# Probability Theory

Muhammad Waliji

August 11, 2006

**Abstract**

This paper introduces some elementary notions in Measure-Theoretic Probability Theory. Several probabalistic notions of the convergence of a sequence of random variables are discussed. The theory is then used to prove the Law of Large Numbers. Finally, the notions of conditional expectation and conditional probability are introduced.

## 1 Heuristic Introduction

Probability theory is concerned with the outcome of experiments that are random in nature, that is, experiments whose outcomes cannot be predicted in advance. The set of possible *outcomes*, $\omega$, of an experiment is called the *sample space*, denoted by $\Omega$. For instance, if our experiment consists of rolling a dice, we will have $\Omega = \{1, 2, 3, 4, 5, 6\}$. A subset, $A$, of $\Omega$ is called an *event*. For instance $A = \{1, 3, 5\}$ corresponds to the event 'an odd number is rolled'.

In elementary probability theory, one is normally concerned with sample spaces that are either finite or countable. In this case, one often assigns a probability to every single outcome. That is, we have probability function $P: \Omega \to [0, 1]$, where $P(\omega)$ is the probability that $\omega$ occurs. Here, we inssist that

$$\sum_{\omega \in \Omega} P(\omega) = 1.$$

However, if the sample space is uncountable, then this condition becomes nonsensible. Two elementary types of problems come into this category and hence cannot be dealt with by elementary probability theory: an infinite number of repeated coin tosses (or dice rolls), and a number drawn at random from $[0, 1]$. This illustrates the importance of uncountable sample spaces.

The solution to this problem is to use the theory of measures. Instead of assigning probabilities to outcomes in the sample space, one can restrict himself to a certain class of events that form a structure known as a $\sigma$-*field*, and assign probabilities to these special kinds of events.

## 2 $\sigma$-Fields, Probability Measures, and Distribution Functions

**Definition 2.1.** A class of subsets of $\Omega$, $\mathcal{F}$, is a *$\sigma$-field* if the following hold:

(i) $\emptyset \in \mathcal{F}$ and $\Omega \in \mathcal{F}$

(ii) $A \in \mathcal{F} \Longrightarrow A^c \in \mathcal{F}$

(iii) $A_1, A_2, \ldots \in \mathcal{F} \Longrightarrow \bigcup_n A_n \in \mathcal{F}$

Note that this implies that $\sigma$-fields are also closed under countable interesections also.

**Definition 2.2.** The $\sigma$-field *generated* by a class of sets, $\mathcal{A}$, is the smallest $\sigma$-field containing $\mathcal{A}$. It is denoted $\sigma(\mathcal{A})$.

**Definition 2.3.** Let $\mathcal{F}$ be a $\sigma$-field. A function $P : \mathcal{F} \to [0,1]$ is a *probability measure* if $P(\emptyset) = 0$, $P(\Omega) = 1$, and whenever $(A_n)_{n\in\mathbb{N}}$ is a disjoint collection of sets in $\mathcal{F}$, we have

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n).$$

Throughout this paper, unless otherwise noted, the words increasing, decreasing, and monotone are always meant in their weak sense. Suppose $\{A_n\}$ is a sequence of sets. We say that $\{A_n\}$ is an *increasing sequence* if $A_1 \subset A_2 \subset \cdots$. We say that $\{A_n\}$ is a *decreasing sequence* if $A_1 \supset A_2 \supset \cdots$. In both of these cases, the sequence $\{A_n\}$ is said to be *monotone*. If $A_n$ is increasing, then set $\lim_n A_n := \bigcup_n A_n$. If $A_n$ is decreasing, then set $\lim_n A_n := \bigcap_n A_n$. The following properties follow immediately from the definitions.

**Lemma 2.4.** *Let $\mathcal{F}$ be a $\sigma$-field, and let $P$ be a probability measure on it.*

*(i)* $P(A^c) = 1 - P(A)$

*(ii) If $A \subset B$, then $P(A) \leq P(B)$.*

*(iii)* $P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$.

*(iv) If $\{A_n\}$ is a monotone sequence in $\mathcal{F}$, then* $\lim_n P(A_n) = P(\lim_n A_n)$.

**Definition 2.5.** Suppose $\Omega$ is a set, $\mathcal{F}$ is a field on $\Omega$, and $P$ is a probability measure on $\mathcal{F}$. Then, the ordered pair $(\Omega, \mathcal{F})$ is called a *measurable space*. The triple $(\Omega, \mathcal{F}, P)$ is called a *probability space*. The a probability space is finitely additive or countably additive depending on whether $P$ is finitely or countably additive.

**Definition 2.6.** Let $(X, \tau)$ be a topological space. The $\sigma$-field, $\mathcal{B}(X, \tau)$ generated by $\tau$ is called the *Borel $\sigma$-field*. In particular, $\mathcal{B}(X, \tau)$ is the smallest $\sigma$-field containing all open and closed sets of $X$. The sets of $\mathcal{B}(X, \tau)$ are called *Borel sets*.

When the topology, $\tau$, or even the space $X$ are obvious from the context, $\mathcal{B}(X,\tau)$ will often be abbreviated $\mathcal{B}(X)$ or even just $\mathcal{B}$.

A particularly important situation in probability theory is when $\Omega = \mathbb{R}$ and $\mathcal{F}$ are the Borel sets in $\mathbb{R}$.

**Definition 2.7.** A *distribution function* is an increasing right-continuous function $F:\mathbb{R}\to[0,1]$ such that

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

We can associate probability functions on $(\mathbb{R},\mathcal{B})$ with distribution functions. Namely, the distribution function associated with $P$ is $F(x) := P((-\infty,x])$. Conversely, each distribution function defines a probability function on the reals.

# 3 Random Variables, Transformations, and Expectation

We now have stated the basic objects that we will be studying and discussed their elementary properties. We now introduce the concept of a *Random Variable*. Let $\Omega$ be the set of all possible drawings of lottery numbers. The function $X:\Omega\to\mathbb{R}$ which indicates the payoff $X(\omega)$ to a player associated with a drawing $\omega$ is an example of a random variable. The *expectation* of a random variable is the "average" or "expected" value of $X$.

