Introduction to Probability
(Lectures 1-2)

1 Probability Spaces and Random Variables

The goal of Lecture 1 and Lecture 2 is twofold. First, we will present the formal definitions of a *probability space* and a *random variable*. Then, we will introduce some more concrete concepts that will be used throughout the course (such as *cumulative distribution function* and *probability density function*).

Overview of Lectures 1-2: The triplet:

$$(\Omega, \mathcal{F}, \mathbb{P}),$$

will be our notation to describe a probability space. We will call

1. Ω: The underlying space of uncertainty or set of states of the world.

2. $\mathcal{F}$: The "σ-algebra" of subsets of Ω. We will think about it is as a collection of "events" whose likelihood of occurrence we would like to analyze.

3. $\mathbb{P}$: The probability measure. We will think about it as a $[0,1]$-valued function summarizing the likelihood of an event.

A real-valued random variable will be denoted by the map

$$X : \Omega \to \mathbb{R},$$

which will satisfy a restriction that we will call "measurability". We will focus on how the probability measure on $(\Omega, \mathcal{F})$ induces a probability over subsets of $\mathbb{R}$.

Once we are done with abstract definitions, we will introduce objects that we will use quite often during this course. In the context of a real-valued random variable we will define

a) The cumulative distribution function (c.d.f.): which we will connect to the "induced" probability measure of a random variable,

b) "Discrete" and "Absolutely Continuous" random variables: this classification will emerge from differences in the c.d.f. of random variables.

c) The probability density function (p.d.f.): a convenient way of summarizing the information in the c.d.f. of an absolutely continuous random variable,

d) The expectation operator $\mathbb{E}_{\mathbb{P}}[\cdot]$ (and then the variance of a real-valued random variable).

e) Finally, the moment generating function.

1.1 Measurable spaces, Probability Measures, and Probability Spaces

1.1.1 σ-algebras and Measurable Spaces

We will start with the definition of a "measurable space". In order to model randomness, we need to have some structure that allows us to specify what the randomness is all about.

Let Ω be any given nonempty set and let $\mathcal{F}$ be a nonempty collection of subsets satisfying the following properties:

1. $\mathcal{F}$ is "closed" under complements: $F \in \mathcal{F} \implies \Omega \backslash F \in \mathcal{F}$
2. $\mathcal{F}$ is "closed" under countable unions: $F_n \in \mathcal{F}$ for all $n \in \mathbb{N} \implies \cup_{n=1}^{\infty} F_n \in \mathcal{F}$.

Definition (**σ-algebra**)**.** (Billingsley (1995), p. 19 and 20) If the nonempty collection $\mathcal{F} \subseteq 2^{\Omega}$ satisfies properties 1,2 above then $\mathcal{F}$ is called a σ-algebra (or a σ-algebra of subsets of Ω)

Definition (**Measurable Space**)**.** (Billingsley (1995), p. 161) The pair $(\Omega, \mathcal{F})$, where $\mathcal{F}$ is a σ-algebra of subsets of Ω is called a measurable space. The subsets of $\mathcal{F}$ are called the events of Ω.

If you are interested in playing with these definitions, here is a simple exercise.

Practice Problem 1 (Optional)**.** Let A be a collection of elements of 2^{Ω}. Define

$$F^*(A) \equiv \{\mathcal{F} \mid \mathcal{F} \text{ is a } \sigma\text{-algebra of } \Omega \text{ containing } A\}$$

i) Show that $F^*(A)$ is non-empty.

ii) Let $\sigma(A)$ denote the intersection over all the σ-algebras contained in $F^*(A)$. Show that $\sigma(A)$ is a σ-algebra. The collection $\sigma(A)$ will be the called the "σ-algebra generated by A". Such collection is unique.

Practice Problem 2 (Optional)**.** Show that σ-algebra is "closed" under countable intersections. That is, if $F_n \in \mathcal{F}$ for all $n \in \mathbb{N}$ then $\cap_{n=1}^{\infty} F_n \in \mathcal{F}$.

1.1.2 Probability Measures

Once we know which are the objects that uncertainty will refer to, we need to have a formal way of expressing the "likelihood" of the different events we have specified. In order to do this we will define a *probability measure.*

Definition. (**Probability Measure**) (Billingsley (1995), p. 22) Let $(\Omega, \mathcal{F})$ be a measurable space. Let $\mathbb{P} : \mathcal{F} \to [0,1]$ be a "set" function mapping the σ-algebra of subsets of Ω into the real line. We say that $\mathbb{P}$ is a probability measure if it satisfies the following properties:

1. $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$ [Normalization]

2. $\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$ for a countable disjoint sequence of elements of $\mathcal{F}$ [σ-additivity]

Definition (**Probability space**). (Billingsley (1995), p. 23) The triplet $(\Omega, \mathcal{F}, P)$ is called a probability space.

