2 Mathematical Foundations: Part I

2.1 Probabilistic Model

We first consider an experiment of which there are n possible outcomes, $\omega_1, \omega_2, \ldots, \omega_n$. We call $\omega_1, \omega_2, \ldots, \omega_n$ sample points, and the set $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ the sample space. For example, for a single toss of a coin, $\Omega = \{H, T\}$, where H denotes that the coin lands heads up and T denotes that the coin lands tails up. For a single toss of a die, $\Omega = \{1, 2, 3, 4, 5, 6\}$, and for n tosses of a die, $\Omega = \{\omega : \omega = (a_1, a_2, \ldots, a_n), a_i = 1, 2, 3, 4, 5 \text{ or } 6\}$. In this example, there are 6^n sample points in the sample space.

In each of the above settings, a probabilistic model may assign a probability to each sample point. For example, in the experiment of a single toss of a die, if the die is fair, we may assign a probability of 1/6 to each of the sample points in $\Omega = \{1, 2, 3, 4, 5, 6\}$. That is, if we define $p(\omega)$ to be the probability of ω, then our probability model is $p(1) = p(2) = p(3) = p(4) = p(5) = p(6) = 1/6$. In principle, we may assign any non-negative numbers to the six sample points as their probabilities, as long as these numbers sum up to one. We make the particular assignment above to reflect our assumption that the die is fair. (This is what we mean by "a model".)

With such an assignment of probabilities, we may calculate the probability of any event. An event is denoted by a subset of the sample space. For example, $A = \{3, 4\}$ corresponds to the event that "3 or 4 is facing up in tossing a die", and this event happens with probability $\mathbb{P}(A) = p(3) + p(4) = 1/3$. The subset $B = \{2, 4, 6\}$ corresponds to the event that "an even number is facing up in tossing a die", and it has probability $\mathbb{P}(B) = p(2)+p(4)+p(6) = 1/2$.

We may be able to construct "new" events from "old" events. For example, if A and B are two events, then A^C denotes that A does not happen, $A \cap B$ denotes that A and B happens at the same time, and $A \cup B$ denotes that either A or B happens. We may calculate their probabilities easily. For example, $\mathbb{P}(A^C) = \mathbb{P}(\{1, 2, 5, 6\}) = p(1) + p(2) + p(5) + p(6) = 2/3$.

Such assignments of probabilities are intuitive and convenient. However, difficulties arise when the experiment (or the sample space) becomes complex. Consider an infinite independent tosses of a fair coin. If we label heads-up as 1 and tails-up as 0, the sample space becomes

$$\Omega = \{\omega : \omega = (a_1, a_2, a_3, \ldots), a_i = 0 \text{ or } 1\}.$$

How many points are there in Ω? It is well known that every real number in $[0, 1)$ has a unique binary expansion of the form $0.b_1b_2b_3\ldots$, where $b_i = 0$ or 1. As a result, there is a one-to-one mapping between points in Ω and points in $[0, 1)$[1]. That is, there are as many points in Ω as in $[0, 1)$, and we may simply take Ω as $[0, 1)$, and take our experiment as choosing a number randomly from $[0, 1)$.

If we continue to assign probabilities to each point in $[0, 1)$, since each point should have the same probability, and these probabilities have to sum up to one, we must assign a probability of zero to each of the points in $[0, 1)$. However, such an assignment does not lead to very far. For example, it is not clear how to obtain $\mathbb{P}([0, 1/2))$ from the assignment that each single point has probability zero, although intuitively, we know that the probability should be 1/2.

Given the fact that we don't know how to calculate the probability of an event from the probabilities of sample points but we may still "know" the probability of that event, why don't we assign probabilities directly to the events instead of to the sample points? Of course, any such assignment should be consistent in some sense. For example, a smaller set should be assigned a probability no greater than a larger set that contains the smaller set, and the assigned probability for the union of two disjoint sets should equal to the sum of the assigned probability for the two individual sets. This is the idea behind the axiomatic formulation of probability theory following Kolmogorov.

[1] Strictly speaking, the mapping from Ω to $[0, 1)$ is not exactly one-to-one but instead surjective. For example, $1/2 = 0.1000\ldots = 0.0111\ldots$. That is, we may find two points in Ω that corresponds to the same number in $[0, 1)$. However, this does not affect the main message we would like to deliver. See Billingsley (1995).