**Definition 3.1.** Let $(\Omega_1,\mathcal{F}_1)$ and $(\Omega_2,\mathcal{F}_2)$ be measurable spaces. A function $T:\Omega_1\to\Omega_2$ is a *measurable transformation* if the preimage of any measurable set is a measurable set. That is, $T$ is a measurable transformation if $(\forall A \in \mathcal{F}_2)(T^{-1}(A)\in\mathcal{F}_1)$.

**Lemma 3.2.** *It is sufficient to check the condition in Definition 3.1 for those $A$ in a class that generates $\mathcal{F}_2$. More precisely, suppose that $\mathcal{A}$ generates $\mathcal{F}_2$. Then, if $(\forall A\in\mathcal{A})(T^{-1}(A)\in\mathcal{F}_1)$, then $T$ is a measurable transformation.*

*Proof.* Let $\mathcal{C} := \{A \in 2^{\Omega_2} : T^{-1}(A)\in\mathcal{F}_1\}$. Then, $\mathcal{C}$ is a $\sigma$-field, and $\mathcal{A}\subset\mathcal{C}$. But then, $\sigma(\mathcal{A}) = \mathcal{F}_2 \subset \mathcal{C}$, which is exactly what we wanted. □

**Definition 3.3.** Let $(\Omega,\mathcal{F})$ be a measurable space. A *measurable function* or a *random variable* is a measurable transformation from $(\Omega,\mathcal{F})$ into $(\mathbb{R},\mathcal{B})$.

**Lemma 3.4.** *If $f:\mathbb{R}\to\mathbb{R}$ is a continuous function, then $f$ is a measurable transformation from $(\mathbb{R},\mathcal{B})$ to $(\mathbb{R},\mathcal{B})$.*

**Definition 3.5.** Given a set $A$, the *indicator function for $A$* is the function

$$I_A(\omega) := \begin{cases} 1 & \text{if } \omega\in A \\ 0 & \text{if } \omega\notin A \end{cases}$$

If $A \in \mathcal{F}$, then $I_A$ is a measurable function.

Note that many elementary operations, including composition, arithmetic, max, min, and others, when performed upon measurable functions, again yield measurable functions.

Let $(\Omega_1, \mathcal{F}_1, P)$ be a probability space and $(\Omega_2, \mathcal{F}_2)$ a measurable space. A measurable transformation $T : \Omega_1 \to \Omega_2$ naturally induces a probability measure $PT^{-1}$ on $(\Omega_2, \mathcal{F}_2)$. In the case of a random variable, $X$, the induced measure on $\mathbb{R}$ will generally be denoted $\alpha$. The distribution function associated with $\alpha$ will be denoted $F_X$. $\alpha$ will sometimes be called a *probability distribution.*

Now that we have a notion of measure and of measurable functions, we can develop a notion of the "integral" of a function. The integral will have the probabalistic interpretation of being an expected (or average) value. For the precise definition of the Lebesgue integral, see any textbook on Measure Theory.

**Definition 3.6.** Suppose $X$ is a random variable. Then the *expectation of $X$* is $EX := \int_\Omega X(\omega) \mathrm{d}P$.

We conclude this section with a useful change of variables formula for integrals.

**Proposition 3.7.** *Let $(\Omega_1, \mathcal{F}_1, P)$ be a probability space and let $(\Omega_2, \mathcal{F}_2)$ be a measurable space. Suppose $T : \Omega_1 \to \Omega_2$ is a measurable transformation. Suppose $f : \Omega_2 \to \mathbb{R}$ is a measurable function. Then, $PT^{-1}$ is a probability measure on $(\Omega_2, \mathcal{F}_2)$ and $fT : \Omega_1 \to \mathbb{R}$ is a measurable function. Furthermore, $f$ is integrable iff $fT$ is integrable, and*

$$\int_{\Omega_1} fT(\omega_1) \mathrm{d}P = \int_{\Omega_2} f(\omega_2) \mathrm{d}PT^{-1}.$$

# 4 Notions of Convergence

We will now introduce some notions of the convergence of random variables. Note that we will often not explicitly state the dependence of a function $X(\omega)$ on $\omega$. Hence, sets of the form $\{\omega : X(\omega) > 0\}$ will often be abbreviated $\{X > 0\}$. For the remainder of this section, let $X_n$ be a sequence of random variables.

**Definition 4.1.** The sequence $X_n$ *converges almost surely (almost everywhere)* to a random variable $X$ if $X_n(\omega) \to X(\omega)$ for all $\omega$ outside of a set of probability 0.

**Definition 4.2.** The sequence $X_n$ *converges in probability (in measure)* to a function $X$ if,

$$\text{for every } \epsilon > 0, \lim_{n \to \infty} P\{\omega : |X_n(\omega) - X(\omega)| \geq \epsilon\} = 0.$$

This is denoted $X_n \xrightarrow{P} X$.

**Proposition 4.3.** *If $X_n$ converges almost surely to $X$, then $X_n$ converges in probability to $X$.*

*Proof.* We have $\{\omega : X_n(\omega) \nrightarrow X(\omega)\} \subset N$, $P(N) = 0$. That is,

$$\forall \epsilon > 0 \quad \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} \{|X_m - X| \geq \epsilon\} \subset N.$$

Therefore, given $\epsilon > 0$, we have

$$\begin{aligned} \lim_{n\to\infty} P\{|X_n - X| \geq \epsilon\} &\leq \lim_{n\to\infty} P \bigcup_{m=n}^{\infty} \{|X_m - X| \geq \epsilon\} \\ &= P \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} \{|X_m - X| \geq \epsilon\} \leq P(N) = 0 \end{aligned}$$

thereby completing the proof. □

Note, however, that the converse is not true. Let $\Omega = [0, 1]$ with Lebesgue measure. Consider the sequence of sets $A_1 = [0, \frac{1}{2}]$, $A_2 = [\frac{1}{2}, 1]$, $A_3 = [0, \frac{1}{3}]$, $A_4 = [\frac{1}{3}, \frac{2}{3}]$, and so on. Then, the indicator functions, $I_{A_n}$, converge in probability to 0. However, $I_{A_n}(\omega)$ does not converge for any $\omega$, and in particular the sequence does not converge almost surely. However, the following holds as a sort of converse:

**Proposition 4.4.** *Suppose $f_n$ converges in probability to $f$. Then, there is a subsequence $f_{n_k}$ of $f_n$ such that $f_{n_k}$ converges almost surely to $f$.*

*Proof.* Let $B_n^\epsilon := \{\omega : |f_n(\omega) - f(\omega)| \geq \epsilon\}$. Then,

$$f_{n_i} \to f \text{ almost surely} \quad \text{iff} \quad P(\bigcap_i \bigcup_{j>i} B_{n_j}^\epsilon) = 0.$$

We know that for any $\epsilon$,

$$\lim_{n\to\infty} P(B_n^\epsilon) = 0.$$

Now, notice that

$$P(\bigcap_n \bigcup_{m\geq n} B_m^\epsilon) \leq \inf_n P(\bigcup_{m\geq n} B_m^\epsilon) \leq \inf_n \sum_{m=n}^{\infty} P(B_m^\epsilon) = \lim_{n\to\infty} \sum_{m=n}^{\infty} P(B_m^\epsilon).$$

Furthermore, $\epsilon_1 < \epsilon_2 \Rightarrow B_n^{\epsilon_1} \supset B_n^{\epsilon_2} \Rightarrow P(B_n^{\epsilon_1}) \geq P(B_n^{\epsilon_2})$.

Let $\delta_i := 1/2^i$. Now, note that $(\forall i)(\exists n_i^\epsilon)(\forall n \geq n_i^\epsilon)(P(B_n^\epsilon) < \delta_i)$. Let $n_i := n_i^{\delta_i}$. Choose $\epsilon \geq 0$. Note, $(\exists m)(\delta_m < \epsilon)$. Hence,

$$P(\bigcap_i \bigcup_{j\geq i} B_{n_j}^\epsilon) \leq \lim_{i\to\infty} \sum_{j=i}^{\infty} P(B_{n_j}^\epsilon) \leq \lim_{i\to\infty} \sum_{j=i}^{\infty} P(B_{n_j}^{\delta_j}) = \lim_{i\to\infty} \sum_{j=i}^{\infty} \delta_j = 0$$

which is what we wanted. □

**Definition 4.5.** A sequence of probability measures $\{\alpha_n\}$ on $\mathbb{R}$ *converges weakly* to $\alpha$ if whenever $\alpha(a) = \alpha(b) = 0$, for $a < b \in \mathbb{R}$, we have

$$\lim_{n\to\infty} \alpha_n[a,b] = \alpha[a,b].$$

A sequence of random variable $\{X_n\}$ converges weakly to $X$ if the induced probability measures $\{\alpha_n\}$ converge weakly to $\alpha$. This is denoted $\alpha_n \Rightarrow \alpha$ or $X_n \Rightarrow X$.

**Lemma 4.6.** *Suppose $\alpha_n$ and $\alpha$ are probability measures on $\mathbb{R}$ with associated distribution functions $F_n$ and $F$. Then, $\alpha_n \Rightarrow \alpha$ iff $F_n(x) \to F(x)$ for each continuity point $x$ of $F$.*

*Proof.* First, note that $x$ is a continuity point of $F$ iff $\alpha(x) = 0$. Let $a < b$ be continuity points of $F$. Suppose $F_n(x) \to F(x)$ for each continuity point $x$ of $F$. Then,

$$\lim_{n\to\infty} \alpha_n[a,b] = \lim_{n\to\infty} F_n(b) - F_n(a) = F(b) - F(a) = \alpha[a,b].$$

For the converse, suppose $\alpha_n \Rightarrow \alpha$. Then,

$$\lim_{n\to\infty} F_n(b) - F_n(a) = \lim_{n\to\infty} \alpha_n[a,b] = \alpha[a,b].$$

Now, we can let $a \to -\infty$ in such a way that $a$ is always a continuity point of $F$. Then, we get, $\lim_n F_n(b) = \alpha(-\infty, b]$. $\square$

The next result shows that weak convergence is actually 'weak':

**Proposition 4.7.** *Suppose $X_n$ converges in probability to $X$. Then, $X_n$ converges weakly to $X$.*

*Proof.* Let $F_n$, $F$ be the distribution functions of $X_n$, $X$ respectively. suppose $x$ is a continuity point of $F$. Note that

$$\{X \le x - \epsilon\} \backslash \{|X_n - X| \ge \epsilon\} \subset \{X_n \le x\}$$

and

$$\begin{aligned} \{X_n \le x\} &= \{X_n \le x \text{ and } X \le x + \epsilon\} \cup \{X_n \le x \text{ and } X > x + \epsilon\} \\ &\subset \{X \le x + \epsilon\} \cup \{|X_n - X| \ge \epsilon\} \end{aligned}$$

Therefore,

$$\begin{aligned} P\{X \le x - \epsilon\} - P\{|X_n - X| \ge \epsilon\} &\le P\{X_n \le x\} \\ &\le P\{X \le x + \epsilon\} + P\{|X_n - X| \ge \epsilon\} \end{aligned}$$

Since for each $\epsilon > 0$, $\lim_n P\{|X_n - X| \ge \epsilon\} = 0$, when we let $n \to \infty$, we have

$$F(x - \epsilon) \le \liminf_{n\to\infty} F_n(x) \le \limsup_{n\to\infty} F_n(x) \le F(x + \epsilon).$$

Finally, since $F$ is continuous at $x$, letting $\epsilon \to 0$, we have

$$\lim_{n\to\infty} F_n(x) = F(x)$$

so that $X_n \Rightarrow X$. $\square$

The converse is not true in general. However, if $X$ is a degenerate distribution (takes a single value with probability one), then the converse is true.

