Notes on Measure and Probability

(BIO 250: Probability II)

Yi Li
Biostatistics Department
Harvard University

Contents

Acknowledgement

This note is based on the notes written and used by L.J. Wei and Robert Gray in 2000 when this course was first offered in Harvard Biostatistics department. A lot of materials, in fact, are verbatim from these sources. Some other good references include

Ash, R. (1972) *Real Analysis and Probability*, Academic Press.

Billingsley, P. (1995). *Probability and Measure, 3rd Ed.*. Wiley.

Chung, K.L. (2001). *A Course in Probability Theory, 3rd Ed.*. Academic Press.

Feller, W. (1966). *An Introduction to Probability Theory and Applications, 1st Ed.* Wiley.

Shiryave (1996). *Probability, 2nd Ed.* Springer-Verlag.

The intention of this note is to list and highlight some important definitions and theorems in probability theory. No attempt has been made to include detailed proof for each theorem or claim, however, some parts or outlines of proofs are given if they help illustrate some particular points.

1 Some Set Theory

1.1 Set and Sample Space

Set is a basic concept in mathematics and probability. We define a *set*, often denoted by a capital letter in this note, as a collection of some elements. For example, R and R^n denote the set of real numbers and the n-dimensional Euclidean space, respectively.

Some commonly used notations involving sets include:

- ϕ: empty set.
- $x \in A$: x is an element of the set A.
- $\{x\} \subset A$: the set consisting of the singleton $x \in A$ is a subset of A.
- $\{x :$ a statement$\}$: the set of all elements x for which the statement holds. For example: the open interval (a, b) can be defined as $\{x : a < x < b\}$.

It is necessary to define an 'abstract space', often denoted by Ω, as a nonempty set of all the elements concerned. These elements are called 'points' and denoted usually by ω. A set containing only part of these elements is called a *subset* (of Ω). In the probability literature, we often use Ω to denote the sample space, which is the collection of all possible distinct realizations of a non-deterministic experiment and an element of Ω, say ω, is called a *simple event* or a *sample point* in Ω. We shall decide in latter sections on what class of (sub)sets probabilities are defined for and on what class of functions are acceptable for random variables.

The choice of the sample space is the first step in formulating a probabilistic model for an experiment. Let us consider several examples of sample space.

(1) A patient's survival status at the end of a clinical trial.
$\Omega = \{dead, alive\}$, which is a finite sample space.

(2) Number of a patient's seizures observed in a clinical trial.
$\Omega = \{0, 1, 2, 3, \ldots, \}$, which is a countably infinitive sample space.

(3) A cancer patient's survival time after treatment.
$\Omega = \{T : \ T \geq 0\}$, which is an uncountably infinitive sample space.

Exercise 1.1 Suppose Ω has exactly n sample points. Find the number of all possible subsets of Ω.

Exercise 1.2 (courtesy of R. Strawderman) Describe the elements of the sample space for the following experiment: Players A, B, and C take turns at a game subject to the following restrictions:

- To start the game, A and B play while C sits out.
- The loser of the 1st round sits out and is replaced by C in the 2nd round (i.e. the 2nd round is played by the winner of the 1st round and player C).
- The game continues in a similar fashion (i.e. the loser is replaced by the one who is sitting out) until one player wins two successive rounds.

1.2 Revisit of Set Calculations

Roughly speaking, subsets of Ω are called events. If Ω is uncountable, we cannot (need not) handle all possible subsets. Instead, we restrict to a "well-behaved" class of subsets, broad enough for our purpose. Only those subsets in such a restricted class will be called events. For example, we may require the interaction (or union) of events is also an event. In this section, we review the basic calculations involving in sets.

Definition: For any two sets in Ω, we define

(1) Union: $A \cup B = \{\ w :\ w \in A \text{ or } w \in B\}$.

(2) Intersection: $A \cap B = \{\ w :\ w \in A \text{ and } w \in B\}$.

(3) Complementation (with respect to Ω): $A^c = \{\ w \in \Omega :\ w \notin A\}$.

(4) Difference: $A - B = \{\ w :\ w \in A,\ w \notin B\} = A \cap B^c$.

(5) Symmetric Difference:
$A \triangle B = \ (A - B) \cup (B - A) = \ \{w :\ w \in$ exactly one of A and $B\}$.

It is straightforward to show that the operation of union has the following properties.

Property 1.1 associative: $(A_1 \cup A_2) \cup A_3 = A_1 \cup (A_2 \cup A_3)$.

Property 1.2 distributive: $(A_1 \cup A_2) \cap A_3 = (A_1 \cap A_3) \cup (A_1 \cap A_3)$.

Property 1.3 commutative: $A_1 \cup A_2 = A_2 \cup A_1$.

If A and B are disjoint, i.e. $A \cap B = \phi$, we sometimes write the disjoint union as

$$A \cup B = A + B.$$

Exercise 1.3 Find A^c, with respect to Ω,

(a) $\Omega = \{x : 0 < x < 1\}, A = \{x : 0.5 < x < 1\}$;

(b) $\Omega = \{(x, y) : |x| + |y| \leq 2\}, A = \{(x, y) : x^2 + y^2 < 2\}$;

(c) $\Omega = R^1, A = \cap_{n=1}^{\infty} B_n$, where $B_n = \{x : x \in (0, 1/n)\}$.

We say two sets are equal if they contain exactly the same elements. Hence, to show $A = B$, one needs to demonstrate in two steps that $A \subset B$ and $B \subset A$.

Exercise 1.4 For any three sets, A, B and C, show $A \triangle B = C$ if and only if $A = B \triangle C$.

We define countable infinite unions and intersections as follows.

Definition: Let $\{A_n\}$ be an infinite sequence of sets in Ω. We define
$\sup_{n \geq 1} A_n = \bigcup_{n=1}^{\infty} A_n = \{w : \ w \in A_n \text{ for some } n\}$ and
$\inf_{n \geq 1} A_n = \bigcap_{n=1}^{\infty} A_n = \{w : \ w \in A_n \text{ for any } n \geq 1\}$.

Similarly, we often use $\sum_{n=1}^{\infty} A_n$ to denote a union of a countable sequence of pairwise disjoint sets. Furthermore, we may extend the intersection or union over a set of integers to any arbitrary set. For $\{A_t, \ t \in T\}$, where T is an index set, $\bigcup_{t \in T} A_t = \{w : \ w \in A_t \text{ for some } t \in T\}$ and similarly for the intersection.

Example 1.1 Let $T = [0, 1]$ and $A_t = \{t + 1\}$. Then $\bigcup_{t\in T} A_t = [1, 2]$.

It may be easy to prove

Theorem 1.1 *(De Morgan's rule)* $(\bigcup_{t\in T} A_t)^c = \bigcap_{t\in T} A_t^c$ *and* $(\bigcap_{t\in T} A_t)^c = \bigcup_{t\in T} A_t^c$*, where* T *is any index set, e.g. finite, countably infinite or uncountably infinite.*

Proof: homework. □

One can also easily show, under complementation, $\subset$ and $\supset$ are interchanged, i.e. $A \subset B$ implies $A^c \supset B^c$.

Exercise 1.5 Show

$$B \bigcap \left(\bigcup_{n=1}^{\infty} A_n \right) = \bigcup_{n=1}^{\infty} (B \cap A_n) \text{ and } B \bigcap \left(\bigcap_{n=1}^{\infty} A_n \right) = \bigcap_{n=1}^{\infty} (B \cap A_n).$$

1.3 Limit Sets

For an infinite sequence of real numbers, we know how to study its convergence and define its limit (if it exists). In this section, we discuss the convergence of an infinite series of sets. We begin with the concepts of "liminf" and "limsup" for a sequence of sets.

Definition: Let $\{A_n\}$ be a sequence of sets in Ω. Define

$$A^* = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$$

and

$$A_* = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k.$$

A^* and A_* are termed *upper limit* and *lower limit* of the sequence $\{A_n\}$, usually denoted by $\limsup_n A_n$ and $\liminf_n A_n$ respectively. If $A_* = A^*$, we say $\{A_n\}$ is convergent and write $\lim_{n\to\infty} A_n = A_* = A^*$.

Example 1.2 Let $\Omega = [0,1]$, $A_n = [0, \frac{1}{n}]$ if n is even and $A_n = [1 - \frac{1}{n}, 1)$ if n is odd. Then by definition, $A_* = \phi$ and $A^* = \{0\}$. Hence, $A_* \neq A^*$, which implies $\lim A_n$ doesn't exist.

However, if $A_n = [0, \frac{1}{n}]$ for all n, then $A_* = A^* = \{0\}$, which implies $\lim A_n = \{0\}$.

Theorem 1.2 *(a) $\omega \in \limsup_n A_n$ if and only if ω is in infinitely many of the A_n. (b) $\omega \in \liminf_n A_n$ if and only if there is an m such that $\omega \in A_n$ for all $n \geq m$.*

Proof: (a) Set $B_m = \bigcup_{n=m}^{\infty} A_n$, and note that $\limsup_n A_n = \bigcap_m B_m$. Suppose ω belongs to infinitely many of the A_n. Then for any m, there is an $n > m$ such that $\omega \in A_n \subset B_m$, so $\omega \in \bigcap_m B_m$. On the other hand, suppose that $\omega \in \limsup_n A_n$, and that ω is in only a finite number of the A_n. Then there is an N such that $\omega \notin A_n$ for all $n > N$, which implies $\omega \notin B_m$ for $m > N$, so $\omega \notin \bigcap_m B_m$, which is a contradiction. Thus ω must be in infinitely many of the A_n.
(b) Let $C_m = \bigcap_{n=m}^{\infty} A_n$, and note that $\bigcup_{m=1}^{\infty} C_m = \liminf_n A_n$. If $\omega \in A_n$ for all $n \geq m$, then $\omega \in C_m \subset \liminf_n A_n$. On the other hand, if $\omega \notin C_m$ for any m, then $\omega \notin \bigcup_{m=1}^{\infty} C_m$. So if $\omega \in \liminf_n A_n$ then ω must be in C_m for some m, and thus in A_n for all $n \geq m$. □

Because of part (a), the terminology 'A_n occurs infinitely often' is often used to refer to $\limsup_n A_n$. For example, suppose $X_n(\omega)$ is the number of heads in n flips of a coin, and $A_n = \{\omega : X_n(\omega) > n/2\}$, which consists of all outcomes where the proportion of heads in the first n flips is $> 1/2$. In this case, $\omega \in \limsup_n A_n$ if $X_n(\omega)/n > 1/2$ for infinitely many n, or equivalently, if the proportion of heads is $> 1/2$ infinitely often. On the other hand, if $\omega \in \liminf_n A_n$, then ω must be in all but a finite number of the A_n. In the example, this means that there is an m (which can be different for different ω), such that $X_n(\omega)/n > 1/2$ for all $n \geq m$. Equivalently, the proportion of heads is $> 1/2$ for all but a finite number of n.

Exercise 1.6 Show $(\limsup_n A_n) \bigcap (\limsup_n B_n) \supset \limsup_n (A_n \bigcap B_n)$ and $(\limsup_n A_n) \bigcup (\limsup_n B_n) = \limsup_n (A_n \bigcup B_n)$.

Using the De Morgan's rule, one may show $\liminf_n A_n = (\limsup_n A_n^c)^c$. In fact,

$$\liminf_n A_n = \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} A_n = \left(\bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n^c \right)^c = \left(\limsup_n A_n^c \right)^c.$$

Exercise 1.7 Show $\liminf_n A_n \subset \limsup_n A_n$.

Intuitively, this follows as if w lies in all but a finite number of members of the sequence, it surely lies in infinitely many members of the sequence.

Exercise 1.8 Let $\Omega = R^2$ and A_n the interior of the circle with center at $((-1)^n/n, 0)$ and radius 1. Find $\liminf_n A_n$ and $\limsup_n A_n$.