2.2 Probability Spaces

To construct a model for an experiment, we need three ingredients: the sample space, the events, and the probabilities assigned to the events. The sample space Ω usually depends on the design of the experiment, and we should know it immediately after we understand what the experiment is. From the section above, we know that we may express an event simply as a subset A of Ω. The question is, which subsets of Ω should be included in the system of events so that the resulting probability model becomes useful and convenient to describe probabilistic phenomena in the real world? From the discussion in the previous section, we know that if A is an event, it is better that A^C, which denotes that the event A does not happen, is an event. Moreover, if A and B are events, it is better that $A \cap B$, denoting that A and B happen at the same time, and $A \cup B$, denoting that either A or B happens, are also events. This motivates the following definition.

Definition 2.1. Let Ω be a nonempty set. A system $\mathcal{A}$ of subsets of Ω is called an algebra if it satisfies the following conditions:

1. $\Omega \in \mathcal{A}$.
2. If $A \in \mathcal{A}$, then $A^C \in \mathcal{A}$.
3. If $A \in \mathcal{A}$ and $B \in \mathcal{A}$, then $A \cup B \in \mathcal{A}$.

We note here that 2 and 3 together implies that if $A \in \mathcal{A}$ and $B \in \mathcal{A}$, then $A \cap B \in \mathcal{A}$.

Now we may take sets in an algebra $\mathcal{A}$ as events, and start to assign probabilities to these events. We may have different ways to assign numbers in different models, but any reasonable assignment $\mathbb{P}$ should make a probabilistic sense in that

1. if $A \in \mathcal{A}$, then $\mathbb{P}(A) \geq 0$,
2. $\mathbb{P}(\Omega) = 1$, and
3. if $A \in \mathcal{A}$ and $B \in \mathcal{A}$ are disjoint, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.

Note that $\mathbb{P}$ may be viewed as a set function from $\mathcal{A}$ to $\mathbb{R}$. A set function satisfying the above three conditions is called a finitely additive probability measure.

Now we have all the three ingredients and we may start to define a probabilistic model (in the extended sense).

Definition 2.2. A probabilistic model (in the extended sense), is an ordered triple $(\Omega, \mathcal{A}, \mathbb{P})$ where

1. Ω is a nonempty set,
2. $\mathcal{A}$ is an algebra of subsets of Ω, and
3. $\mathbb{P}$ is a finitely additive probability measure on $\mathcal{A}$.

It turns out that this model is too broad to lead to a fruitful theory. Instead, we need to restrict both the class of subsets of Ω and the class of set functions $\mathbb{P}$ we consider.

Definition 2.3. A system $\mathcal{F}$ of subsets of Ω is a σ-algebra if it satisfies the following conditions:

1. $\Omega \in \mathcal{F}$.
2. If $A \in \mathcal{F}$, then $A^C \in \mathcal{F}$.
3. If $A_1, A_2, \ldots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

If $A \in \mathcal{F}$, then A is called $\mathcal{F}$-measurable.

Definition 2.4. Let Ω be a nonempty set and $\mathcal{F}$ a σ-algebra of subsets of Ω. A set function $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ is a probability measure if it satisfies the following conditions:

1. If $A \in \mathcal{A}$, then $\mathbb{P}(A) \geq 0$.
2. $\mathbb{P}(\Omega) = 1$.
3. If $A_1, A_2, \ldots \in \mathcal{A}$ are disjoint, then

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

If a set function μ has only the properties 1 and 3, it is called a measure.

Definition 2.5. A probabilistic model, or a probability space, is an ordered triple $(\Omega, \mathcal{F}, \mathbb{P})$ where

1. Ω is a nonempty set,
2. $\mathcal{F}$ is a σ-algebra of subsets of Ω, and
3. $\mathbb{P}$ is a probability measure on $\mathcal{F}$.

2.3 Remarks on σ-algebras

In this section we make some remarks about σ-algebras. Given a sample space Ω, a σ-algebra on it could be very simple, containing only two sets, or it could be very complicated, containing more sets than we could handle. For example, if $\Omega = \mathbb{R}$, the simplest σ-algebra on Ω is the collection of sets $\{\varnothing, \mathbb{R}\}$. However, this σ-algebra is too small to generate any useful results. On the other end, the collection of all subsets of $\mathbb{R}$ is a σ-algebra. It turns out that this σ-algebra is so large that it is difficult to define "lengths" for all sets in this σ-algebra in a consistent way. So a lot of times we want to keep a balance. That is, we want to construct σ-algebras that contain certain sets, or in the language of probability, events, that we care about, while keeping the σ-algebra small enough so that it is easy to work with. This motivates us to define the concept of generated σ-algebra.