**Proposition 4.8.** *Suppose $X_n \Rightarrow X$, and $X$ is a degenerate distribution such that $P\{X = a\} = 1$. Then, $X_n \xrightarrow{P} X$.*

*Proof.* Let $\alpha_n$ and $\alpha$ be the distributions on $\mathbb{R}$ induced by $X_n$ and $X$ respectively. Given $\epsilon > 0$, we have

$$\lim_{n\to\infty} \alpha_n[a-\epsilon, a+\epsilon] = \alpha[a-\epsilon, a+\epsilon] = 1.$$

Hence,

$$\lim_{n\to\infty} P\{|X_n - X| \le \epsilon\} = 1,$$

and so

$$\lim_{n\to\infty} P\{|X_n - X| > \epsilon\} = 0$$ $\square$

# 5 Product Measures and Independence

Suppose $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ are two measurable spaces. We want to construct a product measurable space with sample space $\Omega_1 \times \Omega_2$.

**Definition 5.1.** Let $\mathcal{A} = \{A \times B : A \in \mathcal{F}_1, B \in \mathcal{F}_2\}$. Let $\mathcal{F}_1 \times \mathcal{F}_2$ be the $\sigma$-field generated by $\mathcal{A}$. $\mathcal{F}_1 \times \mathcal{F}_2$ is called the *product $\sigma$-field* of $\mathcal{F}_1$ and $\mathcal{F}_2$.

If $P_1$ and $P_2$ are probability measures on the measurable spaces above, then $P_1 \times P_2(A \times B) := P_1(A)P_2(B)$ gives a probability measure on $\mathcal{A}$. This can be extended in a canonical way to the $\sigma$-field $\mathcal{F}_1 \times \mathcal{F}_2$.

**Definition 5.2.** $P_1 \times P_2$ is called the *product probability measure* of $P_1$ and $P_2$.

Let $\Omega := \Omega_1 \times \Omega_2$, $\mathcal{F} := \mathcal{F}_1 \times \mathcal{F}_2$, and $P := P_1 \times P_2$.

Note that when calculating integrals with respect to a product probability measure, we can normally perform an iterated integral in any order with respect to the component probability measures. This result is know as *Fubini's Theorem*.

Before we define a notion of independence, we will give some heuristic considerations. Two events $A$ and $B$ should be independent if $A$ occurring has nothing to do with $B$ occurring. If we denote by $P_A(X)$, the probability that $X$ occurs given that $A$ has occurred, then we see that $P_A(X) = \frac{P(A\cap X)}{P(A)}$. Now, suppose that $A$ and $B$ are indeed independent. This means that $P_A(B) = P(B)$. But then, $P(B) = \frac{P(A\cap B)}{P(A)}$, so that $P(A \cap B) = P(A)P(B)$. This leads us to define,

**Definition 5.3.** Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $A_i \in \mathcal{F}$ for every $i$. Let $X_i$ be a random variable for every $i$.

(i) $A_1, \ldots, A_n$ are *independent* if $P(A_1 \cap \cdots \cap A_n) = P(A_1) \cdots P(A_n)$.

(ii) A collection of events $\{A_i\}_{i \in I}$ is *independent* if every finite subcollection is independent.

(iii) $X_1, \ldots, X_n$ are *independent* if for any $n$ sets $A_1, \ldots, A_n \in \mathcal{B}(\mathbb{R})$, the events $\{X_i \in A_i\}_{i=1}^n$ are independent.

(iv) A collection of random variables $\{X_i\}_{i \in I}$ is *independent* if every finite subcollection is independent.

**Lemma 5.4.** *Suppose $X, Y$ are random variables on $(\Omega, \mathcal{F}, P)$, with induced distributions $\alpha, \beta$ on $\mathbb{R}$ respectively. Then, $X$ and $Y$ are independent if and only if the distribution induced on $\mathbb{R}^2$ by $(X, Y)$ is $\alpha \times \beta$.*

**Lemma 5.5.** *Suppose $X, Y$ are independent random variables, and suppose that $f, g$ are measurable functions. Then, $f(X)$ and $g(Y)$ are also independent random variables.*

**Proposition 5.6.** *Let $X, Y$ be independent random variables, and let $f, g$ be measurable functions. Suppose that $E|f(X)|$ and $E|g(Y)|$ are both finite. Then, $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$.*

*Proof.* Let $\alpha$ be the distribution on $\mathbb{R}$ induced by $f(X)$, and let $\beta$ be the distribution induced by $g(Y)$. Then, the distribution on $\mathbb{R}^2$ induced by $(f(X), g(Y))$ is $\alpha \times \beta$. So,

$$\begin{aligned} E[f(X)g(Y)] &= \int_\Omega f(X(\omega))g(Y(\omega))\,\mathrm{d}P = \int_{\mathbb{R}} \int_{\mathbb{R}} uv\,\mathrm{d}\alpha\,\mathrm{d}\beta \\ &= \int_{\mathbb{R}} u\,\mathrm{d}\alpha \int_{\mathbb{R}} v\,\mathrm{d}\beta = E[f(X)]E[g(Y)] \end{aligned}$$

□

# 6 Characteristic Functions

The inverse Fourier transform of a probability distribution plays a central role in probability theory.

**Definition 6.1.** Let $\alpha$ be a probability measure on $\mathbb{R}$. Then, the *characteristic function of* $\alpha$ is

$$\phi_\alpha(t) = \int_{\mathbb{R}} e^{\imath t x} \mathrm{d}\alpha$$

If $X$ is a random variable, the characteristic function of the distribution on $\mathbb{R}$ induced by $X$ will sometimes be denoted $\phi_X$. These results demonstrate the importance of the characteristic function in probability.

**Proposition 6.2.** *Suppose $\alpha$ and $\beta$ are probability measures on $\mathbb{R}$ with characteristic functions $\phi$ and $\psi$ respectively. Suppose further that for each $t \in \mathbb{R}$, $\phi(t) = \psi(t)$. Then, $\alpha = \beta$.*

**Theorem 6.3.** *Let $\alpha_n$, $\alpha$ be probability measures on $\mathbb{R}$ with distribution functions $F_n$ and $F$ and characteristic functions $\phi_n$ and $\phi$. Then, the following are equivalent*

*(i)* $\alpha_n \Rightarrow \alpha$.