Some implications of the definition of a probability measure.

1. **Monotonicity:** For events A, B, $A \subseteq B$ implies $\mathbb{P}(A) \leq \mathbb{P}(B)$.

   Proof. Write $B = A \cup (B \backslash A)$. By assumption, $A \in \mathcal{F}$; Practice Problem 2 implies $B \setminus A \in \mathcal{F}$ because $B \setminus A = B \cap (\Omega \setminus A)$. By definition $A \cap (B \backslash A) = \emptyset$. Therefore, by finite additivity (which is implied by σ-additivity) we have $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \backslash A)$, which implies $\mathbb{P}(B) - \mathbb{P}(A) = \mathbb{P}(B \backslash A) \geq 0$, as $\mathbb{P}$ is a $[0,1]$-valued set function. □

2. **Boole's Inequality:** (Billingsley (1995) p. 24) Boole's inequality is a pretty useful result that will show up quite often. Let A_n be any sequence of events in $\mathcal{F}$. Then

$$\mathbb{P}(\cup_{n=1}^{\infty} A_n) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_n)$$

   We provide a proof of Boole's inequality for only two events A, B. This is, we show that:

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$$

   Proof. Let $A_1 = A$, $A_2 = B \backslash A$. By finite additivity:

$$\begin{aligned} \mathbb{P}(A \cup B) &= \mathbb{P}(A_1 \cup A_2) \\ &= P(A_1) + P(A_2) \end{aligned}$$

   Since $B \backslash A \subseteq B$, then $\mathbb{P}(B \backslash A) \leq P(B)$. Therefore, $\mathbb{P}(A \cup B) \leq P(A) + P(B)$. □

In order for you to manipulate the definition of a probability measure, you will be asked to write down proofs of the following statements:

Practice Problem 3 (Properties of a probability measure)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Show that a probability measure satisfies:

1. $\mathbb{P}(F_1 \cup F_2) = \mathbb{P}(F_1) + \mathbb{P}(F_2) - \mathbb{P}(F_1 \cap F_2)$ for any $F_1, F_2 \in \mathcal{F}$
2. $\mathbb{P}(\cup_{n\in\mathbb{N}} F_n) \leq \sum_{n\in\mathbb{N}} \mathbb{P}(F_n)$ for any countable collection $\{F_n\}$

1.2 Random Variables

Along with a probability measure, a central concept in probability and statistics is that of a *random variable.*

Definition (S-valued random variable)**.** Let $(\Omega, \mathcal{F})$, $(S, \mathcal{S})$ be two measurable spaces. The map:

$$X : \Omega \to S$$

is called an S-valued random variable [relative to $(\Omega, \mathcal{F})$-$(S, \mathcal{S})$] if for all $A \in \mathcal{S}$:

$$X^{-1}(A) \equiv \{\omega \in \Omega \mid X(\omega) \in A\} \in \mathcal{F},$$

this is, if $X^{-1}(A)$ is *measurable* for all $A \in \mathcal{S}$.

Note that, by definition, a random variable takes events in the space S to well-defined events in the space Ω. If there is already a well-defined probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$, then there is a natural "induced" probability measure in $(S, \mathcal{S})$:

Definition ("Induced" Probability Measure or "Law" of a Random Variable)**.** (see Billingsley (1995) p. 185 on Transformation of measures and see if you can make the connection) Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. The probability measure over $(S, \mathcal{S})$ induced by the random variable $X : \Omega \to S$ is given by:

$$P_X(A) \equiv \mathbb{P}(X^{-1}(A)) \quad \forall\, A \in \mathcal{S}.$$

1.2.1 Cumulative Distribution Function for an $\mathbb{R}$-valued random variable

In this section, we will analyze the statement: "X is a real-valued random variable with a *cumulative distribution function* $F : \mathbb{R} \to [0, 1]$". We will explain the relation between the c.d.f. of a random variable and the induced probability measure we discussed in the last subsection. We will focus on real-valued random variables (relative to the Borel σ-algebra on the real line, which is the smallest σ-algebra containing all open sets).

