

LECTURE 1: PROBABILITY THEORY

MECO 7312.

INSTRUCTOR: DR. KHAI CHIONG

Probability Theory is the foundation of modern statistics, it dates back to the 1600s, to Fermat and Pascal. We start by introducing the mathematical concept of a *probability space*, which has three components (Ω, \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. $\Omega = \{“HH”, “HT”, “TT”, “TH”\}$.
- (iii) Tossing a single die. $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- (iv) The experiment consists of observing the amount spent by a particular customer, $\Omega = [0, 10000] \subset \mathbb{R}$.
- (v) The experiment consists of recording the action of a user upon seeing a mobile ad. $\Omega = \{“Click”, “Skip”\}$.
- (vi) The experiment consists of observing the profit of a given firm, $\Omega = \mathbb{R}$.
- (vii) U.S. Census survey of household incomes. $\Omega = [0, 10^9] \subset \mathbb{R}$.

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, $A = \{5, 6\}$.
- (iii) Obtains a number greater than 6 when rolling a die, $A = \{\emptyset\}$.
- (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 the amount spent by a customer, $\Omega = [0, 10000]$, the event that the customer qualifies for the Silver Rewards Program, $A = (100, 250]$.
- (vi) In the experiment of observing the profit of a given firm $\Omega = \mathbb{R}$, the event that the firm makes a positive profit, $A = (0, \infty)$.
- (vii) In the experiment of observing US household incomes, the event that a household is classified as under poverty, $A = [0, \$12140]$.

If A and B are two events, then $A \cup B$ is the event that either A **or** B happens. For example, $A = \{5\} \cup \{6\} = \{5, 6\}$ is the event that rolling the die gives 5 **or** 6. Event $A = \{HH, TH, HT\} = \{HH\} \cup \{TH\} \cup \{HT\}$ is the event that the outcome is either HH **or** TH **or** HT .

$A \cap B$ is the event that both event A **and** B happens. When rolling a die, the intersection of the two events $\{1, 2, 3\}$ and $\{3, 4, 5\}$ is $\{1, 2, 3\} \cap \{3, 4, 5\} = \{3\}$. However $\{5\} \cap \{6\} = \emptyset$, that is, a die cannot turn up to be both 5 **and** 6 at the same time.

1.3. Event Space

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

Example of an Event Space:

- (i) For rolling a die, $\mathcal{B} = \{\{1, 2\}, \{1, 4, 5\}, \{2\}, \dots\}$
- (ii) For tossing a coin, $\mathcal{B} = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$.
- (iii) For tossing two coins, $\mathcal{B} = \{\emptyset, \{HH\}, \{TH\}, \{HH, TH\}, \{HH, TH, HT\}, \dots\}$.
- (iv) For the customer's spending, $\mathcal{B} = \{[12.5, 19.7], [0, 5], (56, 99), \dots\}$

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 $A^c \equiv \Omega \setminus A \in \mathcal{B}$.
- (3) It must be closed under countable unions. If $A_1, A_2, \dots, \in \mathcal{B}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$. 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} .

Any set that satisfies the above properties is a σ -algebra of Ω . The event space \mathcal{B} must be a σ -algebra of Ω . In summary, **a σ -algebra on a set Ω is a collection of subsets of Ω that includes Ω itself, is closed under complement, and is closed under countable unions.** Intuitively, in the die tossing example, we need to make sense of the event “not 2”, “either 1 or 3”, etc. Further note that properties (2) and (3) imply that if $A_1, A_2 \in \mathcal{B}$, then $A_1 \cap A_2 \in \mathcal{B}$. By De Morgan’s Law: $(A \cup B)^c = A^c \cap B^c$, hence, $(A_1^c \cup A_2^c)^c = A_1 \cap A_2 \in \mathcal{B}$. By induction, a σ -algebra is closed under countable intersection.

It is easy to construct a valid event space when the sample space 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 of us. The trivial σ -algebra $\{\emptyset, \Omega\}$ is not of interest.

It is more tricky when Ω is uncountable. For instance, 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 events such as $[a, b] \cup [c, d]$. It will be impossible to explicitly write down the σ -algebra of $[0, 1]$, but we can define it generatively. However, for the purpose of this course, we will simply state: let \mathcal{B} be the smallest σ -algebra that contains all open sets from $[0, 1]$. This σ -algebra is given the name: Borel-algebra of $[0, 1]$. One can show that this σ -algebra will contain events of interest such as $[a, b]$ or (a, b) .

In practice, we do not need to explicitly construct the σ -algebra, we can just say: *let Ω be the sample space of interest, and let \mathcal{B} be a sigma-algebra of Ω ...*

FIGURE 1. Lars Peter Hansen's GMM estimator in 1982 Econometrica.