*(ii) for any bounded continuous function $f : \mathbb{R} \to \mathbb{R}$,*

$$\lim_{n\to\infty} \int_{\mathbb{R}} f(x)\mathrm{d}\alpha_n = \int_{\mathbb{R}} f(x)\mathrm{d}\alpha.$$

*(iii) for every $t \in \mathbb{R}$,*

$$\lim_{n\to\infty} \phi_n(t) = \phi(t).$$

**Theorem 6.4.** *Suppose $\alpha_n$ is a sequence of probability measures on $\mathbb{R}$, with characteristic functions $\phi_n$. Suppose that for each $t \in \mathbb{R}$, $\lim_n \phi_n(t) =: \phi(t)$ exists and $\phi$ is continuous at 0. Then, there is a probability distriubtion $\alpha$ such that $\phi$ is the characteristic function of $\alpha$. Furthermore, $\alpha_n \Rightarrow \alpha$.*

Next, we show how to recover the moments of a random variable from its characteristic function.

**Definition 6.5.** Suppose $X$ is a random variable. Then, the *kth moment* of $X$ is $EX^k$. The *kth absolute moment* of $X$ is $E|X|^k$.

**Proposition 6.6.** *Let $X$ be a random variable. Suppose that the kth moment of $X$ exists. Then, the characteristic function $\phi$ of $X$ is $k$ times continuously differentiable, and*

$$\phi^{(k)}(0) = \imath^k EX^k.$$

Now, a result on affine transforms of a random variable:

**Proposition 6.7.** *Suppose $X$ is a random variable, and $Y = aX + b$. Let $\phi_X$ and $\phi_Y$ be the characteristic functions of $X$ and $Y$. Then, $\phi_Y(t) = e^{\imath tb}\phi_X(at)$.*

We will often be interested in the sums of independent random variables. Suppose that $X$ and $Y$ are independent random variables with induced distributions $\alpha$ and $\beta$ on $\mathbb{R}$ respectively. Then, the induced distribution of $(X, Y)$ on $\mathbb{R}^2$ is $\alpha \times \beta$. Consider the map $f : \mathbb{R}^2 \to \mathbb{R}$ given by $f(x, y) = x + y$. Then, the distribution on $\mathbb{R}$ induced by $\alpha \times \beta$ is denoted $\alpha * \beta$, and is called the *convolution* of $\alpha$ and $\beta$. $\alpha * \beta$ is the distribution of the sum of $X$ and $Y$.

**Proposition 6.8.** *Suppose $X$ and $Y$ are independent random variables with distributions $\alpha$ and $\beta$ respectively. Then, $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$.*

*Proof.*

$$\begin{aligned}\phi_{\alpha*\beta}(t) &= \int_{\mathbb{R}} e^{\imath tz}\mathrm{d}\alpha * \beta = \int_{\mathbb{R}}\int_{\mathbb{R}} e^{\imath t(x+y)}\mathrm{d}\alpha\,\mathrm{d}\beta \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} e^{\imath tx}e^{\imath ty}\mathrm{d}\alpha\,\mathrm{d}\beta = \int_{\mathbb{R}} e^{\imath tx}\mathrm{d}\alpha \int_{\mathbb{R}} e^{\imath ty}\mathrm{d}\beta \\ &= \phi_\alpha(t)\phi_\beta(t)\end{aligned}$$

□

# 7 Useful Bounds and Inequalities

Here, we will prove some useful bounds regarding random variables and their moments.

**Definition 7.1.** Let $X$ be a random variable. Then, the *variance* of $X$ is $\mathrm{Var}(X) := E[(X - EX)^2] = EX^2 - (EX)^2$.

The variance is a measure of how far spread $X$ is on average from its mean. It exists if $X$ has a finite second moment. It is often denoted $\sigma^2$.

**Lemma 7.2.** *Suppose $X, Y$ are independent random variables. Then,* $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

**Proposition 7.3** (Markov's Inequality)**.** *Let $\epsilon > 0$. Suppose $X$ is a random variable with finite kth absolute moment. Then, $P\{|X| \geq \epsilon\} \leq \frac{1}{\epsilon^k}E|X|^k$.*

*Proof.*

$$P\{|X| \geq \epsilon\} = \int_{\{|X|\geq\epsilon\}} \mathrm{d}P \leq \frac{1}{\epsilon^k}\int_{\{|X|\geq\epsilon\}} |X|^k\mathrm{d}P \leq \frac{1}{\epsilon^k}\int_\Omega |X|^k\mathrm{d}P = \frac{1}{\epsilon^k}E|X|^k$$

□

**Corollary 7.4** (Chebyshev's Inequality)**.** *Suppose $X$ is a random variable with finite 2nd moment. Then,*

$$P\{|X - EX| \geq \epsilon\} \leq \frac{1}{\epsilon^2}\mathrm{Var}(X).$$

The following is also a useful fact:

**Lemma 7.5.** *Suppose $X$ is a nonnegative random variable. Then,*

$$\sum_{m=1}^{\infty} P\{X \geq m\} \leq EX$$

*Proof.*

$$\begin{aligned}E\lfloor X\rfloor &= \sum_{n=1}^{\infty} nP\{n \leq X < n+1\} = \sum_{m=1}^{\infty}\sum_{n=m}^{\infty} P\{n \leq X < n+1\} \\ &= \sum_{m=1}^{\infty} P\{X \geq m\} \leq EX\end{aligned}$$

□

# 8 The Borel-Cantelli Lemma

First, let us introduce some terminology. Let $A_1, A_2, \ldots$ be sets. Then,

$$\limsup_{n\to\infty} A_n := \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m.$$

$\limsup_n A_n$ consists of those $\omega$ that appear in $A_n$ infinitely often (i.o.). Also,

$$\liminf_{n\to\infty} A_n := \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_n.$$

$\liminf_n A_n$ consists of those $\omega$ that appear in all but finitely many $A_n$.

