(A) \times P(B) \tag{3}$$

Equivalently, two events $A, B \in \mathcal{B}$ are statistically independent if and only if

$$P(A|B) = P(A) \tag{4}$$

Or equivalently,

$$P(B|A) = P(B) \tag{5}$$

Example (2-coin toss):

(1) Let A be the event that "first coin toss is heads", $A = \{HH, HT\}$. Let B be the event that "second coin toss is heads", $B = \{HH, TH\}$. Are A and B independent? Yes. $P(A \cap B) = P(\{HH\}) = 0.25$ which is equal to $P(A) \times P(B) = 0.5 \times 0.5 = 0.25$

(2) Let A be the event that "first coin toss is heads", $A = \{HH, HT\}$. Let B be the event that "at least one tail", $B = \{TT, TH, HT\}$. Are A and B independent? No. $P(A \cap B) = P(\{HT\}) = 0.25$ which is not equal to $P(A) \times P(B) = 0.5 \times 0.75 = 0.375$.

Example (rolling a die):

(1) What is the probability of getting at least 1 six in 4 rolls of a die?

$$\begin{aligned} P(\text{at least 1 six in 4 rolls}) &= 1 - P(\text{no six in 4 rolls}) \\ &= 1 - \prod_{i=1}^{4} P(\text{no six on roll } i) \\ &= 1 - \left(\frac{5}{6}\right)^4 \\ &= 0.518 \end{aligned}$$

For a collection of n events $A_1, A_2, \ldots, A_n$, the events are mutually independent if and only if, for any subcollection $A_{i_1}, A_{i_2}, \ldots, A_{i_k}$, the following holds:

$$P\left(\bigcap_{j=1}^{k} A_{i_j}\right) = \prod_{j=1}^{k} P(A_{i_j})$$

For any possible combination of events from the collection, the probability of all those events happening together is equal to the product of their individual probabilities.

To be mutually independent, the events must satisfy a more stringent condition than pairwise independence. In pairwise independence, each pair of events is independent of each other, but this does not necessarily mean the whole collection is mutually independent.

For example, let's consider three events A, B, and C such that $P(A \cap B) = P(A)P(B)$, $P(A \cap C) = P(A)P(C)$, and $P(B \cap C) = P(B)P(C)$. These events are pairwise independent. However, they are mutually independent only if they also satisfy $P(A \cap B \cap C) = P(A)P(B)P(C)$.

3. Random variables

Let $(\Omega, \mathcal{B}, P)$ be a probability space. A random variable X is a (measurable) function $X : \Omega \to E$, where $E \subseteq \mathbb{R}$. For example, E can be the set of real numbers or the set of integers.[3]

In many cases, we want to transform the original probability space into another probability space that is more convenient for analysis. For example, consider the experiment where we record the action of a user upon seeing an online advertisement. $\Omega = \{\text{“Click”}, \text{“Skip”}\}$. Then define the random variable X that maps from Ω to real numbers such that:

Ω	X
“Skip”	0
“Click”	1

TABLE 1. The random variable X maps the sample space $\{\text{“Click”}, \text{“Skip”}\}$ to $\{0, 1\}$

[3] A measurable function is a valid function mapping from one measurable space to another. A measurable space consists of a set and a σ-algebra. Let (X, Σ) and (Y, T) be two measurable spaces. A function $f : X \to Y$ is said to be measurable if for every $E \in T$, the preimage of E under f is in Σ .

Mapping the sample space from {"Click", "Skip"} to $\{0, 1\}$ allows us to run regressions, and conduct statistical inference.

We often define a random variable through a query about the experiment, for example, revisiting the experiment of tossing two fair coins, define X to be the number of heads obtains. Then, the random variable X realizes the value 0 with an induced probability of 1/4; realizes the value 1 with an induced probability of 1/2, and so on, according to Table 2.

Ω	$P(\cdot)$	X
HH	1/4	2
HT	1/4	1
TH	1/4	1
TT	1/4	0

TABLE 2. Random variable X defined as a mapping from possible outcomes in Ω to real numbers.