Definition. (Billingsley (1995), p. 188 and 256) The cumulative distribution function (c.d.f.) of a real-valued random variable $X : \Omega \to \mathbb{R}$ is defined as the real-valued function $F : \mathbb{R} \to [0,1]$ given by

$$F_X(x) \equiv \mathbb{P}\{\omega \in \Omega \mid X(\omega) \leq x\} = \mathbb{P}\{X^{-1}(-\infty, a]\}$$

The c.d.f. of a random variable X is nothing else than its induced probability measure evaluated at sets of the form $(-\infty, a]$. The c.d.f. satisfies the following properties:

Proposition 1. *If F_X is the c.d.f. of a random variable $X : \Omega \to \mathbb{R}$ then*

1. *F_X is non-decreasing.*
2. $\lim_{x\uparrow\infty} F_X(x) = 1$
3. $\lim_{x\downarrow-\infty} F_X(x) = 0$
4. $\lim_{h\to0^+} F(x+h) = F(x)$

Furthermore, if F is a function satisfying 1,2,3,4, then there is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a random variable $X : \Omega \to \mathbb{R}$ such that F coincides with F_X.

Practice Problem 4. This week's problem set will walk you through the proof of Proposition 1.

Here are some commonly used examples of c.d.f.s of real-valued random variables:

- *Uniform Distribution*: The random variable X is said to have a uniform distribution in $[a,b]$ if the c.d.f. is given by:

$$F(x) = \begin{cases} 0 & \text{if} \quad x < a \\ x - a/[b-a] & \text{if} \quad x \in [a,b) \\ 1 & \text{if} \quad x \geq b \end{cases}$$

- *Bernoulli Distribution*: The random variable X is said to have a Bernoulli distribution with parameter $p \in [0,1]$ if its c.d.f. is given by:

$$F(x) = \begin{cases} 0 & \text{if} \quad x < 0 \\ 1-p & \text{if} \quad x \in [0,1) \\ 1 & \text{if} \quad x \geq 1 \end{cases}$$

- *Geometric Distribution*: The random variable X is said to have a Geometric Distribution with parameter p if its c.d.f. is given by:

$$F(x) = \begin{cases} 0 & \text{if} \quad x < 1 \\ 1-(1-p)^k & \text{if} \quad x \in [k,k+1) \text{ for } k \in \mathbb{N} \end{cases}$$

- *Exponential Distribution*: The random variable X is said to have an exponential distribution with parameter λ>0 if

$$F(x) = \begin{cases} 0 & \text{if} \quad x < 0 \\ 1-e^{-\lambda x} & \text{if} \quad x \geq 0 \end{cases}$$

- *Gaussian Distribution:* The random variable X is said to have a Gaussian distribution with parameters (μ, σ^2) if

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{1}{2\sigma^2}(u-\mu)^2 \Big) du$$

- *Pareto Distribution:* The random variable X is said to have a Pareto distribution with parameters (x_m, α) if

$$F(x) = \begin{cases} 0 & \text{if} \quad x < x_m \\ 1-\left(\frac{x_m}{x}\right)^{\alpha} & \text{if} \quad x \geq x_m \end{cases}$$

One common way to classify the random variables and their c.d.f.s is the following:

Definition (Discrete and Absolutely Continuous Random Variables)**.**

1. **Discrete Random Variables:** A random variable X is discrete if there is a countable set $\{x_1, x_2, \ldots\}$, $x_1 < x_2 < \ldots$ such that

$$\mathbb{P}_X(\{x_1, x_2 \ldots\}^c) = 0$$

   The smallest set $\{x_1, x_2, \ldots\}$ for which $\mathbb{P}_X(\{x_1, x_2 \ldots\}^c) = 0$ is called the support of X. The map $p : \text{Supp} \to [0,1]$ defined by $p(s) \equiv \mathbb{P}_X(s)$ is called the probability mass function.

2. **Absolutely Continuous Random Variables (w.r.t. Lebesgue Measure):** A random variable X is absolutely continuous if:

$$F(x) = \int_{-\infty}^{x} f(y)dy$$

for some integrable, nonnegative function f that we will call the **probability density function** (p.d.f.) of X. If F is continuously differentiable the p.d.f. will coincide with the derivative of F (see p. 257 in Billingsley (1995)).

A Bernoulli r.v. is an example of a discrete distribution with finite support. A Geometric r.v. is an example of a discrete random variable with countable support. The gaussian, pareto, and exponential distribution are both continuous and absolutely continuous random variables. Note that not every random variable falls into one of the categories above (see, for example, the entry for Cantor's function in wikipedia) .

1.2.2 Moments of Discrete and Absolutely Continuous Random Variables

In this section we introduce the definition of expectation for only discrete and absolutely continuous random variables. For a more detailed exposition you can see Chapters 1.4-1.6 in Durrett (2010).