**Theorem 8.1** (Borel-Cantelli Lemma)**.** *Let $A_1, A_2, \ldots \in \mathcal{F}$. If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $P(\limsup_n A_n) = 0$. Furthermore, suppose that the $A_i$ are independent. Then, if $\sum_{n=1}^{\infty} P(A_n) = \infty$, then $P(\limsup_n A_n) = 1$.*

*Proof.* Suppose $\sum_{n=1}^{\infty} P(A_n) < \infty$. Then,

$$P\left(\bigcap_{n=1}^{\infty} \bigcup m = n^{\infty} A_m\right) = \lim_{n\to\infty} P\left(\bigcup_{m=n}^{\infty} A_m\right) \leq \lim_{n\to\infty} \sum_{m=n}^{\infty} P(A_m) = 0.$$

For the converse, it is enough to show that

$$P\left(\bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m^c\right) = 0,$$

and so it is also enough to show that

$$P\left(\bigcap_{m=n}^{\infty} A_m^c\right) = 0$$

for all $n$. By independence, and since $1 - x \leq e^{-x}$, we have

$$P\left(\bigcap_{m=n}^{\infty} A_m^c\right) \leq P\left(\bigcap_{m=n}^{n+k} A_m^c\right) = \prod_{m=n}^{n+k} (1 - P(A_m)) \leq \exp\left\{-\sum_{m=n}^{n+k} P(A_m)\right\}$$

Since the last sum diverges, taking the limit as $k \to \infty$, we get

$$P\left(\bigcap_{m=n}^{\infty} A_m^c\right) = 0 \qquad \square$$

# 9 The Law of Large Numbers

Let $X_1, X_2, \ldots$ be random variables that are independent and indentically distributed (iid). Let $S_n := X_1 + \cdots + X_n$. We will be interested in the asymptotic behavior of the average $\frac{S_n}{n}$. If $X_i$ has a finite expectation, then we would think that $\frac{S_n}{n}$ would settle down to $EX_i$. This is known as the Law of Large Numbers. There are two varieties of this law: the Weak Law of Large Numbers and the Strong Law of Large Numbers. The weak law states that the average converges in probability to $EX_i$. The strong law, however states that the average converges almost surely to $EX_i$. However, the strong law is significantly harder to prove, and requires a bit of additional machinery. For the rest of this section, fix a probability space $(\Omega, \mathcal{F}, P)$.

**Theorem 9.1** (The Weak Law of Large Numbers)**.** *Suppose $X_1, X_2, \ldots$ are iid random variables with mean $EX_i = m < \infty$. Then, $\frac{S_n}{n} \to_P m$.*

*Proof.* Let $\phi$ be the characteristic function of $X_i$. Then, the characteristic function of $S_n$ is $[\phi(t)]^n$. Then, by 6.7, the characteristic function of $\frac{S_n}{n}$ is $\psi_n(t) = [\phi(\frac{t}{n})]^n$. Furthermore, by 6.6, $\phi$ is differentiable, and $\phi'(0) = \imath m$. Therefore, we can form the Taylor expansion,

$$\phi\left(\frac{t}{n}\right) = 1 + \frac{\imath m t}{n} + o\left(\frac{1}{n}\right),$$

and so

$$\psi_n(t) = \left[1 + \frac{\imath m t}{n} + o\left(\frac{1}{n}\right)\right]^n.$$

Taking the limit as $n \to \infty$, we get

$$\lim_{n\to\infty} \psi_n(t) = e^{\imath m t}$$

which is the characteristic function for the distribution degenerate at $m$. Therefore, by Proposition 4.8, $\frac{S_n}{n}$ converges in probability to $m$. $\square$

**Theorem 9.2** (The Strong Law of Large Numbers)**.** *Suppose $X_1, X_2, \ldots$ are iid random variables with $EX_i = m < \infty$. Let $S_n = X_1 + \cdots + X_n$. Then, $\frac{S_n}{n}$ converges to $m$ almost surely.*

*Proof.* We can decompose an arbitrary random variable $X_i$ into its positive and negative parts: $X_i^+ := X_i I_{\{X_i \geq 0\}}$ and $X_i^- := -X_i I_{\{X_i < 0\}}$, so that $X_i = X_i^+ - X_i^-$. Then, we have $S_n = X_1^+ + \cdots + X_n^+ - (X_1^- + \cdots + X_n^-) =: S_n^+ - S_n^-$. Hence, it is enough to prove the Theorem for nonnegative $X_i$.

Now, Let $Y_i := X_i I_{\{X_i \leq i\}}$. Let $S_n^* := Y_1 + \ldots Y_n$. Furthermore, let $\alpha > 1$, and set $u_n := \lfloor \alpha^n \rfloor$. We shall first establish the inequality

$$\sum_{n=1}^{\infty} P\left\{\left|\frac{S_{u_n}^* - ES_{u_n}^*}{u_n}\right| \geq \epsilon\right\} < \infty \qquad \forall \epsilon > 0 \tag{9.1}$$

Since the $X_i$ are independent, we have

$$\mathrm{Var}(S_n^*) = \sum_{k=1}^{n} \mathrm{Var}(Y_k) \leq \sum_{k=1}^{n} EY_k^2 \leq \sum_{k=1}^{n} E[X_i^2 I_{\{X_i \leq k\}}] \leq nE[X_i^2 I_{\{X_i \leq n\}}]$$

By Chebyshev's inequality, we have

$$\begin{aligned} \sum_{n=1}^{\infty} P\left\{ \left| \frac{S_{u_n}^* - ES_{u_n}^*}{u_n} \right| \geq \epsilon \right\} &\leq \sum_{n=1}^{\infty} \frac{\mathrm{Var}(S_{u_n}^*)}{\epsilon^2 u_n^2} \\ &\leq \frac{1}{\epsilon^2} \sum_{n=1}^{\infty} \frac{E[X_i^2 I_{\{X_i \leq u_n\}}]}{u_n} \\ &= \frac{1}{\epsilon^2} E\left[ X_i^2 \sum_{n=1}^{\infty} \frac{1}{u_n} I_{\{X_i \leq u_n\}} \right] \qquad (9.2) \end{aligned}$$