In practice, we need not worry about measurability. If the query about the experiment is defined in a reasonable and sensible way, then it can be supported as a random variable. We will work with random variables for the rest of this class. Although not explicitly stated, bear in mind that behind every random variable is an underlying probability space of some experiment, which gives rise to the probabilistic nature of this random variable. In particular, we will mostly deal with continuous random variable that has the probability space $(\mathbb{R}, \mathcal{B}, P_X)$, where $\mathcal{B}$ is the Borel σ-algebra of $\mathbb{R}$.

In terms of notation, random variables will be denoted with uppercase letters (such as X), and the realized values of the variable will be denoted by the corresponding lowercase letters (such as x).

4. Cumulative Density Functions (CDF)

With every random variable X, we associate a function called the *cumulative distribution function* (cdf) of X. The cdf of a random variable X is denoted by $F_X(x)$, and is defined by:

$$F_X(x) \equiv P_X(X \leq x) \text{ for all } x \in \mathbb{R} \tag{6}$$

Consider the experiment tossing two fair coins and let X = number of heads observed. To see what the cdf of X looks like, try to evaluate $F_X(x)$ at different values, say at $x = -1, 0, 0.5, 1, 2, 3$, as follows (and using Table 2 for reference):

$$\begin{aligned} F_X(-1) &= P_X(X \leq -1) = 0 \\ F_X(0) &= P_X(X \leq 0) = \frac{1}{4} \\ F_X(1) &= P_X(X \leq 1) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4} \\ F_X(2) &= P_X(X \leq 2) = \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1 \end{aligned}$$

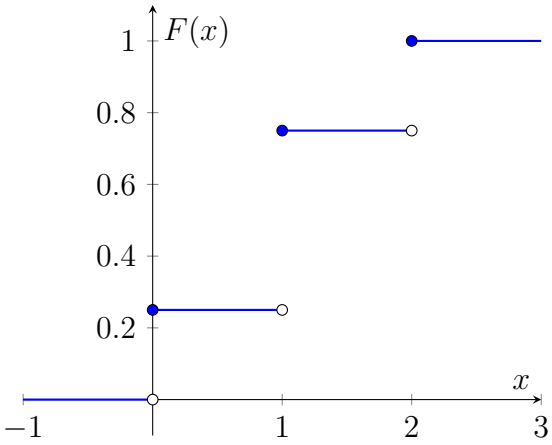

FIGURE 1. CDF of a discrete random variable

The complete cdf is:

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ \frac{1}{4} & \text{if } 0 \leq x < 1, \\ \frac{3}{4} & \text{if } 1 \leq x < 2, \\ 1 & \text{if } x \geq 2. \end{cases} \tag{7}$$

4.1. Properties of a cdf

The function $F(x)$ is a cdf if and only if the following three conditions hold:

(i) $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.

(ii) $F(x)$ is non-decreasing. If $a > b$, then $F(a) \geq F(b)$.

(iii) $F(x)$ is right-continuous (continuous when approached from the right).[4]

A random variable X is continuous if $F_X(x)$ is a continuous function of x. A random variable X is discrete if $F_X(x)$ is a step function of x.

The cdf completely determines the probability distribution of a random variable. The random variable X and Y are identically distributed if and only if $F_X(x) = F_Y(x)$ for every x.

Let X be a continuous random variable with the following cdf, $F_X(x) = \frac{1}{1+e^{-x}}$. Is $F_X(x)$ a valid cdf? First, check that $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$, as $\lim_{x\to-\infty} e^{-x} = \infty$ and $\lim_{x\to\infty} e^{-x} = 0$. Then check that $\frac{d}{dx}F_X(x) = \frac{e^{-x}}{(1+e^{-x})^2} \geq 0$. In fact, $F_X(x) = \frac{1}{1+e^{-x}}$ is a well-known distribution called the *logistic distribution*. Plotting the cdf in Mathematica using: `Plot[1/(1 + Exp[-x]), {x, -10, 10}]`

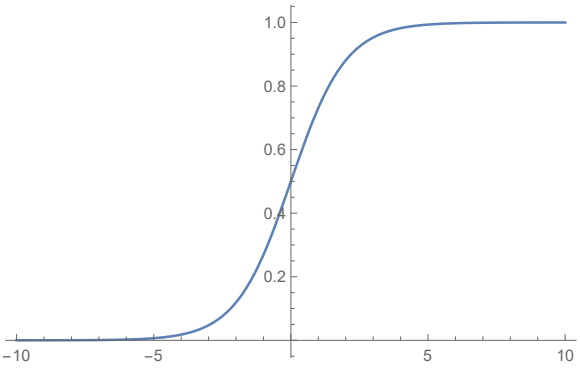

FIGURE 2. The cdf of the Logistic Distribution.