Exercise 1.9 Let a_n be a sequence of real numbers, and let $A_n = (-\infty, a_n)$. Find the connection between $\limsup_n a_n$ and $\limsup_n A_n$, and similarly for $\liminf$ (Recall that, for a sequence of real numbers $\{a_n\}$, $\liminf_n a_n = \lim_{n\to\infty} inf_{k\geq n}\{a_k\}$ and $\limsup_n a_n = \lim_{n\to\infty} sup_{k\geq n}\{a_k\}$.)

Parallel to a series of real numbers, we also introduce the concept of monotonicity to a sequence of sets.

Definition: A monotone sequence of sets is defined as follows

- $\{A_n\}$ is called monotone increasing if and only if $A_n \subset A_{n+1}$ for any n.
- $\{A_n\}$ is called monotone decreasing if and only if $A_n \supset A_{n+1}$ for any n.

Then we prove

Theorem 1.3 *A monotone sequence of sets is convergent.*

Proof: First suppose that A_n is increasing. Then

$$\liminf_n A_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k = \bigcup_{n=1}^{\infty} A_n.$$

On the other hand,

$$\limsup A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k \subset \bigcup_{k=n}^{\infty} A_k \subset \bigcup_{n=1}^{\infty} A_n = B = A_*.$$

As $A_* \subset A^*$, we have that $A_* = A^*$.

Now we suppose that A_n is decreasing. By definition, we have that

$$A_* = A^* = \bigcap_{n=1}^{\infty} A_n,$$

which completes the proof. □

1.4 Zorn's Lemma and Zermelo Choice Axiom

In this section, we introduce some fundamental set theorems.

Definition: A partial ordering on a set is a relation "$\leq$" that is

- reflexive: $a \leq a$.
- antisymmetric: if $a \leq b$, $b \leq a$, then $a = b$.
- transitive: if $a \leq b$, $b \leq c$, then $a \leq c$.

Definition: A set equipped with such a relation "$\leq$" is called a *partially ordered* set.

Note "$a \leq b$" can also be written as "$b \leq a$" and "$a < b$" means "$a \leq b$ but $a \neq b$".

Example 1.3 Let Ω be a nonempty set and denote by $\mathcal{X}$ the class of all the subsets of Ω. It follows that $\mathcal{X}$ is a partially ordered set by the ordinary set inclusion relationship, That is, for any $A, B \in \mathcal{X}$, $A \leq B$ means $A \subset B$.

Definition: A set C is called *totally ordered* if and only if for all $a, b \in C$, either $a \leq b$ or $b \leq c$. A totally ordered subset of a partially ordered set A is called a *chain* of A.

Example 1.4 Any subset of R is totally ordered if taking the ordinary '$\leq$' relationship between two real numbers. How about the two-dimensional real space?

Definition: Suppose that X is a partially ordered set and that X_0 is a subset of X. An element b is called the *upper bound* of X_0 if that $b \in X$ and any $x \in X_0$

implies $x \leq b$. If b is an upper bound of X_0 and for any upper bound b', $b \leq b'$, then b is called the supremum of X_0. In other words, the supremum of X_0 is the smallest upper bound of X_0 (in the sense of " $\leq$ ").

Remark 1.1 The upper bound and supremum of X_0 are not necessarily in X_0. The lower bound and infimum can be defined similarly. The supremum and infimum of X_0 are usually denoted by $\sup X_0$ and $\inf X_0$.

Example 1.5 Suppose that X_0 is a class of some subsets of X. Then $\sup X_0 = \bigcup_{A \in X_0} A, \inf X_0 = \bigcap_{A \in X_0} A$.

We state below without proof a theorem needed in the further development of basic set theorems.

Theorem 1.4 *Let X be a non-empty partially ordered set and assume that every non-empty chain of X has a supremum. Further, let a mapping $f : X \to X$ satisfy $x \leq f(x)\,(x \in X)$, then there exists $c \in X$ such that $f(c) = c$.*

The proof of this theorem is rather involved. Interested readers should refer to P.J Cohen's *Set theory and the Continuum Hypothesis* (New York, 1966, Amsterdam).

Definition: Let X be a partially ordered set and $x \in X$. If any $y \in X$ such that $x \leq y$ implies $x = y$, then x is called the maximum of X. The minimum of X is defined similarly.

Remark 1.2 The maximum or the minimum of a set may not be unique.

Example 1.6 Let Ω be a nonempty set and denote by $\mathcal{X}$ the class of all the subsets of Ω. It follows that $\mathcal{X}$ is a partially ordered set by the ordinary set inclusion relationship. The maximum of $\mathcal{X}$ is Ω and the minimum of $\mathcal{X}$ is ϕ, the empty set. Consider a subset of $\mathcal{X}$, $\mathcal{X}_1 = \mathcal{X} - \{\phi\}$. It is easy to show that every singleton generated by Ω is the minimum of $\mathcal{X}_1$.

Theorem 1.5 *Every partially ordered set has a maximum chain.*

Proof: Let X be a partially ordered set with the relation "$\leq$". Let $\mathcal{X}$ be the class of all the chains of X. It follows that $\mathcal{X}$ itself is a partially ordered set by the set inclusion relationship (i.e. $\subset$). We prove by contradiction that $\mathcal{X}$ has a maximum.

Assume that $\mathcal{X}$ does not have a maximum. Then for any $A \in \mathcal{X}$, as A is not a maximum, there exists an A_1 such that $A \subset A_1$ and $A \neq A_1$. Let f be a mapping for such a realization, i.e. $f : \mathcal{X} \to \mathcal{X}$ satisfying $A \subset f(A)$ and $f(A) \neq A$. In addition, every non-empty chain of $\mathcal{X}$ has a supremum (think of the union!). Hence, by Theorem 1.4, there exists an $A_0 \in \mathcal{X}$ such that $f(A_0) = A_0$, which contradicts to the definition of f! The theorem is thus proved. □

Now we can easily apply the result to prove Zorn's lemma, one of the fundamental theorems in the set theory.

Theorem 1.6 *(Zorn's lemma) Let X be a set with a partial ordering '$\leq$'. If every chain in X has a supremum, then X has a maximum.*

Proof: By Theorem 4, there exists a maximal chain X_0 of X. Let $x_0 = sup X_0$. For any $x \in X$ such that $x_0 \leq x$, we would like to show $x_0 = x$. If $x \notin X_0$, then $X_0 \cup \{x\}$ is a chain containing X_0, which contradicts to that X_0 is a maximal chain. Therefore, $x \in X_0$. As x_0 is the supremum of X_0, $x \leq x_0$. Hence, $x = x_0$, which implies that x_0 is a maximum. □

Another equivalent form of Zorn's lemma can be stated as

Theorem 1.7 *Let X be a set with a partial ordering '$\leq$'. If every chain in X has an upper bound, then X has a maximum.*

We last introduce the Zermelo's choice axiom.

Theorem 1.8 *Let $\mathcal{X}$ be a class of non-empty sets. Then there exists a mapping $f : \mathcal{X} \to \bigcup_{A \in \mathcal{X}} A$ such that for every $A \in \mathcal{X}$, $f(A) \in A$.*

This theorem tells that one is able to form a set by choosing a point from each set in a class of sets, a seemingly trivial but fundamental action in set theory. In history, a nontrivial application of this theorem was in finding a non-Lebesque measurable set (due to Vitali). More involved investigations can show that Theorems 1.5-1.8 are in fact equivalent. These theorems indeed form the axiomatic basis for set theory.

1.5 Field and σ-field(algebra)

Length, area, volume, as well as probability are examples of measure that we will discuss. A measure is a set function, which assigns a number $\mu(A)$ to each set A in s certain class. Some structure must imposed on the class of sets on which the set function μ is define. Probability considerations give a good motivation for the required structure.

In this section, we concentrate on the underlying class of sets and define the essential structure. Finally, we shall give a precise definition for events in a well defined probability space.

Definition: A class of sets $\mathcal{X}$ is closed under an operation $*$ (e.g. union, intersection, etc.) if $*$ performed on any member(s) of $\mathcal{X}$ yields a set which also belongs to the class.

Example 1.7 $\mathcal{X} = \{\phi, A, B, C, A \cup B, A \cup C, B \cup C, A \cup B \cup C\}$ is closed under Union.

Definition: A class $\mathcal{X}$ of sets in Ω is called a *field (Boolean field)* if (i) $\mathcal{X}$ is non-empty; and (ii) $\mathcal{X}$ is closed under finite union and complementation.

This definition means that (i) there exists $A \subset \Omega$ such that $A \in \mathcal{X}$; and (ii) if $A_i \in \mathcal{X}$ for $i = 1 \cdots n$, then $\bigcup_{i=1}^{n} A_i \in \mathcal{X}$, and if $A \in \mathcal{X}$, then $A^c \in \mathcal{X}$.

Some properties of a field are summerized below.

Property 1.4 A field $\mathcal{X}$ is also closed under finite intersection.

Proof: For $A_i \in \mathcal{X}$, $i = 1 \cdots n$, as $A_i^c \in \mathcal{X}$, hence, $\cup A_i^c \in \mathcal{X}$.

Since $\left(\bigcap_{1}^{n} A_i \right) = (\bigcup A_i^c)^c$, therefore, $\bigcap A_i \in \mathcal{X}$. □

Property 1.5 $\phi \in \mathcal{X}$ and $\Omega \in \mathcal{X}$ (homework)

Example 1.8 (1) $\mathcal{X}_1 = \{\phi, \Omega\}$ is a field; (2) $\mathcal{X}_2 =$ all subsets of Ω is a field; (3) Let $\Omega = (-\infty, \infty)$. $\mathcal{X}_3 =$ class of all finite interval (a, b) is not a field (why?)

Definition of σ-field: A class $\mathcal{X}$ of sets is a σ-field if (i) $\mathcal{X}$ is non-empty (ii)$\mathcal{X}$ is closed under complementation and countable unions. A σ-field is often called as a σ-algebra as well.

Example 1.9 $\mathcal{X}_1$ is a trivial σ-field, the 'poorest' σ-field, whereas $\mathcal{X}_2$ is the 'richest' σ-field, containing all subsets of the sample space.

Note that, similar to a field, the empty set, ϕ, and the sample space, Ω, must be contained in a σ-field as well. It is trivial to see that a σ-field is a field, but the converse is not true.

Exercise 1.10 Find a field which is not a σ-field.

Exercise 1.11 Let $\Omega = R, show \mathcal{F} = \{A : A \text{ is countable or } A^c \text{ is countable}\}$ is a σ-field.

Exercise 1.12 Let $\mathcal{X}$ be a nonempty class defined by
$\mathcal{X} = \{A : \ x \in A \Rightarrow x \pm 1, \ x \pm 2 \cdots \text{ are all in } A\}$. Verify that $\mathcal{X}$ is a σ-field.

1.6 σ-field Generated by a Class of Sets

Definition: Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be two σ-fields. We define their intersection as

$$\mathcal{X}_1 \cap \mathcal{X}_2 = \{A : \ A \in \mathcal{X}_1 \text{ and } A \in \mathcal{X}_2\}.$$

We can easily show

Theorem 1.9 *$\mathcal{X}_1 \cap \mathcal{X}_2$ itself is a σ-field.*

Proof: one may verify that conditions (i) and (ii) hold (homework). □

Similarly we define countable intersections and arbitrary intersections and can show that arbitrary intersections of σ-fields are σ-fields. This result is particularly useful as it allows us to construct a 'smallest' σ-field containing a given class of sets.

But it is not the case for unions. For example, given two σ-fields, $\mathcal{X}_1$ and $\mathcal{X}_2$, their union

$$\mathcal{X}_1 \cup \mathcal{X}_2 = \{A : \ A \in \mathcal{X}_1 \text{ or } A \in \mathcal{X}_2\}$$

is not necessarily a σ-field.

Example 1.10 Let $\mathcal{X}_1 = \{\phi, A, A^c, \Omega\}$ and $\mathcal{X}_2 = \{\phi, B, B^c, \Omega\}$. But $\mathcal{X}_1 \cup \mathcal{X}_2 = \{\phi, A, A^c, B, B^c, \Omega\}$ is not a σ-field.

The next theorem establishes the existence of a minimal σ-field containing a given class of sets.