Now, let $K := \frac{2\alpha}{\alpha - 1}$. Let $x > 0$, and let $N := \inf\{n : u_n \geq x\}$. Then, $\alpha^N \geq x$. Also, note that $\alpha^n \leq 2u_n$, and so $u^{-n} \leq 2\alpha^{-n}$. Then,

$$\sum_{u_n \geq x} \frac{1}{u_n} \leq 2 \sum_{n \geq N} \frac{1}{\alpha^n} = 2\alpha^{-N} \sum_{n=0}^{\infty} \left( \frac{1}{\alpha} \right)^n = K\alpha^{-N} \leq Kx^{-1},$$

and hence,

$$\sum_{n=1}^{\infty} \frac{1}{u_n} I_{\{X_i \leq u_n\}} \leq KX_1^{-1} \qquad \text{if } X_1 > 0$$

and so, putting this into (9.2), we get

$$\frac{1}{\epsilon^2} E\left[ X_i^2 \sum_{n=1}^{\infty} \frac{1}{u_n} I_{\{X_i \leq u_n\}} \right] \leq \frac{1}{\epsilon^2} E\left[ X_i^2 K X_i^{-1} \right] = \frac{K}{\epsilon^2} EX_i < \infty$$

thereby establishing inequality (9.1).

Therefore, by the Borel-Cantelli Lemma, we have

$$P\left( \limsup_n \left\{ \left| \frac{S_{u_n}^* - ES_{u_n}^*}{u_n} \right| \geq \epsilon \right\} \right) = 0 \qquad \forall \epsilon > 0.$$

Taking an intersection over all rational $\epsilon$, we get that

$$\frac{S_{u_n}^* - ES_{u_n}^*}{u_n} \to 0 \qquad \text{almost surely.}$$

However, $\frac{1}{n} ES_n^* = \frac{1}{n} \sum_{k=1}^{n} EY_k$, and since $EY_k \to EX_i$, taking the limit as $n \to \infty$, we have that $\frac{1}{n} ES_n^* \to EX_i$. Therefore, we have that

$$\frac{S_{u_n}^*}{u_n} \to EX_i \qquad \text{almost surely.} \qquad (9.3)$$

Now, notice that by Lemma 7.5,

$$\sum_{n=1}^{\infty} P\{X_n \neq Y_n\} = \sum_{n=1}^{\infty} P\{X_i > n\} \leq EX_i < \infty$$

Again, by the Borel-Cantelli Lemma, we have

$$P\left(\limsup_n \{X_n \neq Y_n\}\right) = 0.$$

Therefore, $\frac{S_n^* - S_n}{n} \to 0$ almost surely, and so by (9.3),

$$\frac{S_{u_n}}{u_n} \to EX_i \qquad \text{almost surely.} \tag{9.4}$$

Now, to get that the entire sequence $\frac{S_n}{n} \to EX_i$ almost surely, note that $S_m$ is an increasing sequence. Suppose $u_n \leq k \leq u_{n+1}$. Then,

$$\frac{u_n}{u_{n+1}} \frac{S_{u_n}}{u_n} \leq \frac{S_k}{k} \leq \frac{u_{n+1}}{u_n} \frac{S_{u_{n+1}}}{u_{n+1}}$$

and so,

$$\frac{1}{\alpha} EX_i \leq \liminf_k \frac{S_k}{k} \leq \limsup_k \frac{S_k}{k} \leq \alpha EX_i \qquad \text{almost surely.}$$

Taking $\alpha \to 1$, we get by (9.4)

$$\lim_{k \to \infty} \frac{S_k}{k} = EX_i \qquad \text{almost surely} \qquad \square$$

# 10 Conditional Expectation and Probability

Before defining condtional expectation and probability, we will make a few observations about the probabalistic interpretation of $\sigma$-fields.

Consider a process where a random number between zero and one is chosen. More precisely, an outcome $\omega$ is chosen according to some probability law from the set of all possible outcomes, $\Omega = [0, 1)$. We may be able to observe this number $\omega$ to some amount of precision, say up to one digit. The $\sigma$-field that represents this amount of precision is $\mathcal{F}_1 := \sigma\{[0, .1), [.1, .2), \ldots, [.9, 1)\}$. The $\sigma$-field $\mathcal{F}_1$ represents all the information that we can know about $\omega$ by observing it to one digit of precision. That is, an observer who can observe the number $\omega$ to one digit will be able to determine exactly which sets $A \in \mathcal{F}_1$ that $\omega$ belongs to, but he will not be able to give any information more precise than that. Similarly, if we can observe $\omega$ up to $n$ digits of precision, the $\sigma$-field which corresponds to this is: $\mathcal{F}_n := \sigma\left\{\left[\frac{i}{10^n}, \frac{i+1}{10^n}\right) : 0 \leq i < 10^n\right\}$.

This example illustrates a general concept: The $\sigma$-field that is used represents the amount of *information* that an observer has about the random process.

**Definition 10.1.** If $\mathcal{F}$ is a $\sigma$-field, a *$\mathcal{F}$-observer* is an observer who can determine precisely which sets $A \in \mathcal{F}$ that a random outcome $\omega$ belongs to but has no more information about $\omega$.

Therefore, a $2^\Omega$-observer has complete information about the outcome $\omega$, whereas a $\mathcal{F}$-observer has less information. Similarly, if $\Sigma \subset \mathcal{F}$, then a $\mathcal{F}$-observer has more information than a $\Sigma$-observer.

Suppose that a random variable $X$ is $\mathcal{F}$-measurable. This means that the preimage of any Borel set under $X$ in $\mathcal{F}$. Therefore, a $\mathcal{F}$-observer will have complete information about $X$, or any other $\mathcal{F}$-measurable random variable. Note that if $\Sigma \subset \mathcal{F}$, a $\Sigma$-measurable function is also $\mathcal{F}$-measurable.

Suppose that $X$ is a $\mathcal{F}$-measurable random variable, and that you are a $\Sigma$-observer. You do not have complete information about $X$. However, given your information $\Sigma$, you would like to make a "buest guess" about the value of $X$. That is, you want to create another random variable, $Y$, that is $\Sigma$-measurable, but which approximates $X$. $Y$ is called the *conditional expectation of $X$ wrt $\Sigma$*, and is denoted $E[X|\Sigma]$.