Definition (Expectation of the transformation of an absolutely continuous random variable)**.** Let X be an absolutely continuous real-valued random variable with c.d.f F and p.d.f. $f(x)$. Let $g : X \to \mathbb{R}$ be a real-valued function. The expectation of $g(X)$ under F is defined as:

$$\mathbb{E}_F[g(X)] \equiv \int_{\mathbb{R}} g(x)f(x)dx$$

Definition (Expectation of the transformation of a discrete random variable)**.** Let $\{x_1, x_2, \ldots\}$ be the support of the discrete random variable X. Let $g : X \to \mathbb{R}$ be a real-valued function. The expectation of $g(X)$ under the probability mass function p is given by:

$$\mathbb{E}_F[g(X)] \equiv \sum_{x_i \in \text{Supp}} g(x_i)p(x_i).$$

Now, some important observations:

1. **Mean of a random variable:** $E_F[X] \equiv \mu$ is called the *mean* or *first moment* of X. If $E_f[|X|] < \infty$ then X is said to be integrable.

2. **Variance of a random variable:** $E_F[(X-\mu)^2] \equiv \sigma^2$ is called the *variance* or *second centered moment* of X. The mean and the variance of random variables will play an important role

in Weak Laws of Large numbers and Central Limit Theorems. For certain random variables the variance need not be finite (the pareto distribution with parameter $\alpha = 2$ is an example of this fact)

3. **Jensen's Inequality:** For any convex function g: $\mathbb{E}_F[g(X)] \geq g(\mathbb{E}[X])$. This is an important result. We will not prove it, but we will be using it quite often.

4. **Markov's Inequality:** For any random variable X and $c > 0$, $c\mathbb{P}_X[|X| > c] \leq \mathbb{E}_F[|X|]$

5. **k-th Moment:** $E_F[X^k]$ is called the k-th uncentered moment of the random variable X, $k = 1, 2, \ldots$. Note that Jensen's inequality implies that if $E_F[|X|^k] < \infty$ for some $k \in \mathbb{N}$ then $E_F[|X|^j] < \infty$ for all $j < k$. Likewise, if $E_F[|X|^k] = \infty$, then $E_F[|X|^j] = \infty$ for all $j > k$.

6. **Layer Representation of k-th moments:** if X is an absolutely continuous random variable and $\mathbb{E}[|X|^k] < \infty$ then $\mathbb{E}[|X|^k] = k\int_0^\infty c^{k-1}\mathbb{P}_X(|x| > c)dc$.

7. **Moment Generating Function:** The random variable X is said to have a moment generating function $m_X(t)$ if

$$m_X(t) \equiv \mathbb{E}_F[\exp(tX)] < \infty \quad \text{for all } t \in (-\epsilon, \epsilon)$$

   for some $\epsilon > 0$. Important: Two random variables with the same moment generating function have the same distribution.

The last question of this week's problem set will ask you to go over some of these concepts.

Quick Summary of Lectures 1 and 2: We have provided formal definitions of a probability space and a real-valued random variable.

We defined discrete and absolutely continuous random variables. We used our classification to provide a definition of the expectation of a function $g(X)$. We defined the mean of a real-valued random variable (μ), the variance of a real-valued random variable (σ^2), the k-th moment of a real-valued random variable, and at the very end of Lecture 2, the *moment generating function.*

The moment generating function contains—in a sense that we did not make precise—the same information as the c.d.f. of a real-valued random variable. Here are some results to keep in mind.

- Two discrete random variables with the same moment generating function have the same support and the same probability mass function. See Billingsley (1995), p.147, second paragraph.

- Two positive random variables with the same moment generating function have the same distribution. See Billingsley (1995), Theorem 22.2. Can you fill in the details of the proof?

Optional: Suppose that F_X has a density f. Suppose that $m_X(t) = \mathbb{E}[\exp(tX)]$ is well defined for every $t \in \mathbb{R}$. The question is: how do we recover f from the moment generating function? An sketch of the answer is the following:

1. Extend the image of the m.g.f. to the *complex* plane: Let $i = \sqrt{-1}$. For $t \in \mathbb{R}$, define:

$$\phi(t) \equiv m_X(it) = \mathbb{E}[\exp(itX)]$$

   The function $\phi(t) : \mathbb{R} \to \mathbb{C}$ is called the *characteristic function* (in fact, this is always well defined: even if the m.g.f. does not exist.).

2. Recover the p.d.f. from the characteristic function: It turns out that one can prove the following equality (*or inversion formula*):

$$f(x) = \int_{-\infty}^{\infty} \exp(-it)\phi(t)dt$$

You can found details in Section 26 (342-349) in Billingsley (1995) or Section 3.3 (106-111) in Durrett (2010). The main take away (and we will use this result later) is that under certain conditions two random variables with the same moment generating function have the same distribution over the Borel σ-algebra in the real line. In this sense, the moment generating function provides a characterization of the c.d.f.

References

Billingsley, P. (1995): *Probability and Measure*, John Wiley & Sons, New York, 3rd ed.

Durrett, R. (2010): *Probability: Theory and Examples*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 4th ed.