Theorem 1.10 *Given a class of sets, there is a minimum σ-field containing it.*

Proof: Let S be a given class of sets in Ω and $G = \{\mathcal{X} : \ \mathcal{X}$ be a σ-field $\mathcal{X} \supset S\}$.

G is nonempty as the σ-field of all the subsets of Ω contains S. Hence, $\bigcap_{\mathcal{X} \in G} \mathcal{X}$ is the smallest σ-field containing S. □

Definition: We say that a σ-field is *generated by* a class S of sets if it is an intersection of all σ-fields which contain S, denoted by $\sigma(S)$.

We list some properties of $\sigma(S)$.

Property 1.6 $\sigma(S)$ is itself a σ-field.

Property 1.7 $S \subset \sigma(S)$.

Property 1.8 $S_1 \subset S_2$ implies $\sigma(S_1) \subseteq \sigma(S_2)$.

Property 1.9 If S itself is a σ-field, then $\sigma(S) = S$.

Exercise 1.13 Given a series of σ-fields $\mathcal{F}_j$, we know that $\cup_{j=1}\mathcal{F}_j$ may not be a σ-field. We use $\wedge_{j=1}\mathcal{F}_j$ to denote the smallest σ-field containing each $\mathcal{F}_j$. A σ-field is said to be countably generated if and only if it is generated by a countable collection of sets. Prove if each $\mathcal{F}_j$ is countably generated, so is $\wedge_{j=1}\mathcal{F}_j$.

Let $\Omega = R = (-\infty, \infty)$. Consider the following 4 types of finite intervals.

$$
\begin{aligned}
S_1 &= \{[a, b) : \ a < b, \ a, b \in R \ \}, \\
S_2 &= \{(a, b) : \ \}, \\
S_3 &= \{(a, b] : \ \}, \\
S_4 &= \{[a, b] : \ \}.
\end{aligned}
$$

Let $S = \bigcup_{i=1}^{4} S_i$. In other words, S is a class of all finite intervals. But S is neither a field nor a σ-field. An important extension of S to a σ-field is through the following definition.

Definition: The minimal σ-field over S is called the *Borel* σ-field on R, denoted by $\mathfrak{B} = \sigma(S)$. Any set in $\mathfrak{B}$ is called a *Borel* set.

The following shows that any single type of intervals is actually enough to generate the Borel σ-field.

Theorem 1.11 $\mathfrak{B} = \sigma(S_1) = \sigma(S_2) = \sigma(S_3) = \sigma(S_4)$.

Proof: We only prove $\sigma(S_1) = \mathfrak{B}$.

As $S_1 \subset S$, hence $\sigma(S_1) \subset \sigma(S) = B$. On the other hand, if we can prove $S \subset \sigma(S_1)$, then $\sigma(S) \subset \sigma(S_1)$.

In fact, we know $S_1 \subset \sigma(S_1)$. In addition, for any set in S_2, one can write

$$(a, b) = \bigcup_{n=1}^{\infty} \left[a + \frac{1}{n}, b \right).$$

Therefore, $S_2 \subset \sigma(S_1)$. Similarly, we can show $S_3 \subset \sigma(S_1),\ S_4 \subset \sigma(S_4)$. Thus,

$$S = \bigcup_{i=1}^{4} S_i \subset \sigma(S_1).$$

□

Exercise 1.14 Show that $\mathfrak{B}$ is not generated by all the singletons on R.

We return, at the end of this section, to give the formal definition of an event in a probability space.

Definition of Events: Let Ω be a sample space endowed with a σ-field $\mathcal{F}$ of subsets of Ω. An event is defined to be an element of $\mathcal{F}$, i.e. a set in $\mathcal{F}$.

2 Measure Theory

We are now in a position to introduce a measure, *Lebesgue measure,* on the *Borel* sets of R. Based on the essential properties of Lebesgue measure, we will introduce a general measure in any abstract space and, specifically, we will discuss the probability measure in a probability space.

2.1 Lebesgue Measure

A Lebesgue measure on R is a generalization of "length of interval" in a real line to more general sets, e.g. Borel sets. For simplicity, we restrict the sample space to $[0, 1]$.

Definition: Let $\mathcal{I}$ be the class of subintervals $(a, b]$ on $\Omega = [0, 1]$ and define $\lambda(I) = |I| = b - a$, the ordinary length for $I \in \mathcal{I}$. Let $\mathcal{B}_0$ be the field of finite unions of such subintervals. Then for each $A \in \mathcal{B}_0$, there exists disjoint $I_i, i = 1, \ldots, n$ such that $A = \cup I_i$. Define

$$\lambda(A) = \sum \lambda(I_i) = \sum_{i=1}^{n} |I_i|.$$

We call λ *Lebesgue measure.*

Theorem 2.1 *Lebesgue measure λ is countably additive on the field $\mathcal{B}_0$.*

Proof: Suppose that $A = \cup_{k=1}^{\infty} A_k$, where A and A_k are $\mathcal{B}_0$- sets. Then $A = \cup_{i=1}^{n} I_i$ and $A_k = \cup_{j=1}^{m_k} J_{kj}$ are disjoint unions of $\mathcal{I}$-sets. By definition,

$$\begin{aligned} \lambda(A) &= \sum_{i=1}^{n} |I_i| = \sum_{i=1}^{n} \sum_{k=1}^{\infty} \sum_{j=1}^{m_k} |I_i \cap J_{kj}| \\ &= \sum_{k=1}^{\infty} \sum_{j=1}^{m_k} |J_{kj}| = \sum_{k=1}^{\infty} \lambda(A_k). \end{aligned}$$

□

A natural question is how to measure a set which is not in $\mathcal{B}_0$. We thus consider *outer measure*, an extension of λ.

Definition: For each $A \in \Omega$, define its *outer measure* by

$$\lambda^*(A) = \inf \sum_n \lambda(A_n)$$

where the infimum extends over all finite and infinite sequences $A_1, A_2, \ldots,$ of $\mathcal{B}_0$ satisfying $A \subset \cup_n A_n$. Here $\{A_n\}$ need not be disjoint.

There are four properties associated with the set function λ^*

Property 2.1 $\lambda^*(\phi) = 0$

Property 2.2 nonnegativity: for any $A \subset \Omega$ $\lambda^*(A) \geq 0$.

Property 2.3 monotonicity: $A \subset B$ implies $\lambda^*(A) \leq \lambda^*(B)$.

Property 2.4 countably subadditivity: $\lambda^*(\cup_n A_n) \leq \sum_n \lambda^*(A_n)$

Proof: For a given ϵ, choose $\mathcal{B}_0$-sets B_{nk} such that $A_n \subset \cup_k B_{nk}$ and $\sum_k \lambda(B_{nk}) < \lambda^*(A_n) + \epsilon 2^{-n}$ (by the definition of λ^*). Now $\cup_n A_n \subset \cup_{n,k} B_{nk}$ so that $\lambda^*(\cup_n A_n) \leq \sum_{nk} \lambda(B_{nk}) < \sum_n \lambda^*(A_n) + \epsilon$. Hence, the statement follows as ϵ is arbitrary. □

It is also natural in approximating A from the inside by approximating its complement A^c from outside. We define the *inner measure* by

$$\lambda_*(A) = 1 - \lambda^*(A^c).$$

Definition: If $\lambda_*(A) = \lambda^*(A)$, we call A is *(Lebesgue) measurable.*

We have another equivalent definition for a measurable set.

Definition: If for any $E \subset \Omega$,

$$\lambda^*(A \cap E) + \lambda^*(A^c \cap E) = \lambda^*(E) \tag{1}$$

we call A *(Lebesgue) measurable.*

It will not be difficult to show

Theorem 2.2 *Any countable set has Lebesgue measure 0.*

Hence, the Lebesgue measure of the set of all rational numbers is actually zero!

Another natural question might be: does there exist a non-measurable set? Surprisingly, finding a non-measurable set turns out to be a non-trivial task. Only one non-measurable set was ever found (due to Vitali) in history, which was constructed based on the Zermelo's choice axiom.

Finally, we show that

Theorem 2.3 *The class of (Lebesgue) measurable sets in* $\Omega = [0, 1]$ *is a* $\sigma-$*algebra.*

Proof: Let $\mathcal{M}$ be the class of measurable sets in Ω. It is easy to see that $\Omega \in \mathcal{M}$.

If $A \in \mathcal{M}$, as (1) is symmetric (with regard to complementation), $A^c \in \mathcal{M}$.

Our goal now is to show $\mathcal{M}$ is closed under countably infinite intersections. Toward this end, we first show that $\mathcal{M}$ is closed under finite intersections.

Suppose that $A, B \in \mathcal{M}$ and $E \subset \Omega$, Then

$$\begin{aligned}
\lambda^*(E) &= \lambda^*(B \cap E) + \lambda^*(B^c \cap E) \\
&= \lambda^*(A \cap B \cap E) + \lambda^*(A^c \cap B \cap E) \\
&\quad +\lambda^*(A \cap B^c \cap E) + \lambda^*(A^c \cap B^c \cap E) \\
&\geq \lambda^*(A \cap B \cap E) \\
&\quad +\lambda^*((A^c \cap B \cap E) \cup (A \cap B^c \cap E) \cup (A^c \cap B^c \cap E)) \\
&= \lambda^*((A \cap B) \cap E) + \lambda^*((A \cap B)^c \cap E)
\end{aligned}$$

The equality hence follows by subadditivity. hence, $A \cap B \in \mathcal{M}$ and $\mathcal{M}$ is closed under finite intersections (so $\mathcal{M}$ is a field.)

Now suppose $A_1, A_2, \ldots$ are disjoint $\mathcal{M}$ sets with union A. Since $F_n = \cup_{k=1}^n A_k$ lies in $\mathcal{M}$,

$$\lambda^*(E) = \lambda^*(E \cap F_n) + \lambda^*(E \cap F_n^c).$$

Applying to the first term an equality for a finite or infinite sequence of disjoint sets (exercise)

$$\lambda^*(E \cap (\bigcup_k A_k)) = \sum_k \lambda^*(E \cap A_k), \tag{2}$$

and applying the monotonicity to the second term (as $(F_n^c \supset A^c)$, we have

$$\lambda^*(E) \geq \sum_{k=1}^{n} \lambda^*(E \cap A_k) + \lambda^*(E \cap A^c).$$

Let $n \to \infty$ and apply (2) again. Then we end up with

$$\lambda^*(E) \geq \sum_{k=1}^{\infty} \lambda^*(E \cap A_k) + \lambda^*(E \cap A^c) = \lambda^*(E \cap A) + \lambda^*(E \cap A^c).$$

Hence $A = \cup_n A_n$ lies in $\mathcal{M}$. Now we have shown that $\mathcal{M}$ is closed under countable disjoint unions. For any set B_k in $\mathcal{M}$, let $A_1 = B_1$ and $A_k = B_k \cap B_1^c \ldots \cap B_{k-1}^c$, then the A_k are disjoint $\mathcal{M}$ sets and $\cup_k B_k = \cup_k A_k$ hence lies in $\mathcal{M}$, which completes the proof. □

This theorem indicated that applying finite or countably infinite operations (e.g. intersection, union or complementation) on measurable sets would still yield a measurable set.

Exercise 2.1 Prove (2). (Hint: use induction to prove that (2) holds for a finite sequence. For the infinite case, use monotonicity,
$\lambda^*(E \cap (\cup_{k=1}^{\infty} A_k)) \geq \lambda^*(E \cap (\cup_{k=1}^{n} A_k)) = \sum_{k=1}^{n} \lambda^*(E \cap A_k)$. Then let $n \to \infty$.)

In the end, we summarize the development of Lebesgue measure.