2. CONSISTENCY OF THE GMM ESTIMATOR

In this section we specify our first form of the GMM estimator and provide some sufficient conditions that insure its almost sure convergence to the parameter vector that is being estimated. Let Ω denote the set of sample points in the underlying probability space used in our estimation problem, and let E denote the associated expectations operator. We will be working with a p component stochastic process $\{x_n; n \geq 1\}$ defined on this probability space. A finite segment of one realization of this process, i.e., $\{x_n(\omega_0): 1 \leq n \leq N\}$ for sample size N and for some $\omega_0 \in \Omega$, can be thought of as the observable data series that the econometrician employs.

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 $\mathcal{B} \rightarrow [0, 1]$ satisfying:

- (i) $P(A) \geq 0$, for all $A \in \mathcal{B}$.
- (ii) $P(\Omega) = 1$
- (iii) Countable additivity: If $A_1, A_2, \dots \in \mathcal{B}$ are pairwise disjoint, then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

The sets are disjoint if $A \cap B = \emptyset$. 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
\vdots	\vdots
{1,2}	1/3
{1,3,6}	1/2
{2,3,4,6}	2/3
{2,3,4,5,6}	5/6
\vdots	\vdots
\emptyset	0
{1,2,3,4,5,6}	1

1.5. Additional properties of the probability function

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:
 $P(\cup_{i=1}^n A_i) \leq \sum_{i=1}^n P(A_i)$

(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:

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

(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 ($0.8 + 0.9 - 1 = 0.7$).

(viii) Frechet bounds:

For intersection of events:

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

The percentage of students who like *both* cats *and* dogs is between 70% and 80%. For union of events:

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

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

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

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? $P(\text{Mac}|\text{Marketing}) = \frac{P(\text{Mac and Marketing})}{P(\text{Marketing})} = 0.15/0.25 = 0.6$.

(3) Monty Hall problem. Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 2, which has a goat. He then says to you, "Do you want to pick door No. 3?" Is it to your advantage to switch your choice?

More formally, define two random variables D (for door behind which the prize is) and M (denoted the door which Monty opens). Consider a comparison of the conditional probabilities $P(D = 1|M = 2)$ vs. $P(D = 3|M = 2)$. Note that these two sum to 1, so you will switch to $D = 3$ if $P(D = 3|M = 2) > 0.5$.

D	M	Probability
1	1	0
1	2	$\frac{1}{3} \times \frac{1}{2}$
1	3	$\frac{1}{3} \times \frac{1}{2}$
2	1	0
2	2	0
2	3	$\frac{1}{3} \times 1$
3	1	0
3	2	$\frac{1}{3} \times 1$
3	3	0

$$\begin{aligned}
 P(D = 3|M = 2) &= \frac{P(D = 3, M = 2)}{P(M = 2)} \\
 &= \frac{P(D = 3, M = 2)}{\sum_i^3 P(D = i, M = 2)} \\
 &= \frac{1/3}{1/6 + 1/3} \\
 &= \frac{2}{3}
 \end{aligned}$$

2. Independence

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

$$(2) \quad P(A \cap B) = P(A) \times P(B)$$

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?

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

3. Random variables

In many cases, we want to transform the original probability space into another probability space that is more convenient for analysis.

Let (Ω, \mathcal{B}, P) be a probability space. A random variable X is a (measurable) function $X : \Omega \rightarrow E$, where $E \subseteq \mathbb{R}$. For example E can be the set of real numbers or the set of integers.¹

For example, define X to be the number of heads obtains when tossing two fair coins. Is X is a random variable? What is the probability space of X ? The random variable X has the sample space $\{0, 1, 2\}$ with the probability function P_X (see Table 2).

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

TABLE 1. The probability space of tossing two fair coins.

X	$P_X(\cdot)$
0	1/4
1	1/2
2	1/4

TABLE 2. Probability space of the random variable X , defined as the number of heads in tossing two fair coins.

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.

¹A 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 $f : X \rightarrow Y$ is said to be measurable if for every $E \in T$, the preimage of E under f is in Σ .

Although not explicitly stated, just 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.

Now consider the experiment where we record the action of a user upon seeing a mobile ad. $\Omega = \{\text{“Click”}, \text{“Skip”}\}$. Then define the random variable X that maps from Ω to real numbers such that:

Ω	X
“Skip”	0
“Click”	1

TABLE 3. The random variable X maps the sample space $\{\text{“Click”}, \text{“Skip”}\}$ to $\{0, 1\}$