5. Probability Density Functions (pdf) and Probability Mass Functions (pmf)

Associated with a random variable X and its cdf F_X is another function, called either the pdf (for continuous random variable) or the pmf (for discrete random variable).

The pmf of a discrete random variable X is given by $f_X(x) = P_X(X = x)$. For the experiment where we toss two fair coins, the pmf is given by Equation 8 below.

[4]That is, $\lim_{x\to c^+} F_X(x) = F_X(c)$ for all c.

$$f_X(x) = \begin{cases} \frac{1}{4} & \text{if } x = 0, 2 \\ \frac{1}{2} & \text{if } x = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

The pmf is related to the cdf. The cdf of X at x, $P_X(X \leq x)$ equals to the sum of $P_X(X = k)$ at all k smaller than x. Hence, the cdf can be interpreted as the sum of point probabilities (pmfs).

For continuous random variable, the pdf is defined differently, where we now substitute integrals for sums.

$$P_X(X \leq x) = F_X(x) = \int_{-\infty}^{x} f_X(t)dt \tag{9}$$

Hence, the probability density function (pdf) for a continuous random variable X is a function $f_X(x)$ such that

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt \text{ for all } x \tag{10}$$

Using the Fundamental Theorem of Calculus, for a continuous random variable we have the relationship:

$$\frac{d}{dx}F_X(x) = f_X(x) \tag{11}$$

Pdf is a useful object because from it we can derive cdf and compute probability such as $P(a \leq X \leq b) = \int_a^b f_X(t)dt$. Note that for a continuous random variable X, $P_X(X = x)$ is always zero, as a result of $F_X(x)$ being continuous if X is a continuous random variable.[5]

Example. Let X be a random variable with the cdf $F_X(x) = \frac{1}{1+e^{-x}}$. Recall that F_X is the well-known Logistic Distribution. The pdf of X is given by $f_X(x) = \frac{d}{dx}F_X(x) = \frac{e^{-x}}{(1+e^{-x})^2}$. Using Mathematica: `Plot[Exp[-x]/(1 + Exp[-x])^2, {x, -10, 10}]`.

[5] $X = x$ is a measure-zero event, i.e. the probability measure assigns zero to the event $X = x$. A continuous random variable can realize infinitely many outcomes (an interval of real numbers is uncountably infinite). $P(X = x) = \lim_{\delta \to 0}(P(X \leq x+\delta) - P(X \leq x)) = \lim_{\delta \to 0}(F(x+\delta) - F(x)) = 0$

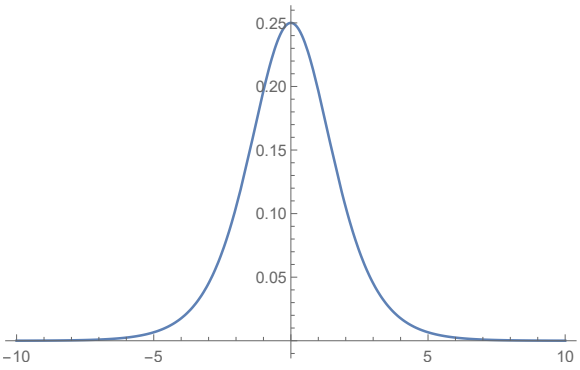

FIGURE 3. The pdf of the logistic distribution.

A function $f_X(x)$ is considered a valid probability density function (pdf) if it satisfies the following conditions:

(i) Non-Negative: The function must be non-negative for all x in the domain.

$$f_X(x) \geq 0 \quad \text{for all } x$$

(ii) Normalization: The area under the curve of the pdf must equal 1. For continuous random variables, this is expressed as:

$$\int_{-\infty}^{\infty} f_X(x)\, dx = 1$$

For discrete random variables, this becomes:

$$\sum_{\text{all } x} f_X(x) = 1$$