- Let $\mathcal{I} = \{[a, b)\}$, where a, b can be $-\infty, \infty$, respectively, and λ is defined on $\mathcal{I}$ by $\lambda((a, b]) = b - a$.
- $\mathcal{B}_0 = \{I : I$ is a finite union of intervals in $\mathcal{I}\}$ is a field. (Why?)
- λ on $\mathcal{B}_0$ is monotone, i.e., $B_1 \subseteq B_2 \Rightarrow \lambda(B_1) \leq \lambda(B_2)$.
- λ is countable additive, hence is a measure in $\mathcal{B}_0$.
- λ is extended to λ^*, the outer measure to measure sets outside of $\mathcal{B}_0$.
- the class of λ^* measurable sets is a σ-field, denoted by $\mathcal{M}$.
- As $\mathcal{M}$ contains $\mathcal{I} = \{(a, b]\}$, $\mathfrak{B}$ (Borel-σ-field) is measurable. (why?)

2.2 Abstract Measure

In the last section, we have talked about Lebesgue measure, which is defined on a real line. We now discuss a measure for an abstract space.

Definition: A measurable space is a space Ω endowed with a σ-field $\mathcal{F}$ of subsets of Ω, often denoted by a pair $(\Omega, \mathcal{F})$.

Definition: A measure μ defined on a measurable space $(\Omega, \mathcal{F})$ is set function:

$$\mu : \mathcal{F} \to R^+,$$

which is non-negative and countably additive. That is, (1) $\mu(A) \geq 0$ for any $A \in \mathcal{F}$, and (2) if A_n is a sequence of disjoint sets in $\mathcal{F}$,

$$\mu(\sum_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n).$$

Similarly, we can define μ for any class of σ-field, say S, but $\sum_{n=1}^{\infty} A_n$ must be in S.

Exercise 2.2 For any countably infinite set Ω, the collection of its finite subsets and their complements forms a field $\mathcal{F}$. If we define a set function $\mu(E)$ on $\mathcal{F}$ to be 0 or 1 according as E is finite or not, then show μ is finitely additive but not countably so.

Example 2.1 Lebesgue measure is a measure by the definition in this section.

Example 2.2 Let $\Omega = \{1, 2, \cdots, n, \cdots\}$, a set of all positive integers, $\mathcal{F}$ = all subsets of Ω, let $\mu(A) = \#$ of integers in A. μ is a measure on $(\Omega, \mathcal{F})$, we call it "counting measure".

Example 2.3 Let $\Omega = R, \mathcal{F} = \{A : A \text{ is countable or } A^c \text{ is countable}\}$

$$\mu(A) = \begin{cases} 1, & \text{if } A \text{ is countable;} \\ 0, & \text{if } A^c \text{ is countable;} \end{cases}$$

Let A_1 and A_2 be two disjoint countable sets. Then $A_1 + A_2$ is countable too¿ Hence $\mu(A_1 + A_2) = 1 \neq \mu(A_1) + \mu(A_2) = 2$. So μ is **NOT** a measure!

Exercise 2.3 We redefine

$$\mu(A) = \begin{cases} 0, & \text{if } A \text{ is countable;} \\ 1, & \text{if } A^c \text{ is countable;} \end{cases}$$

show that μ is indeed a measure.

Definition: A measure μ is finite, if $\mu(\Omega) < \infty$, otherwise μ is a infinite measure. A measure μ is called $\sigma - finite$, if there exists a countable partition of Ω,

$$\Omega = \sum_{n=1}^{\infty} A_n, A_n \in \mathcal{F}$$

such that

$$\mu(A_n) < \infty$$

for any $n = 1, 2, \cdots$.

By definition, a finite measure is a $\sigma-$finite measure, but not vice versa, as a $\sigma-$finite measure can be infinite.

Remark 2.1 Example 2.2 is a σ-finite measure, but not a finite measure.

Theorem 2.4 *A measure μ defined on a field F can be extended to a measure on $\sigma(F)$. If μ is σ-finite, the extension is unique.*

A detailed proof can be found in Billingsley (1995, p.37-43), wherein the construction of the extension is similar to the development of the outer measure in the context of Lebesgue measure. Similarly, this theorem holds for Lebesgue measure in higher Euclidean spaces, i.e. $R^2, R^3, \cdots$.

2.3 Probability Measure

Given that an "event" in a probability space $(\Omega, \mathcal{F})$ is defined as a member of the σ-field $\mathcal{F}$, our interest here is to measure the possibility that an event occurs. For this purpose, we introduce the concept of a *probability measure*.

Definition: A set function $P : \mathcal{F} \to [0, 1]$ is a *probability measure* on $(\Omega, \mathcal{F})$ if it satisfies

(i) (nonnegativity)for any $A \in \mathcal{F}$ $P(A) \geq 0$.

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

(iii) (countable additivity) If $\{A_n\}$ is a countable collection of (pairwise) disjoint sets in $\mathcal{F}$, then $P(\bigcup_n A_n) = \sum_n P(A_n)$.

In some literature, the conditions above are termed Kolmogorov's Axioms for probability.

Exercise 2.4 Let Ω be a sample space for a given experiment, say, E. Suppose that E is repeated n times. Consider the following set function: for any $A \subset \Omega$, define $P_n(A) = \frac{1}{n}\sum_{i=1}^{n} I\{\text{the outcome of experiment i is in A}\}$, where $I\{\cdot\}$ is the indicator function (i.e. equal to one if the argument is true, and zero otherwise). Note that $P_n(A)$ is just the proportion of times the event A occurs in n replications of the experiment. Prove that $Pn(\cdot)$ is a probability measure.

It is easy to prove that the probability measure $P(\cdot)$ has the following properties:

Property 2.5 $P(\phi) = 0$

Proof: As $\Omega = \Omega + \phi + \phi + \cdots$, hence $P(\Omega) + P(\phi) + P(\phi) + \cdots = 1$. Therefore, $P(\phi) = 0$. □

Property 2.6 $P(A^c) = 1 - P(A)$

Property 2.7 $A \subseteq B \Rightarrow P(A) \leq P(B)$

Proof: consider $B = A + A^c \cap B$. □

Property 2.8 $P(\bigcup_{n=1}^{k} A_n) \leq \sum P(A_n)$.

Proof: consider the decomposition

$$\bigcup A_n = A_1 + A_1^c A_2 + A_1^c A_2^c A_3 + \cdots + A_1^c A_2^c \cdots A_k,$$

together with $P(A_1^c A_2) \leq P(A_2)$. □

Property 2.9 (continuity) (i) If $A_n \uparrow$ and $A_n \in \mathcal{F}, then P(\lim_n A_n) = \lim_n P(A_n)$. (ii) Similarly, if $\{A_n\} \downarrow$, then $P(\lim A_n) = \lim P(A_n)$.

Proof: (i) As $A_n \uparrow$, $\lim_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} A_n$. We then write $\lim_{n\to\infty}$ in a disjoint union, i.e.

$$\lim_{n\to\infty} A_n = A_1 + A_1^c A_2 + A_1^c A_2^c A_3 + \cdots .$$

So,

$$\begin{aligned} P(\lim_{n\to\infty} A_n) &= P(A_1) + P(A_1^c A_2) + \cdots \\ &= \lim_{n\to\infty} P(A_1) + \cdots + P(A_1^c A_2^c \cdots A_n) \\ &= \lim_{n\to\infty} [P(A_1 + \cdots + A_1^c A_2^c \cdots A_n)] \\ &= \lim_{n\to\infty} P(A_n). \end{aligned}$$

(ii) homework. □

The following indicates that the definition of probability has another equivalent form.

Exercise 2.5 Show conditions (ii) and (iii) in the definition of probability measure can be replaced by "If $\Omega = \cup_{i=1}^{\infty} A_i$, where A_i are (pairwise disjoint), then $\sum_{i=1}^{\infty} P(A_i) = 1$".

Proof: homework (Hint: write
$\Omega = (\sum A_n)^c + (\sum A_n) = (\sum A_n)^c + A_0 + A_1 + A_2 + \cdots .$)

Using the continuity properties we are able to prove a theorem concerning the probabilities of the lower and upper limits of a sequence of sets.

We recall that, for a series of sets $\{A_n\}$, we defined $\liminf_n A_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k$ and $\limsup_n A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$.

Theorem 2.5 *Suppose $A_n \in \mathcal{F}$ for all n. Then*

$$P(\limsup_n A_n) = \lim_{m\to\infty} P\left(\bigcup_{n=m}^{\infty} A_n\right), \tag{3}$$

$$P(\liminf_n A_n) = \lim_{m\to\infty} P\left(\bigcap_{n=m}^{\infty} A_n\right). \tag{4}$$

Proof: Letting $B_m = \bigcup_{n=m}^{\infty} A_n$ and $C_m = \bigcap_{n=m}^{\infty} A_n$, then B_m is a decreasing sequence of sets and C_m is an increasing sequence, with $B_m \downarrow \limsup_n A_n$ and $C_m \uparrow \liminf_n A_n$. (3) and (4) then follow from the continuity properties of probability measures. □

We are then able to prove an important theorem in the probability literature. This theorem indicates what can be expected by interchanging the order of limit and probability measure.

Theorem 2.6 *(***Fatou-Lebesgue Theorem***) For any sequence* $\{A_n\} \in \mathcal{F}$

$$P(\liminf_n A_n) \le \liminf_n P(A_n) \le \limsup_n P(A_n) \le P(\limsup_n A_n).$$

In addition, if $\lim A_n$ *exists,* $P(\lim A_n) = \lim P(A_n)$.

Proof: Let $B_n = \bigcap_{k=n}^{\infty} A_k$. Then

$$\liminf_n A_n = \bigcup_{n=1}^{\infty} B_n = \lim_{n\to\infty} B_n.$$

As $B_n \subseteq A_n$, so $P(B_n) \le P(A_n)$. Therefore,

$$\lim P(B_n) = \liminf_n P(B_n) \le \liminf_n P(A_n).$$

Also notice that

$$P(\liminf_n A_n) = P(\lim B_n).$$

Hence, if $\lim A_n$ exists, $P(\lim A_n) = \lim P(A_n)$. □

Another useful theorem we shall prove is *the first Borel-Cantelli lemma.*

Theorem 2.7 *If* $\sum_n P(A_n)$ *converges, then* $P(\limsup_n A_n) = 0$.

Proof: From $\limsup_n A_n \subset \cup_{k=m}^{\infty} A_k$ follows

$$P(\limsup_n A_n) \leq P(\cup_{k=m}^{\infty} A_k) \leq \sum_{k=m}^{\infty} P(A_k),$$

and this sum tends to 0 as $\sum_n P(A_n)$ converges. □

The first Borel-Cantelli lemma is useful in establishing the strong law of large numbers for a sequence of i.i.d random variables with a finite mean; see Bilingsley (1995, p.85).

2.4 Independence and Conditional Probability

Intuitively, two events A and B are independent if a statement concerning the occurrence or nonoccurrence of one of the events does not change the odds about the other event. This leads us to introduce fundamental new concept peculiar to the theory of probability, that of 'independence'.

Definition: Two events A and B are *independent*, if and only if

$$P(AB) = P(A)P(B)$$

or

$$P(A|B) = P(A), P(A|B^c) = P(A).$$

Naturally we may extend the definition for a (finite) series of events.

Definition: Events $A_1, \cdots, A_n$ are *completely independent* if and only if $P(A_{k_1} \cdots A_{k_s}) = P(A_{k_1}) \cdots P(A_{k_s})$ for every $1 \leq k_1 < \cdots < k_s \leq n$.

Exercise 2.6 For events $A_1, \ldots, A_n$, consider the 2^n equations

$$P(B_1 \cap \ldots \cap B_n) = P(B_1) \ldots P(B_n)$$

with $B_i = A_i$ or A_i^c for each i. Show that $A_1, \ldots, A_n$ are independent if all these equations hold.

But pairwise independence is not equivalent to completely independence as indicated in the following exercise.

Exercise 2.7 Toss two fair coins. Denote by A_1=H in the first toss, A_2=H in the second toss and A_3=HH or TT. Then show $P(A_iA_j) = P(A_i)P(A_j)$ but $P(A_1A_2A_3) \neq P(A_1)P(A_2)P(A_3)$.