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

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:

$$(3) \quad F_X(x) \equiv P_X(X \leq x) \text{ for all } x$$

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):

$$F_X(-1) = P_X(X \leq -1) = 0$$

$$F_X(0) = P_X(X \leq 0) = \frac{1}{4}$$

$$F_X(0.5) = P_X(X \leq 0.5) = \frac{1}{4}$$

$$F_X(1) = P_X(X \leq 1) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}$$

$$F_X(2) = P_X(X \leq 2) = \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1$$

$$F_X(3) = P_X(X \leq 3) = 1$$

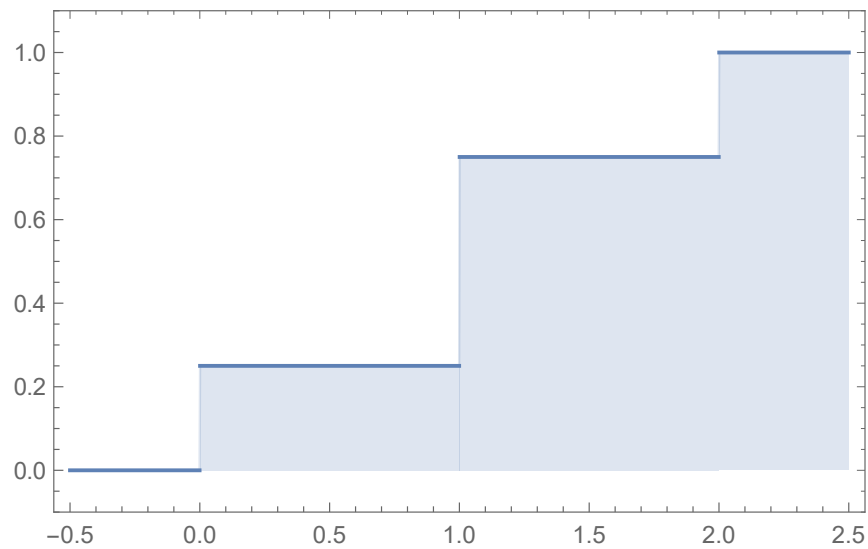


FIGURE 2. CDF of a discrete random variable

4.1. Properties of a cdf

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

- (i) $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow \infty} F(x) = 1$.
- (ii) $F(x)$ is non-decreasing.
- (iii) $F(x)$ is right-continuous (continuous when approached from the right).²

²That is, $\lim_{x \rightarrow c^+} F_X(x) = F_X(c)$

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 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 $\lim_{x \rightarrow -\infty} F_X(x) = 0$ and $\lim_{x \rightarrow \infty} F_X(x) = 1$. 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*. Plot this cdf using the following Mathematica command: `Plot[1/(1 + Exp[-x]), {x, -10, 10}]`

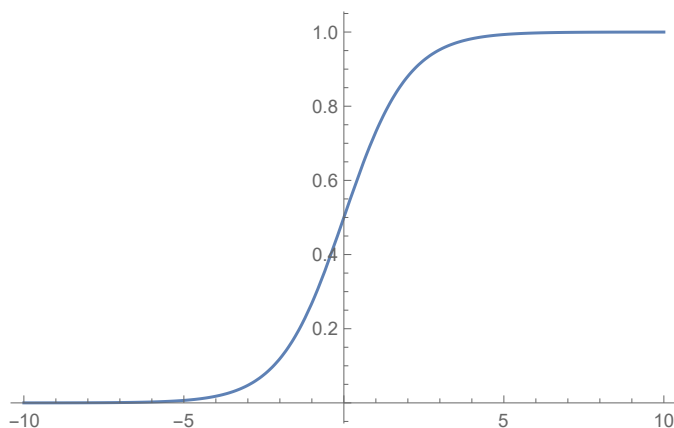


FIGURE 3. The cdf of the Logistic Distribution.

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 trivially given by Table 2. For instance, $P_X(X = 2) = \frac{1}{4}$.

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.

$$(4) \quad P_X(X \leq x) = F_X(x) = \int_{-\infty}^x f_X(t) dt$$

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

$$(5) \quad F_X(x) = \int_{-\infty}^x f_X(t) dt \text{ for all } x$$

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

$$(6) \quad \frac{d}{dx} F_X(x) = f_X(x)$$

Note that for a continuous random variable X , $P_X(X = x)$ is always zero.

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}]`.

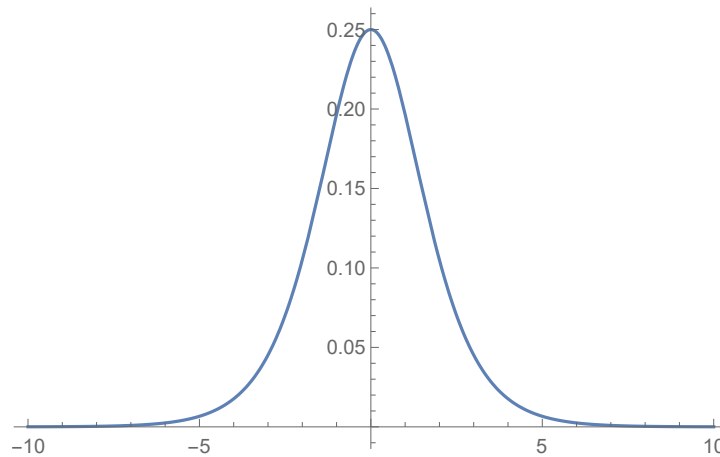


FIGURE 4. The pdf of the logistic distribution.