Definition 2.6. Let Ω be a nonempty set and $\mathcal{E}$ a collection of subsets of Ω. The smallest σ-algebra that contains $\mathcal{E}$, denoted by $\sigma(\mathcal{E})$, is called the σ-algebra generated by $\mathcal{E}$.

In the above definition, "smallest" means that $\sigma(\mathcal{E})$ is contained in any σ-algebra that contains $\mathcal{E}$. In other words, it is the intersection of all σ-algebras that contain $\mathcal{E}$. It is easy to show that this intersection is indeed a σ-algebra.

There is a very useful σ-algebra on $\mathbb{R}$, called the Borel σ-algebra, that is generated by all open subsets of $\mathbb{R}$. We usually denote this σ-algebra by $\mathcal{B}(\mathbb{R})$. A Borel-measurable set is called a Borel set. It is easy to show that $\mathcal{B}(\mathbb{R})$ contains the singletons $\{a\}, a \in \mathbb{R}$ and all sets of the forms (a, b), $[a, b]$, $(a, b]$, $[a, b)$, $(-\infty, a)$, $(-\infty, a]$, (a, ∞), $[a, \infty)$. Of course, it contains much more sets that the ones specified above. Actually, the Borel σ-algebra suffices most of our purposes, and it is not easy to think of a set that is not Borel.

There is a special measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, called the Lebesgue measure, that has the interpretation of "length". We define a set function λ such that for any set of the form $(a, b]$, $\lambda((a, b]) = b - a$. Now that λ measures the length of the sets of the form $(a, b]$. It is well known that these sets generate the σ-algebra $\mathcal{B}(\mathbb{R})$, and the set function λ extends in a unique way to a measure on $\mathcal{B}(\mathbb{R})$ in the sense that the extension now 1) is a set function defined for all sets in $\mathcal{B}(\mathbb{R})$, 2) gives the length for the sets of the form $(a, b]$, and 3) is a measure that satisfies the conditions in the definition above. This measure is called the Lebesgue measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and is denoted also by λ.

2.4 Random Variables

Intuitively, a "random" variable is a variable whose value is not deterministic. But how to formally model such a variable? Think about a gamble whose return depends on the result of tossing a die. Suppose the rule of the game is the following. First toss a die. If an odd number is facing up, you get one dollar. If an even number is facing up, you lose one dollar. Then your return of playing this game is a "random" variable. We know that we could write the sample space as $\Omega = \{1, 2, 3, 4, 5, 6\}$. Let X be your return of playing the game. Then it is tempting to write something like

$$\begin{aligned}
&X(\omega = 1) = 1, \\
&X(\omega = 2) = -1, \\
&X(\omega = 3) = 1, \\
&X(\omega = 4) = -1, \\
&X(\omega = 5) = 1, \\
&X(\omega = 6) = -1.
\end{aligned}$$

We see that actually we may view the random variable X simply as a function from the sample space Ω to $\mathbb{R}$. Now the question is, can an arbitrary function from the sample space to $\mathbb{R}$ be regarded as a random variable? It turns out that for a theory of probability to be useful, we need to restrict the functions we consider a little bit. Think about the following question: what is the probability of earning one dollar in the game? To be able to get the probability, we need to figure out the event, which is defined to be a subset of Ω, that corresponds to the result of earning one dollar in the game. Obviously, the event is given by $X^{-1}(\{1\}) = \{1, 3, 5\}$. To be able to talk about the probability of the event, according to the construction of probabilistic models in the previous section, we need to require that the set $\{1, 3, 5\}$, which is $X^{-1}(\{1\})$, to be in the σ-algebra $\mathcal{F}$ if $(\Omega, \mathcal{F}, \mathbb{P})$ is the probability space corresponding to the game.

Now consider a general random variable $X : \Omega \to \mathbb{R}$ with the underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$. As we mentioned above, most of the time the Borel sets suffice our purposes. So the question we always asks is, what is the probability of X taking values in a Borel set B? The corresponding event is $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$. To be able to talk about the probability of this event $X^{-1}(B)$, we need $X^{-1}(B) \in \mathcal{F}$.

Now things become clear. A random variable should be defined as a special kind of function.