Given a series of independent events, we can prove the following *second Borel-Cantelli lemma.*

Theorem 2.8 *If A_n is an independent sequence of events and $\sum_n P(A_n)$ diverges, then $P(\limsup_n A_n) = 1$.*

Proof: It suffices to show that $P(\cup_{n=1}^{\infty} \cap_{n=1}^{\infty} A_n^c) = 0$ can hence is enough to show $P(\cap_{n=1}^{\infty} A_n^c) = 0$ for every n. Since $1 - x \leq e^{-x}$,

$$P\left(\bigcap_{k=n}^{n+j} A_k^c\right) = \prod_{k=n}^{n+j}(1 - P(A_k)) \leq \exp\left[-\sum_{k=n}^{n+j} P(A_k)\right].$$

Since $\sum_k P(A_k)$ diverges, the last expression tends to 0 as $j \to \infty$, and therefore

$$P(\left(\bigcap_{k=n}^{\infty} A_k^c\right) = \lim_j P(\left(\bigcap_{k=n}^{n+j} A_k^c\right) = 0.$$

□

By Theorem (2.6) (the Fatou-Lebesgue Theorem), $\limsup_n P(A_n) > 0$ implies $P(\limsup_n A_n) > 0$, whereas in the theorem above the hypothesis $\sum_n P(A_n) = \infty$ is weaker but the conclusion is stronger because of the additional assumption of independence.

For arbitrary number (e.g. infinite or uncountable) of events, we define their independence as follows.

Definition: Events in a class C are *independent*, if and only if events in all finite subclass are independent.

Definition: Let C_1 and C_2 be two class of events. C_1 and C_2 are independent if and only if for any $A_1 \in C_1$ and any $A_2 \in C_2$, A_1 and A_2 are independent.

Noting that a $\sigma-$field itself is a class of events, we have the following theorem.

Theorem 2.9 *Let C_1 and C_2 be two class of events such that $C_i, i = 1, 2$ are closed under finite intersection and C_1 and C_2 are independent. Then $\sigma(C_1)$ and $\sigma(C_2)$ are independent.*

Proof: see Billingsley (1995, p.55). □

We have considered the situation of independence for two events, where the occurrence or nonoccurrence of one event does not alter the odds about the other. In the absence of independence, the odds are altered and the concept of conditional probability measures quantitatively the change.

For example, in a series of independent trials where two events are observed, suppose that $P(A) = 0.4$ and $P(A \cap B) = 0.1$. Given a large number of trials, if we restrict to the trials on which A has occurred, B will occur roughly 25% of the time. In general, the ration $P(A \cap B)/P(A)$ is a measure of the probability of B under the condition that A is known to have occurred. In other words, the concept of 'conditional probability' arises when we restrict to only parts of the sample space.

Definition of Conditional Probability: Given a probability space $(\Omega, \mathcal{F}, P)$, for a $B \in \mathcal{F}$ such that $P(B) > 0$, and any $A \in \mathcal{F}$, the conditional probability of A given B is defined to be

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

In the next section, we will also consider the definition of conditional probability $P(A|B)$ when the event B has probability 0. Of course, the definition of (5) is no long valid. An indirect approach will be used.

Exercise 2.8 $P(\cdot|B)$ is a probability measure on $(\Omega, \mathcal{F})$.

Theorem 2.10 (i) If $P(B) > 0$, A and B are independent if and only if $P(A|B) = P(A)$. (ii) (law of multiplication law) If $P(A_1 \cdots A_{n-1}) > 0$, then

$$P(A_1 \cdots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2) \cdots P(A_n|A_1 \cdots A_{n-1}).$$

Proof: (i) follows by definition. To prove (ii), observe that $P(A_1 \cdots A_{n-1}) > 0$ implies that $P(A_1) > 0, P(A_1 \cap A_2) > 0, \cdot, P(A_1 \ldots \cap A_{n-1}) > 0$. Hence, all the conditional probabilities involved are well defined. Using

$$P(A_1 \cdots A_n) = P(A_n|A_1 \ldots \cap A_{n-1})P(A_1 \ldots \cap A_{n-1}),$$

we complete the proof by induction. □

3 Random Variables

We begin with some basic concepts involving mappings.

Definition: Let X be a mapping from $\Omega \to \Omega'$. For any $\omega \in \Omega$, there is a unique $X(\omega) \in \Omega'$. Ω is called the *domain* of X and Ω' the *range* of X. If $\omega \neq \omega'$ implies $X(\omega) \neq X(\omega')$, X is called *injective*. For any
$A \subset \Omega, X(A) = \{y : y \in \Omega', y = X(\omega) \text{ for some } w \in A\}$ is called the *image* of A under X. If $X(\Omega) = \Omega'$, X is called *surjective* or *onto* and Ω' is called *range* or *strict range* of X. For $B \subseteq \Omega'$,

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$$

Remark 3.1 $X(A) = B$ doesn't necessarily imply $X^{-1}(B) = A$.

The next proposition, a standard exercise on inverse mapping, is essential.

Theorem 3.1 *X^{-1} commutable with set operations*

- $X^{-1}(\bigcup_1^\infty B_n) = \bigcup_n X^{-1}(B_n)$.
- $X^{-1}(\bigcap_1^\infty B_n) = \bigcap_n X^{-1}(B_n)$.
- $X^{-1}(B^c) = (X^{-1}(B))^c$.
- $X^{-1}(B_1 - B_2) = X^{-1}(B_1) - X^{-1}(B_2)$.

Proof: Homework. □

Now, let C be a class of events in Ω' and define
$X^{-1}(C) = \{A : A \in \Omega, A = X^{-1}(B), \text{for some } B \in C\}$. Then we have two relevant theorems.

Theorem 3.2 *Let C be a σ-field in Ω', then $X^{-1}(C)$ is also a σ-field.*

Proof: homework. □

Theorem 3.3 *Let C be any class of sets in Ω' and $\sigma(C)$ is the minimal σ-field, then*

$$X^{-1}(\sigma(C)) = \sigma(X^{-1}(C))$$

Proof: homework. □

3.1 Random Variables

A random variable is a quantity associated with a random experiment. If an experiment is carried out in a probability space $(\Omega, \mathcal{F}, P)$ and the outcome corresponds to a sample point $\omega \in \Omega$, a measuring process is performed to obtain a number $X(\omega)$. Thus X is a function mapping Ω to the real space.

We are often interested in measuring the probability of events involving X. For example, we may want to know the probability of X belongs to B, a Borel set. Thus we want to compute $P(\omega : X(\omega) \in B)$. To let this make sense, we need to require $\{\omega : X(\omega) \in B\}$ is an event. In other words, we need to have $\{\omega : X(\omega) \in B\} \in \mathcal{F}$. This leads to the following definition of random variable.

Definition 1: Let $(\Omega, \mathcal{F}, P)$ be a probability space. A random variable X is a function $\Omega \to R$ such that the inverse image of all Borel sets are in $\mathcal{F}$, i.e., $X^{-1}(B) \in \mathcal{F}$ for any $B \in \mathfrak{B}$.

Definition 2: A random variable X is a function: $\Omega \to R$ such that $X^{-1}(-\infty, b) \in \mathcal{F}$ for any real number b.

Theorem 3.4 *Definitions 1 and 2 are equivalent.*

Proof: homework. (Hint: Consider the collection $\mathcal{A}$ of all subsets S of R such that $X^{-1}(S) \in \mathcal{F}$ and prove $\mathcal{A}$ is a σ-field.) □

It is customary to omit the argument ω in $X(\omega)$ in probability theory. Thus, with no confusion, X stands for a general value of $X(\omega)$ of the function as well as the function itself and $[X < b]$ is short for $\{\omega : X(\omega) < b.\}$ Hence, we write

$$X^{-1}(-\infty, b) = [X < b] = \{\omega : X(\omega) < b\}.$$

Finally, for two random variables, X, Y, we define their sum $Z = X + Y$ if $Z(\omega) = X(\omega) + Y(\omega)$ for any $\omega \in \Omega$.

Definition of Pointwise Convergence of R.V.: Let$\{X_n\}$ be a sequence of random variables. $Y = \lim X_n$is defined as any $\omega, Y(\omega) = \lim X_n(\omega)$. In a similar way, we define $Y = \limsup_n X_n$ and $Y = \liminf_n X_n$.

One of the basic reasons why measurable functions are useful is that a pointwise limit of measurable functions is still measurable.

Theorem 3.5 *Let $\{X_n\}$ be a sequence of random variables defined on $(\Omega, \mathcal{F}, P)$, $\liminf_n X_n, \limsup_n X_n, \lim X_n$, are random variables provided they are finite function and are defined pointwise.*

Proof: To prove that $Y = \liminf_n X_n$ is a random variable, we need to show

$$Y^{-1}(-\infty, b) \in \mathcal{F}.$$

In fact,

$$[Y < b] = [\liminf_n X_n < b] = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} [X_k < b].$$

Noting that $\limsup_n X_n = -(\liminf_n(-X_n))$, one may also show $Y = \limsup_n X_n$ is a random variable. □

We have previously discussed independence of events; we now consider independent for random variables. Intuitively, independence of $\{X_i, i = 1, \ldots, n\}$ means a statement about one or more X_i does not change the odds concerning the remaining X_i. A formal definition is as follows.

Definition: The random variables $\{X_i, i = 1, \ldots, n\}$ are said to be *independent* if and only if for any *Borel* sets $\{B_i, i = 1, \ldots, n\}$, we have

$$P\left\{\bigcap_{i=1}^{n}(X_i \in B_i)\right\} = \prod_{i=1}^{n} P(X_i \in B_i).$$

Let I_A be the indicator function of the set $A \subseteq \mathcal{F}$, i.e.

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

Then it is easy to note that, for a series of events, say, $\{A_i, i = 1, \ldots, n\}$, a definition of independence equivalent to that in Section 2.4 is that their indicators are independent.

3.2 Simple Random Variable

The previous section dealt with general random variables, i.e. measurable functions on Ω with arbitrary range. We here introduce the simplest but extremely important random variables with only finite range.

Definition: X is a simple random variable if and only if there exists a finite measurable partition of Ω, i.e. $\Omega = \sum_{i=1}^{n} A_i, A_i \in \mathcal{F}$. and $X(\omega) = x_i$ for any $\omega \in A_i, i = 1, \cdots, n$. Here, x_i are real numbers (not necessarily distinct).

Example 3.1 Let I_A be the indicator function of the set A. Then I_A is a simple random variable.

Exercise 3.1 Check I_A is a random variable.

Example 3.2 $X = \sum_{i=1}^{n} x_i I_{A_i}$ is a simple random variable.

Remark 3.2 A set A is called P-null if $P(A) = 0$. Empty set is a null set, but the converse is not true.

3.3 Approximation Theorem

Simple random variables are easy to handle. Further, as shown in the next theorem, each random variable can actually be approximated by a series of simple random variables. This result is very useful for the development of *expectation* in the later sections.

Theorem 3.6 *(Approximation Theorem)*

(a) Every random variable is the limit of a sequence of simple random variables.

(b) Every non-negative random variable X can be approximated by a sequence of non-negative and monotone increasing sequence of simple random variable, i.e,

if $X \geq 0$, we can construct a sequence of $\{X_n\}$ such that for each n $\{X_n\}$ is a simple random variable, $X_n \geq 0$ and $X_n \uparrow X$ as $n \to \infty$.

Proof: (a) Let X be a random variable. We construct a sequence of simple random variable.

For a fixed $\omega \in \Omega$, define

$$X_n(\omega) = -nI_{[X<-n]}(\omega) + \sum_{k=-n2^n+1}^{n2^n} \frac{k-1}{2^n} I(\omega)_{[\frac{k-1}{2^n} \leq X < \frac{k}{2^n}]} + nI_{[X \geq n]}(\omega).$$

This is equal to

$$X_n(\omega) = \begin{cases} -n & \text{if} X(\omega) < -n \\ \frac{k-1}{2^n} & \text{if} \frac{k-1}{2^n} \leq X(\omega) < \frac{k}{2^n} \\ n & \text{if} X(\omega) \geq n. \end{cases}$$

Obviously, for $n = 1, 2, \ldots$, X_n are all simple random variables.