We will require that

$$\int_A X(\omega)\mathrm{d}P = \int_A E[X|\Sigma](\omega)\mathrm{d}P \qquad \text{for all } A \in \Sigma \tag{10.1}$$

**Lemma 10.2.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $\Sigma$ be a sub-$\sigma$-field of $\mathcal{F}$. Let $P_{|\Sigma}$ denote the restriction of $P$ to $\Sigma$. Suppose $f$ is a $\Sigma$-measurable function and $A \in \Sigma$. Then,*

$$\int_A f(\omega)\mathrm{d}P_{|\Sigma} = \int_A f(\omega)\mathrm{d}P.$$

Justified by the previous lemma, we will often be sloppy and not explicitly say which $\sigma$-field a particular integral is taken over. In order to prove that a function satisfying (10.1) exists, we will have to discuss the Radon-Nikodym Theorem. First, a definition.

**Definition 10.3.** A *signed measure* $\lambda$ on a measurable space $(\Omega, \mathcal{F})$ is a function $\lambda : \mathcal{F} \to \mathbb{R}$ such that whenever $A_1, A_2, \ldots$ is a finite or countable sequence of disjoint sets in $\mathcal{F}$, we have

$$\lambda\left(\bigcup_i A_i\right) = \sum_i \lambda(A_i)$$

In particular, we have for a signed measure, $\lambda(\emptyset) = 0$. All probability measures are also signed measures. Note that $\lambda$ is permitted to take on negative values. However, it is not permitted to take on the values $+\infty$ or $-\infty$.

**Definition 10.4.** a signed measure $\lambda$ on $(\Omega, \mathcal{F})$ is *absolutely continuous* with respect to a probability measure $P$ if, whenever $P(A) = 0$, we have also $\lambda(A) = 0$. This is denoted $\lambda \ll P$.

For example, if $f$ is an integrable function wrt $P$, then $\lambda(A) = \int_A f(\omega)\mathrm{d}P$ is a signed measure that is absolutely continuous with respect to $P$. In fact, all absolutely continuous signed measures arise in this way.

**Theorem 10.5** (Radon-Nikodym)**.** *Suppose $\lambda \ll P$. Then, there is an integrable function $f$ such that*

$$\lambda(A) = \int_A f(\omega)\mathrm{d}P. \tag{10.2}$$

*Furthermore, if $f'$ is another function satisfying* (10.2)*, then $f = f'$ $P$-almost-everywhere.*

**Definition 10.6.** The function $f$ in Theorem 10.5 is called the *Radon-Nikodym derivative* of $\lambda$ with respect to $P$. It is denoted $\frac{\mathrm{d}\lambda}{\mathrm{d}P}$.

Note that the Radom-Nikodym derivative is only defined up to equality almost everywhere. We can use the Radon-Nikodym derivative to define the conditional expectation satisfying (10.1).

**Definition 10.7.** Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $\Sigma$ be a sub-$\sigma$-field of $\mathcal{F}$. Let $X$ be a $\mathcal{F}$-integrable random variable. Let $\lambda$ be the signed measure defined by $\lambda(A) = \int_A X(\omega)\mathrm{d}P$. The *conditional expectation of $X$ wrt $\Sigma$* is

$$E[X|\Sigma] := \frac{\mathrm{d}\lambda_{|\Sigma}}{\mathrm{d}P_{|\Sigma}}.$$

We now state some of the elementary properties of conditional expectation.

**Lemma 10.8.** *Let $X$ and $X_i$ be random variables on $(\Omega, \mathcal{F}, P)$. Let $\Sigma$ be a sub-$\sigma$-field of $\mathcal{F}$.*

*(i)* $E[E[X|\Sigma]] = E[X]$

*(ii) If $X$ is nonnegative, then $E[X|\Sigma]$ is nonnegative almost surely.*

*(iii) Suppose $a_1, a_2 \in \mathbb{R}$. Then*

$$E[a_1X_1 + a_2X_2|\Sigma] = a_1E[X_1|\Sigma] + a_2E[X_2|\Sigma] \qquad \textit{almost surely.}$$

*(iv)* $\int |E[X|\Sigma]|\mathrm{d}P \leq \int |X|\mathrm{d}P$.

*(v) If $Y$ is bounded and $\Sigma$-measurable, then $E[XY|\Sigma] = YE[X|\Sigma]$ almost surely.*

*(vi) If $\Sigma_2 \subset \Sigma_1 \subset \mathcal{F}$ are sub-$\sigma$-fields, then $E[X|\Sigma_2] = E[E[X|\Sigma_1]|\Sigma_2]$ almost surely.*

As a special case of conditional expectation, we have conditional probability.

**Definition 10.9.** Let $(\Omega, \mathcal{F}, P)$ be a probability space, and let $\Sigma$ be a sub-$\sigma$-field of $\mathcal{F}$. Then, the *conditional probability* of an event $A \in \mathcal{F}$ given $\Sigma$ is $P[A|\Sigma] := E[I_A|\Sigma]$.

$P[A|\Sigma](\omega)$ is also sometimes written $P(\omega, A)$. We now state some of the elementary properties of conditional probability.

**Lemma 10.10.** *The following hold almost surely:*

*(i)* $P(\omega, \Omega) = 1$ *and* $P(\omega, \emptyset)$.

*(ii) For any* $A \in \mathcal{F}$, $0 \leq P(\omega, A) \leq 1$.

*(iii) If* $A_1, A_2 \ldots$ *is a finite or countable sequence of disjoint sets in* $\mathcal{F}$, *then*

$$P\left(\omega, \bigcup_i A_i\right) = \sum_i P(\omega, A_i).$$

*(iv) If* $A \in \Sigma$, *then* $P(\omega, A) = I_A(\omega)$.

Lemma 10.10 in particular implies that given $\omega \in \Omega$, $P(\omega, \cdot)$ is a probability measure on $(\Omega, \mathcal{F})$.

# References


[1] P. Billingsley, *Probability and measure*, John Wiley & Sons, Inc., 1995.

[2] S.R.S. Varadhan, *Probability theory*, Courant lecture notes, 7. American Mathematical Society, 2001.