Definition 2.7. Let f be a function from a set Ω to $\mathbb{R}$, and let $\mathcal{F}$ be a σ-algebra corresponding to Ω. The function f is called $\mathcal{F}$-measurable (or simply measurable if the underlying σ-algebra is clear from the context) if $f^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{B}(\mathbb{R})$.

Definition 2.8. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A random variable is a real function on Ω that is $\mathcal{F}$-measurable.

Defining random variables as measurable functions has the advantage in terms of integration, which we shall introduce soon in a later section. Here we first look at the probability distribution of a random variable X on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

To motivate, we still take the example of the game above. We know that we may construct the probability space for the experiment of tossing a die as a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where

$$\Omega = \{1, 2, 3, 4, 5, 6\},$$

$$\mathcal{F} = \sigma(\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}),$$

and

$$\mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = \mathbb{P}(\{3\}) = \mathbb{P}(\{4\}) = \mathbb{P}(\{5\}) = \mathbb{P}(\{6\}) = \frac{1}{6}.$$

Now we consider the distribution of the values of X on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. We know that the gambler earns one dollar with probability 1/2, earns negative one dollars with probability 1/2, and earns other amounts with probability zero. Therefore, we have a new probability measure $\mathbf{P_X}$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ given by

$$\mathbf{P}_X(\{x\}) = \begin{cases} 1/2, & \text{if } x = 1 \\ 1/2, & \text{if } x = -1 \\ 0, & \text{otherwise.} \end{cases}$$

Now we have a new probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbf{P}_X)$ induced by X. The probability measure $\mathbf{P_X}$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with

$$\mathbf{P}_X(B) = \mathbb{P}\{X \in B\}, \qquad B \in \mathcal{B}(\mathbb{R})$$

is called the probability distribution of X on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The function

$$F_X(x) = \mathbb{P}(\{X \leq x\}) = \mathbf{P}_X((-\infty, x]), \qquad x \in \mathbb{R}$$

is called the (accumulative) distribution function of X. It is easy to see that F_X is a nondecreasing, right continuous function with left limits. Also, we have $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$.

Let X be a random variable and $g : \mathbb{R} \to \mathbb{R}$ be a measurable function. Then $g(X)$ is a random variable. (Try to show this by yourself.)

Let X and Y be random variables. Then $X + Y, X - Y, XY$ and X/Y are random variables (provided that no indeterminate forms such as $\infty - \infty, \infty/\infty, a/0$ appear.)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X_1, X_2, \ldots, X_n$ be random variables. If we stack these random variables together, we get a random vector $(X_1, X_2, \ldots, X_n)'$. We may view a random vector as a measurable mapping from $(\Omega, \mathcal{F}, \mathbb{P})$ to $\mathbb{R}^n$ with the Borel σ-algebra $\mathcal{B}(\mathbb{R}^k)$ generated by all open subsets of $\mathbb{R}^k$.

Let $\{X_i\}_{i\in I}$ be a set of random variables. We say that the set of random variables are (mutually) independent if for every finite set of indices $i_1, i_2, \ldots, i_n$ the random variables $X_{i_1}, X_{i_2}, \ldots, X_{i_n}$ are independent, i.e.,

$$\mathbb{P}(X_{i_1} \in B_1, X_{i_2} \in B_2, \ldots, X_{i_n} \in B_n) = \mathbb{P}(X_{i_1} \in B_1)\mathbb{P}(X_{i_2} \in B_2) \cdots \mathbb{P}(X_{i_n} \in B_n),$$

$B_1, B_2, \ldots, B_n \in \mathcal{F}$.

2.5 Lebesgue Integral and Mathematical Expectation

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We may define the expectation of a random variable as the Lebesgue integral of X with respect to the probability measure $\mathbb{P}$. We first consider a simple case when X can be written as

$$X(\omega) = \sum_{i=1}^{n} x_i I_{A_i}(\omega)$$

where for any $A \in \mathcal{F}, I_A$ is the indicator function taking values either zero or one:

$$I_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A, \\ 0, & \text{otherwise.} \end{cases}$$

We call such random variables simple. Now we define the expectation of a simple random variable X to be

$$\mathbb{E}X = \sum_{i=1}^{n} x_i \mathbb{P}(A_i).$$

Now for any non-negative random variable X, it is well known that there exists a sequence of simple random variables $\{X_n\}$ such that $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega$. Then we define the expectation of X by

$$\mathbb{E}X = \lim_{n\to\infty} \mathbb{E}X_n.$$

We may show that the limit on the right hand side above is independent of the choice of the sequence $\{X_n\}$. Note that it is possible that $\mathbb{E}X = \infty$.