To show $\lim X_n = X$ i.e. for any $\omega, \lim X_n(\omega) = X(\omega)$. Take an arbitrary ω and let $b_\omega = X(\omega)$. Then for any $n > |b_\omega|, |X_n(\omega) - X(\omega)| < \frac{1}{2^n}$. Hence, as $n \to \infty, |X_n(\omega) - X(\omega)| \to 0$.

(b) Consider

$$X_n(\omega) = \sum_{k=1}^{n2^n} \frac{k-1}{2^n} I(\omega)_{[\frac{k-1}{2^n} \leq X < \frac{k}{2^n}]} + nI_{[X>n]}(\omega).$$

□

3.4 Distribution Function

Given a probability space $(\Omega, \mathcal{F}, P)$, let X be a random variable. For any $B \in \mathfrak{B}$, where $\mathfrak{B}$ is the Borel σ-field in R, the set function P_X on $(R, \mathfrak{B})$ defined by

$$P_X(B) = P(X^{-1}(B)) = P([X \in B])$$

is called the *distribution function* of X.

One may verify that P_X is a probability measure on the space $(R, \mathfrak{B})$.

Definition: P_X is called the probability measure *induced* from $(\Omega, \mathcal{F}, P)$ by the random variable X, also the probability distribution of X.

The numbers $P_X(B)$, $B \in \mathcal{F}$, completely characterize the random variable X in that they provide the probabilities of all events involving X. It is useful to know such information can actually be captured by a single real function, namely, *cumulative distribution function*. The equivalence of the cumulative distribution function to P_X will be proved in Theorem (3.11).

Definition: Define a point function $F_X(x) = P_X(-\infty, x) = P([X < x])$ for any real number x. we will write $F(x)$ for $F_X(x)$, which is called the cumulative distribution function of X.

A cumulative distribution function has two fundamental properties.

Property 3.1 $F(x)$ is non-decreasing and left continuous.

Proof: For $x < x'$, we want to show $F(x) \leq F(x')$. As $(-\infty, x) \subseteq (-\infty, x')$ and P_X is a probability measure, hence $P_X(-\infty, x) \leq P_X(-\infty, x') \Rightarrow F(x) \leq F(x')$.

As $(-\infty, x - \frac{1}{n}) \uparrow (-\infty, x)$ when $n \to \infty$, $F(x - \frac{1}{n}) \to F(x), n \to \infty$, i.e.,
$F(x - 0) = F(x)$. □

Property 3.2 $F(-\infty) = 0$ and $F(\infty) = 1$

Proof: Consider $(-\infty, x) \downarrow \phi, x \to -\infty$, and $(-\infty, x) \uparrow R, x \to \infty$. □

Remark 3.3 As $(-\infty, x + \frac{1}{n}) \downarrow (-\infty, x], n \to \infty$, then
$F(x + 0) = P_X(-\infty, x] = P_X(-\infty, x) + P_X(\{x\})$, so $P([X = x]) = F(x + 0) - F(x)$.
If we define $F(x) = P(X \leq x)$ then F is right continuous.

Independence of random variables can be characterized in terms of cumulative distribution function as follows.

Theorem 3.7 *Let $X_1, \ldots, X_n$ be random variables on $(\Omega, \mathcal{F}, P)$. Let F_i be the cumulative distribution function of $X_i, i = 1, \ldots, n$ and F the (joint) cumulative distribution function of $X = (X_1, \ldots, X_n)$. Then $X_1, \ldots, X_n$ are independent if and only*

$$F(x_1, \ldots, x_n) = F_1(x_1) \ldots F_n(x_n)$$

for all real $x_1, \ldots, x_n$.

Proof: see Ash (1972, p.214).

The following theorem states that a cumulative distribution function does not have 'too many' discontinuities. In fact, the set of these discontinuities is a null set, i.e. its Lebesgue measure is 0.

Theorem 3.8 *Every cumulative distribution function* $F(x)$ *has at most a countable numbers of discontinuous points.*

Proof: Consider and intervals $[-k, k]$, let $-k \leq x_1 < x_2 < \cdots < x_n \leq k$ be any n discontinuous points of $F(x)$ in this interval. $F(x_i) < F(x_i + 0)$ Let $P(x_i) = F(x_i + 0) - F(x_i)$ then $\sum_{i=1}^{n} P(x_i) \leq F(k) - F(-k) \leq 1$. So the number of jumps by more than $1/n$ can be at most n, Let $S_{n,k}$ be set of discontinuous points in $[-k, k]$ with jump size greater than $\frac{1}{n}$. It is a finite set. $D =$ set of all discontinuous points$=\bigcup_{k=1}^{\infty} \bigcup_{n=1}^{\infty} S_{n,k}$=countable union of finite sets=countable. Therefore D can always be written as $D = \{x_n\}$. □

We give a simpler proof as well. For each discontinuous point x, consider the open interval $I_x = (F(x-), F(x+))$. If x' is another point of jump and $x < x'$, say, then there is a point $\tilde{x}$ such that $x < \tilde{x} < x'$. Hence by monotonicity, we have

$$F(x+) \leq F(\tilde{x}) \leq F(x'-).$$

It follows that I_x and $I_{x'}$ are disjoint, though they may abut on each other if $F(x+) = F(x'-)$. Now we can associate with the set of discontinuities a set of pairwise disjoint open sets. Such a collection of intervals is countable as each interval can be indexed by a rational number it contains. □

As the simpler proof only uses the monotonicity of $F(x)$, we in fact have proved a more general theorem.

Theorem 3.9 *Any monotone function has at most countable discontinuities.*

We may even further decompose a cumulative distribution function into the sum of a continuous cumulative distribution function and a piece-wise constant cumulative distribution function.

Theorem 3.10 *(Decomposition Theorem) Every cumulative distribution function $F(x)$ can be decomposed (uniquely) as $F(x) = pF_c(x) + (1-p)F_d(x)$, where $0 \leq p \leq 1$ and F_c and F_d are both proper cumulative distribution functions. F_c is continuous and F_d is a pure step function.*

Proof: Let $\{x_n\}_{n=1}^{\infty}$ be the set of discontinuous points of $F(x)$. Write $p(x_k) = F(x_k + 0) - F(x_k)$ and denote by $(1-p) = \sum_{i=1}^{\infty} p(x_i)$, $0 < p < 1$.

Take

$$F_d(x) = \frac{1}{1-p} \sum_{x_n < x} p(x_n)$$

and

$$F_c(x) = \frac{1}{p}[F(x) - (1-p)F_d(x)].$$

We first consider F_d. Obviously, it is increasing and left continuous i.e. $F_d(x-0) = F_d(x)$. Also, $F_d(-\infty) = 0$, since $(-\infty, x) \downarrow \phi$ as $x \to -\infty$. In addition, $F_d(\infty) = 1$. So $F_d(x)$ is a proper cumulative distribution function .

We observe that $F_c(x)$ is left continuous, $F_c(-\infty) = 0$ and $F_c(\infty) = 1$. We next prove F_c is right continuous and increasing. Let $x' > x$, since

$$p(F_c(x') - F_c(x)) = F(x') - F(x) - \sum_{x \leq x_n < x'} p(x_n)$$

$$= (F(x') - F(x+0)) - \sum_{x < x_n < x'} p(x_n) \geq 0$$

and let $x' \downarrow x$, we then have $F_c(x+0) = F_c(x)$.

Now assume there are two decompositions:

$$F(x) = pF_c(x) + (1-p)F_d(x) = p'F_c'(x) + (1-p')F_d'(x).$$

Hence,

$$pF_c(x) - p'F_c'(x) = (1-p')F_d'(x) - (1-p)F_d(x).$$

The contradiction occurs as the left side is continuous function and right side is a step function. □

By definition, the probability measure P_X of a random variable X uniquely determines its cumulative distribution function $F(x)$. The following theorem shows that the other way around is also true. Hence, P_X and $F(x)$ are two equivalent definitions.

Theorem 3.11 *(Correspondence Theorem) The cumulative distribution function $F(x)$ uniquely determines the probability measure P_X on $(R, \mathfrak{B})$.*

Proof: We shall prove this theorem based on the definition of $F(x) = P_X(-\infty, x)$ and the fact that P_X is a probability measure.

Let $S = \{(-\infty, a), [a, b), [b, \infty)\}$. Given only a cumulative distribution function $F(x)$, define

$$P_X(-\infty, a) = F(a), P_X([a, b)) = F(b) - F(a),$$

and

$$P_X[b, \infty) = 1 - F(b), P_X(a, b) = F(b) - F(a + 0).$$

Thus $F(x)$ determines P_X for all intervals and in particular for those in S. Let $F =$ field of finite unions of intervals in S. Let $B \in F$. Thus B has the representation $B = \sum_{j=1}^{n} B_j$. Hence,

$$P_X(B) = \sum_{j=1}^{n} P_X(B_j).$$

For any other representation, we get the same $P_X(B)$. So P_X is uniquely defined on F. That F is a field and P_X is a measure imply P_X is uniquely defined on $(R, \mathfrak{B})$ by the Extension Theorem. □

Theorem 3.12 *If F is a non-decreasing and left continuous function with $F(-\infty) = 0, F(\infty) = 1$, then there exists on some probability space a random variable X for which $P(X < x) = F(x)$.*

Proof: By the Correspondence Theorem, $F(x)$ uniquely determines a probability measure P_X on $(R, \mathfrak{B})$. For the probability space, take $(\Omega, \mathcal{F}, P) = (R, \mathfrak{B}, P_X)$ and for the random variable, take the identity function, i.e. $X(\omega) = \omega$ for any $\omega \in \Omega = R$. Then $P(X < x) = P_X(x) = F(x)$. □

Definition: If $g : R \to R$ such that $g^{-1}(B) \in \mathfrak{B}$ for any $B \in \mathfrak{B}$, we term g a *Borel function* on $R \to R$.

It can be shown that the Borel function of a random variable is also a random variable, i.e. $Y = g(X)(\omega) = g(X(\omega))$, where $g(X)$ is a function $\Omega \to R$, is a random variable. Note that the distribution function of $Y = g(X)$ can be calculated by $P_Y(B) = P_X(g^{-1}(B)) = P(X^{-1}g^{-1}(B))$.

Example 3.3 Let X be a random variable with a cumulative distribution function $F(x)$. Suppose $Y = -X$, then the cumulative distribution function of Y is

$$P([Y < y]) = P([X > -y]) = P_X(-y, \infty) = 1 - F(-y + 0).$$

Finally we consider a multi-dimensional random variable. A $k-$dimension random vector $\mathbf{X}$ is a map $\Omega \to R^k$ such that $\mathbf{X^{-1}}(B^k) \in \mathcal{F}$ for any B^k in the $k-$dimensional Borel $\sigma-$field on R^k, defined as the $\sigma-$field generated by all $k-$dim rectangles. This topic will be discussed in detail in Section 9.

4 Expectation of Random Variables

If X is a random variable on $(\Omega, \mathcal{F}, P)$, the *expectation* of X is defined by

$$E(X) = \int_\Omega X dP, \tag{6}$$

provided such an integral exists. Thus $E(X)$ is the integral of the Borel measurable function with respect to the probability measure.

We shall discuss in the next section how integral (6) is defined and calculated for different types of random variables. We begin with simple random variables and extend to nonnegative random variables. Finally, we define (6) for a general random variable.

4.1 Abstract Lebesgue Integration

Definition for Simple R.V.: Let X be a simple random variable on $(\Omega, \mathcal{F}, P)$, i.e. $X = \sum_{i=1}^n x_i I_{A_i}$, where $A_i \in \mathcal{F}, \sum_i A_i = \Omega$. Then we define

$$E(X) = \int_\Omega X dP = \int_\Omega X(\omega) dP(\omega) \stackrel{\text{def}}{=} \sum_{i=1}^n x_i P(A_i),$$

and for any $C \in \mathcal{F}$

$$\int_C X dP = \int_\Omega X I_C dP = \sum_{i=1}^{n} x_i P(CA_i).$$

Often we write $\int X dP$ for $\int_\Omega X dP$.

We summarize below some properties associated with the expectation defined above.

Property 4.1 $\int X dP$ is uniquely defined. That is, if $X = \sum_{j=1}^{n} x_j I_{A_j}$ and $X = \sum_{i=1}^{m} x'_i I_{C_i}$ are two representations for the simple random variable X, then $\int X dP$ has the same value.

Proof: Let

$$D_{ij} = A_j C_i, \sum_i \sum_j A_j C_i = \Omega, X = \sum_i \sum_j x_{ij} I_{D_{ij}},$$

where $x_{ij} = x_j = x'_i$. Now consider

$$\sum_i \sum_j x_{ij} P(D_{ij}) = \sum \sum x_{ij} P(A_j C_i) = \sum_j x_j P(A_j) = \sum_i x'_i P(C_i).$$

□

Property 4.2 $X \geq 0$ implies $\int X dP \geq 0$.

Property 4.3 If $X \geq 0, \int X dP = 0$ then $P(X = 0) = 1$.

Property 4.4 Linearity: $\int (aX + bY) dP = a \int X dP + b \int Y dP$ or $E(aX + bY) = aE(x) + bE(y)$.

Property 4.5 Monotonicity: If $X \geq Y$, then $\int X dP \geq \int Y dP$ or $E(X) \geq E(Y)$.

Definition for Non-negative R.V.: Let X be a non-negative random variable, i.e. $X(\omega) \geq 0$ for all $w \in \Omega$. By approximation theorem, there exists a sequence $\{X_n\}$ of non-negative simple random variable such that $X_n \uparrow X$. So we define

$$\int X dP = \lim_{n \to \infty} \int X_n dP.$$

Since $X_1 < X_2 \cdots$, $\int X_1 dP \leq \int X_2 dP \cdots$. Hence, $\{\int X_n dP\}$ is a increasing sequence of real number and, therefore, its limit may be infinite or finite. If the limit is finite, we say X is integrable, whereas if the limit is ∞, we say the integral exists but X is not integrable.

The expectation obtained by using the simple random variable approximation is indeed unique as shown in the following theorem.

Theorem 4.1 *(Uniqueness) Let* $\{X_n\}$,$\{X'_m\}$ *be two approximation sequences, such that* $0 \leq X_n \uparrow X$ *and* $0 \leq X'_m \uparrow X$. *Then*

$$\lim_{n\to\infty} \int X_n dP = \lim_{m\to\infty} \int X'_m dP.$$

Proof: Take an arbitrary $\epsilon > 0$ and let $A_{nm} = [X_n > X'_m - \epsilon]$. Fix an m, and consider $A_{nm} \uparrow \Omega$ as $n \to \infty$. Hence,

$$\int X_n dP \geq \int X_n I_{A_{nm}} dP \geq \int (X'_m - \epsilon) I_{A_{nm}} dP$$

$$= \int X'_m I_{A_{nm}} dP - \epsilon P(A_{nm}) = \int X'_m dP - \int_{A^c_{nm}} X'_m dP - \epsilon P(A_{nm})$$

Let K_m be the upper bound of X'_m. Then

$$\int X_n dP \geq \int X'_m dP - K_m P(A^C_{nm}) - \epsilon P(A_{nm})$$

Let $n \to \infty$. Then for a fixed m,

$$\lim_{n\to\infty} \int X_n dP \geq \int X'_m dP - 0 - \epsilon$$

Since ϵ is arbitrary, $\lim \int X_n dP \geq \int X'_m dP$. Letting $m \to \infty$, we have $\lim \int X_n dP \geq \lim \int X'_m dP$.

Exchange the roles of $\{X_n\}$ and $\{X'_m\}$, we have that

$$\lim \int X_n dP \leq \lim \int X'_m dP.$$

Therefore,

$$\lim \int X_n dP = \lim \int X'_m dP.$$

□

Remark 4.1 All the properties of simple random variables hold for positive random variables.

Now we may consider the integral for an arbitrary random variable.

Definition for Arbitrary R.V.: Let X be an arbitrary random variable. Write

$$X^+ = XI_{[X\geq 0]}, X^- = -XI_{[X<0]}.$$

Note $X = X^+ - X^-$. We then define the expectation of X as

$$\int XdP = \int X^+ dP - \int X^- dP.$$

If each of them on right side is ∞, we say $\int XdP$ doesn't exist; if one of them is finite, and the other is ∞, then $\int XdP$ is either $+\infty$ or $-\infty$, we say the integral exists but X is not integrable; if both are finite, $\int XdP$ is finite, we say X is integrable. By definition, X is integrable is equivalent to $|X|$ is integrable as $|X| = X^+ + X^-$.

If $\{X_n\}$ is a sequence of random variables on $(\Omega, \mathcal{F}, P)$ and let $X_n \to X$ a random variable. Then does

$$\lim_{n\to\infty} E(X_n) = E(\lim_{n\to\infty} X_n)$$

hold always?

In fact, this does **NOT** always hold true! Let $\Omega = [0, 1]$. Consider

$$X_n(\omega) = \begin{cases} n & \omega \in [0, \frac{1}{n}] \\ 0 & \omega \in (\frac{1}{n}, 1] \end{cases}$$

For each $\omega \in [0, 1]$, $X_n(\omega) \to X \equiv 0$. But $E(X_n) = 1$ for any n.

A condition is said to hold *almost surely* (or *almost everywhere*) with respect to a measure μ if and only if there is a set $B \in \mathcal{F}$ such that $\mu(B) = 0$ and the condition holds outside of B. From the perspective of integration, functions that differ only on a set of measure 0 are identical as shown in the following theorem.

Theorem 4.2 *Let X and Y be two random variables and $X = Y$ almost surely. Then if $\int X dP$ exists, so does $\int Y dP$ and*

$$\int X dP = \int Y dP.$$

Proof: see Ash (1972, p.46). □

Exercise 4.1 Let X, Y be two random variables in the probability space $(\Omega, \mathcal{F}, P)$. If

$$\int_C X(\omega) dP(\omega) = \int_C Y(\omega) dP(\omega)$$

holds for any $C \in \mathcal{F}$, then $X = Y$ a.e (with respect to P).

(Hint: If not, then $P(X \neq Y) > 0$. Notice that
$[\omega : X(\omega) \neq Y(\omega)] = \cup_{n=1}^{\infty} [\omega : |X(\omega) - Y(\omega)| \geq 1/n]$ and derive contradiction.)

Similar to the expectation of simple random variables, the expectation of arbitrary random variables has two important properties.

Theorem 4.3 (i) *Monotonicity: If X and Y are two random variables and $X \leq Y$ almost surely, then*

$$\int X dP \leq \int Y dP.$$

(ii) *Linearity: If X and Y are two random variables and α, β are finite real numbers, then*

$$\int (\alpha X + \beta Y) dP = \alpha \int X dP + \beta \int Y dP.$$

Proof: see Billingsley (1995, p.206). □

Theorem 4.4 *(monotone convergence theorem) Suppose $0 \leq X_n \uparrow X$, a random variable. Then*

$$\lim_{n \to \infty} \int X_n dP = \int X dP.$$

Proof: see Billingsley (1995, p.201-202).

An equivalent form of the monotone convergence theorem is as follows.

Remark 4.2 $0 \leq X_n, Y = \sum_{n=1}^{\infty} X_n$, and Y is a random variable then

$$E(\sum_{n=1}^{\infty} X_n) = \sum_{n=1}^{\infty} E(X_n).$$

Proof: $0 \leq \sum_{i=1}^{n} X_i \uparrow Y$. This form is useful as it allows a term-wise integration for the limit sum of a series of positive random variables.

An important consequence of the monotone convergence theorem is known as Fatou's lemma, which has the virtue of no assumptions on the integrand with a one-sided conclusion.

Theorem 4.5 *(Fatou's lemma)* (i) For nonnegative X_n,

$$\int (\liminf_n X_n) dP \leq \liminf_n \int X_n dP.$$

(ii) $Z \geq X_n$, where Z is a integrable random variable, then

$$\int (\limsup_n X_n) dP \geq \limsup_n \int X_n dP.$$

Proof: (i) Let $G_n = inf_{k \geq n} X_k$, then $0 \leq G_n \uparrow G = \lim inf_n X_n$. Then applying the monotone convergence theorem and noticing that

$$\int X_n dP \geq \int G_n dP \to \int G dP = \int \liminf_n X_n dP$$

gives the result.

(ii) Consider $Y_n = Z - X_n$ and apply (i). □

Exercise 4.2 Show if $X_1, X_2, \ldots$ are random variables, $X_n \geq X$ for each n and X is integrable, then

$$\int (\liminf_n X_n) dP \leq \liminf_n \int X_n dP.$$

The inequality in Fatou's lemma can actually be strict manifested by the following example.

Example 4.1 On the unit interval, take $X(\omega) \equiv 0$ and $X_n(\omega) = n^2 I_{[0,n^{-1}]}$. Then for each ω $X_n(\omega)$ converges to $X(\omega)$, but $\int X_n(\omega)dP = n \not\to 0 = \int X(\omega)dP$.

Scrutinizing the phenomenon that $\int X_n(\omega)dP$ does not converge to $\int X(\omega)dP$, one may conjecture that it may be due to the unboundness of $X_n(\omega)$. This is indeed the case as shown by the following *dominated convergence theorem* as a direct application of Fatou's lemma.

Theorem 4.6 *(Dominated Convergence Theorem) If $|X_n| \leq Z$, where Z is a integrable random variable and* $\lim X_n = X$*, then*

$$\lim_{n\to\infty} \int X_n dP = \int X dP.$$

Proof: At the outset only assume that the X_n are dominated by an integrable Z. Let $X_* = \liminf_n X_n$ and $X^* = \limsup_n X_n$. As X^*, X_* are random variables and dominated by integrable Z, hence they are integrable. Since $Z + X_n$ and $Z - X_n$ are nonnegative, Fatou's lemma gives

$$\int ZdP + \int X_* dP = \int \liminf_n (Z+X_n)dP \leq \liminf_n \int (Z+X_n)dP = \int ZdP + \liminf_n \int X_n dP$$

and

$$\int ZdP - \int X^* dP = \int \liminf_n (Z-X_n)dP \leq \liminf_n \int (Z-X_n)dP = \int ZdP - \limsup_n \int X_n dP.$$

Therefore,

$$\int \liminf_n X_n dP \leq \liminf_n \int X_n dP \leq \limsup_n \int X_n dP \leq \int \limsup_n X_n dP.$$

Now consider $X_n \to X$, hence X is dominated by integrable Z. Therefore, X is also integrable and

$$\lim \int X_n dP = \int X dp = \int \lim X_n dP.$$

□

An application of the DCT is differentiation under integration, which shall be discussed in the next section. On the other hand, example 4.1 shows that that this theorem fails if no dominating Z exists. The next result, the *bounded convergence theorem*, is a special case of the DCT when the dominating Z is a constant.

Theorem 4.7 *(Bounded Convergence Theorem) If X_n are uniformly bounded and $\lim X_n = X$, then*

$$\lim_{n\to\infty} \int X_n dP = \int X dP.$$

Exercise 4.3 Let $X_1, X_2, \ldots$ be random variables on $(\Omega, \mathcal{F}, P)$. If

$$\sum_{n=1}^{\infty} \int_{\Omega} |X_n| dP \leq \infty,$$

show that $\sum_{n=1}^{\infty} X_n$ converges everywhere a.e with respect to P and

$$\sum_{n=1}^{\infty} \int_{\Omega} X_n dP = \int_{\Omega} \sum_{n=1}^{\infty} X_n dP.$$

(Hint: first show $\sum_{n=1}^{\infty} |X_n|$ is integrable. Thus, $\sum_{n=1}^{\infty} |X_n|$ is finite a.e. Then, $\sum_{n=1}^{\infty} X_n$ converges almost everywhere. Now consider $Y_n = \sum_{k=1}^{n} X_k$ and verify the conditions in the DCT and apply the DCT.)