For a general random variable X, we can decompose it as

$$X = X^+ - X^-$$

where $X^+ = \max(X, 0)$, and $X^- = -\min(X, 0)$. Since both X^+ and X^- are non-negative random variables, we define their expectations as above and define the expectation of X as

$$\mathbb{E}X = \mathbb{E}X^+ - \mathbb{E}X^-$$

if at least one of $\mathbb{E}X^+$ and $\mathbb{E}X^-$ is finite. The expectation $\mathbb{E}X$ is also called the Lebesgue integral of X with respect to the probability measure $\mathbb{P}$, and is denoted $\int_\Omega X \mathrm{d}\mathbb{P}$, or $\int_\Omega X(\omega)\mathbb{P}(\mathrm{d}\omega)$. We say that X is integrable if both $\mathbb{E}X^+$ and $\mathbb{E}X^-$ are finite.

The expectation has the following properties.

1. If X and Y are non-negative random variables, or if $\mathbb{E}\,|X| < \infty$ and $\mathbb{E}\,|Y| < \infty$, then $\mathbb{E}(X+Y) = \mathbb{E}X + \mathbb{E}Y$.
2. If $\mathbb{E}X$ exists and c is a constant, then $\mathbb{E}(cX) = c\mathbb{E}X$.

In fact, we may define the Lebesgue integral for any measurable function with respect to any measure in a similar way by first defining the integral for piecewise constant functions, then for non-negative measurable functions, and then for a general measurable function. Of course, when we talk about the expectation of a random variable, we always mean its expectation with respect to a probability measure.

Now we may compare the Lebesgue integral to the Riemann integral. Recall that the Riemann integral is defined as the limit of the Riemann sum: one partitions the domain of the function, form the rectangles, and calculate the area of the rectangles. One then make the partition finer and finer, and get a sequence of areas. If the limit of the sequence converges, we define it as the Riemann integral. On the other hand, intuitively the Lebesgue integral partitions the range of the function, and the rest are the same. The advantage of partitioning in this new way is that now we are able to define integral for a much larger class of functions. The Riemann integral only works for piecewisely continuous functions, while Lebesgue integral could work for functions that are no where continuous. Another important advantage is that for Riemann integrals, we can only exchange the integral sign and the limit sign under restrictive conditions (for example, uniform convergence). For Lebesgue integral, the exchange of the integral sign and the limit sign is possible under very mild conditions.

Theorem 2.9 (On Monotone Convergence)**.** *Let $X, Y, X_1, X_2, \ldots$ be random variables such that $X_n > Y$ for all $n \geq 1, \mathbb{E}Y > -\infty$, and $X_n \uparrow X$. Then*

$$\lim_{n\to\infty} \mathbb{E}X_n = \mathbb{E}\left(\lim_{n\to\infty} X_n\right) = \mathbb{E}X.$$

Theorem 2.10 (Lebesgue's Dominated Convergence)**.** *Let $X, X_1, X_2, \ldots$ be random vari-*

ables such that $|X_n| < Y$ for all $n \geq 1, \mathbb{E}Y < \infty$, and $\lim_{n\to\infty} X_n = X$. Then

$$\lim_{n\to\infty} \mathbb{E}X_n = \mathbb{E}\left(\lim_{n\to\infty} X_n\right) = \mathbb{E}X.$$

The following inequality (Markov Inequality) is a direct consequence of the definition of the Lebesgue integral.

Theorem 2.11. *Let X be a random variable. Then for any $\epsilon > 0$ and for any integer $n \geq 1$, we have*

$$\mathbb{P}(|X| \geq \epsilon) \leq \frac{\mathbb{E}\,|X|^n}{\epsilon^n}.$$

The most famous special case is the Chebyshev Inequality:

Corollary 2.12. Let X be a random variable, then for any $\epsilon > 0$

$$\mathbb{P}(|X - \mathbb{E}X| \geq \epsilon) \leq \frac{\mathbb{E}(X - \mathbb{E}X)^2}{\epsilon^2}.$$

Note that from introductory probability, $\mathbb{E}(X - \mathbb{E}X)^2$ is defined as the variance of X. So the probability that a random variable taking values far away from its mean is bounded by the random variable's variance.