Now consider the random variable $g(X)$: $\Omega \to R$, where $g(x)$ is a Borel function $R \to R$. We can calculate $E(g(X)) = \int_{\Omega} g(X) dP$ using the definition of expectation of a random variable.

In the following, we also consider another equivalent representation of the integral. Consider $\int_R g(x) dP_X$, and consider $(R, \mathfrak{B}, P_X)$ playing the role of Basic probability space. If g is a simple function, i.e. $g = \sum c_j I_{B_j}$. Then

$$\int_{\Omega} g(x) dP_X = \sum c_j P_X(B_j).$$

For $g \geq 0$, there exists simple $g_n \uparrow g$, we define

$$\int_R g(x) dP_X = \lim \int_R g_n dP_X.$$

For a general g, we consider $g = g^+ - g^-$ and define

$$\int_R g(x)dP_X = \int_R g^+(x)dP_X - \int_R g^-(x)dP_X.$$

Now we show

Theorem 4.8

$$\int_\Omega g(X)dP = \int_R g(x)dP_X(x)$$

Proof: Let $I_1 = \int_\Omega g(X)dP$ and $I_2 = \int_R g(x)dP_X$. We outline the proof using 3 steps.

Step 1. Suppose that $g(x)$ is a simple function i.e. $g(x) = \sum c_j I_{B_j}(x)$, then

$$I_2 = \sum c_j P_X(B_j) = \sum c_j P(X^{-1}(B_j)),$$

and

$$g(X)(\omega) = g(X(\omega)) = \sum c_j I_{B_j}(X(\omega)) = \sum c_j I_{X^{-1}(B_j)}(\omega)$$

since $w \in X^{-1}(B_j)$ is equal to $X(\omega) \in B_j$.

As $g(X) = \sum c_j I_{X^{-1}(B_j)}$, hence
$I_1 = \int_\Omega g(X)dP = \sum c_j P(X^{-1}(B_j)) = sumc_j P_X(B_j) = I_2$.

Step 2. Suppose that g is a non-negative Borel function, there exists a sequence $\{g_n\}$ of simple Borel functions such that $0 \le g_n \uparrow g$, then by Step 1,

$$I_2 = \lim_{n\to\infty} \int_R g_n(x)dP_X(x) = \lim_{n\to\infty} \int_\Omega g_n(X)dP$$

But $0 \le g_n \uparrow g$ and $g_n(X)$ is a simple random variable $g_n(X)(\omega) = g_n(X(\omega))$,

$$\lim_{n\to\infty} \int_\Omega g_n(X)dP = \int_\Omega g(X)dP = I_1.$$

Step 3. Suppose that g is a general Borel function, then $g = g^+ - g^-$.

By Step 2, $\int_R g^+ dP_X(x) = \int_\Omega g^+(X)dP$ and $\int_R g^- dP_X(x) = \int_\Omega g^-(X)dP$

Therefore, $\int_R (g^+ - g^-)dP_X(x) = \int_R g^+ dP_X(x) - \int_R g^- dP_X(x) =$
$\int_\Omega g^+(X)dP - \int_\Omega g^-(X)dP = \int_R (g^+ - g^-)(X)dP.$ □

Example 4.2 Consider a special case when $g(x) = x$. Then
$E(X) = \int_\Omega X dP = \int_R x dP_X \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} x dF(x)$.

Exercise 4.4 If X is a non-negative random variable with finite expectation, then

$$E(X) = \int_0^\infty P(X \geq x) dx.$$

(Hint: use integration by parts to show $\int_0^\infty x dF(x) = \int_0^\infty [1 - F(x)]dx$.)

Example 4.3 $\int_a^{b-0} x dF(x) = \int_{[a,b)} x dP_X = \int_{-\infty}^{\infty} x I_{[a,b)} dF(x)$. Specially,
$\int_a^{a+0} x dF(x) = \int_{\{a\}} x dP_X = a P_X(\{a\}) = a[F(a+0) - F(a)]$. Note that $\int_a^b$ and $\int_{a+0}^b$ may be different.

Remark 4.3 When the space is real line, the integral is called Lebesgue-Stieltjes integral (LS), whereas for a general Ω it is called abstract Lebesgue-integral.

4.2 Riemann-Stieltjes (R-S) Integral

Riemann-Stieltjes integral is a straightforward extension of Riemann integral and is defined in a similar way.

Definition: For a partition on a finite interval $[a, b]$, say,
$a = x_{n1} \leq x_{n2} \leq \cdots \leq x_{n,m_n+1} = b$, denote by $\underline{g_{nk}} = \inf\{g(x) : x_{nk} \leq x \leq x_{n,k+1}\}$
and $\overline{g_{nk}} = \sup\{g(x) : x_{nk} \leq x \leq x_{n,k+1}\}$. Let

$$\underline{S_n} = \sum_{k=1}^{m_n} \underline{g_{nk}}[F(x_{n,k+1}) - F(x_{n,k})]$$

and

$$\overline{S_n} = \sum_{k=1}^{m_n} \overline{g_{nk}}[F(x_{n,k+1}) - F(x_{n,k})]$$

Let $n \to \infty$ such that the max span of subdivision $\to 0$. If

$$\lim \overline{S_n} = \lim \underline{S_n} = S$$

which is finite, then we call S the *Riemann-Stieltjes integral* and denote

$$S = (RS) \int_a^{b-0} g(x) dF(x).$$

It is trivial to see that the ordinary Riemann integral is a special case when $F(x) \equiv x$.

Remark 4.4 The R-S integral may **NOT** be equal to the Lebesgue-Stieltjes integral.

Example 4.4 Let $P(X = 0) = 1$ and $g(x) = I_{\{x>0\}}$. Consider $\int g(x)dF(x)$. Since $\underline{S_n} = 0, \overline{S_n} = 1$, the R-S integral doesn't exist, but $g(0) = 0, \int g(x)dF(x) = 0 \times 1 = 0$

We give below a classical example of a function that is Lebesgue integrable but not R-S integrable.

Example 4.5 $F(x) = x, 0 \leq x \leq 1, g(x) = I_A(x), A =$ set of rational numbers in $[0, 1]$. $\underline{S_n} = 0, \overline{S_n} = 1$ (RS) integral doesn't exist, but $\int g(x)dF(x) = 1 \times P(A) = 0$.

Example 4.6 Define a function $f(x) = \frac{1}{x} sin\frac{1}{x}$ on an open interval $(0, 1)$. Then $\int_0^1 f(x)dx$ is RS integrable (in the sense of improper integral). But as $f(x)$ is not absolutely integrable, its Lebesgue integral does not exist.

Under certain conditions, however, these two types of intergals are equivalent.

Theorem 4.9 *If g is a continuous function a.e (with respect to the Lebesgue measure) on a finite interval $[a, b]$, then $\int_a^b g(x)dF(x)$is the same in both RS and LS. For $(-\infty, +\infty)$, if g is continuous, the two integrals are same, provided that g is LS-integrable. In addition, that g is Lebesgue integrable is equivalent to $|g|$ is Lebesgue integrable.*

Proof: see Ash (1972, p.55). □

Finally, the following exercise shows that the DCT may not always hold for Riemann integrals.

Exercise 4.5 Give an example of a sequence of functions f_n on $[0, 1]$ such that each f_n is Riemman integrable, $|f_n| \leq 1$ for all n and $f_n \to f$ everywhere, but f is not Riemann integrable.

4.3 Moments

Expectations of certain functions of X are of particular interest.

Definition: Let a be a real number, r positive, then $E(|X-a|^r)$ is called the *absolute moment of X of order r about a* . It can be ∞; otherwise, if r is an integer, $E((X-a)^r)$ is the corresponding *moment* about a. (In literature, the *moment* about the origin, $E(X^r)$, is often called *moment*, and the absolute moment about the origin, $E|X|^r$, the *absolute moment*). The moments about the mean is called the *central moments.* That of order 2 is particularly important and is called the *variance*, $var(X)$, and its positive square root the *standard deviation.* We define *factorial moments* by $E(X(X-1)\cdots(X-r+1))$ when r is an integer.

We say the $r-$th moment exists if $E(X^r)$ is finite; otherwise if $E(X^r)$ is either $+\infty$ or $-\infty$, we say that it doesn't exist.

Some properties of moments are

Property 4.6 $E(X^r)$ is finite if and only if $E(|X|^r)$ is finite. Also, finiteness of the r-th moment implies finiteness of lower moments, i.e. $E(X^r)$ is finite implies $E(X^s)$ is finite for any $0 < s < r$.

Proof: The first assertion is true as that X^r is integrable is equivalent to that $|X^r|$ is integrable.

Now Let $0 < s < r$, $E(|X|^r)$ is finite and $|X|^s \leq 1 + |X|^r$. In fact, when $|X| \leq 1, |X|^s \leq 1$ and if $|X| > 1, |X|^s \leq |X|^r$. □

Property 4.7 That $E(X^r)$ is finite implies $P(|X| > n) = o(\frac{1}{n^r})$, as $n \to \infty$. i.e. $n^r P(|X| > n) \to 0$ as $n \to \infty$.

Proof: Write $v_r = E(|X|^r)$. Consider

$$v_r = \int_{[|X| \leq n]} |X|^r dP + \int_{[|X| > n]} |X|^r dP$$

and

$$n^r P(|X| > n) \leq \int_{[|X| > n]} |X|^r dP = v_r - \int_{\Omega} |X|^r I_{[|X| \leq n]} dP.$$

Define $Y_n = |X|^r I_{[|X|\le n]}$. Hence $0 \le Y_n \uparrow |X|^r$. By MCT

$$\int Y_n dP \to \int |X|^r dP = v_r, \text{as } n \to \infty$$

or, equivalently,

$$v_r - \int_\Omega |X|^r I_{[|X|\le n]} dP \to 0, \text{as } n \to \infty$$

Hence,

$$n^r P(|X| > n) \to 0,$$

as $n \to \infty$. □

Remark 4.5 Conversely, $n^r P(|X| > n) \to 0$ as $n \to \infty$ implies that $E(|X|^{r-\epsilon})$ is finite for any $0 < \epsilon < r$.

Proof: home work. (Hint: Consider $\int_\Omega |X|^{r-\epsilon} dP = \sum_{n=0}^\infty \int_{n<|X|\le n+1} |X|^{r-\epsilon} dP$. Then show,

$$\begin{aligned}
\int_\Omega |X|^{r-\epsilon} dP &\le \sum_{n=0}^\infty (n+1)^{r-\epsilon} P(n < |X| \le n+1) \\
&= \sum_{n=0}^\infty (n+1)^{r-\epsilon} (P(|X| > n) - P(|X| > n+1)) \\
&= \sum_{n=0}^\infty ((n+1)^{r-\epsilon} - n^{r-\epsilon}) P(|X| > n) \stackrel{def}{=} \sum_{n=0}^\infty a_n
\end{aligned}$$

Then prove that $\sum_{n=0}^\infty a_n$ is convergent.) □

We end this section with two well-known inequality and we leave their proofs as exercises.

Exercise 4.6 (Holder's Inequality) For $0 \le p < \infty$, let $L^p = L^p(\Omega, \mathcal{F}, P)$ be the class of random variables X for which $E(|X|^p)$ is finite. Define $||X||_p = (E(|X|^P))^{1/p}$. For $1 < p, q < \infty$, if $1/p + 1/q = 1$ and $X \in L^p, Y \in L^q$, show

$$||XY||_1 = E|XY| \le ||X||_p ||Y||_q.$$

Thus the Schwartz inequality is only a special case with $p = q = 2$. (Hint: Using the convexity of $-\log(x)$, show that, for any $c, d > 0$,

$$-\log(\frac{1}{p}c^p + \frac{1}{q}d^q) \leq \frac{1}{p}(-\log c^p) + \frac{1}{q}(-\log d^q).$$

Hence $cd \leq c^p/p + d^q/q$. Then notice that $\frac{|XY|}{||X||_p||Y||_q} \leq \frac{|X|^p}{p||X||_p^p} + \frac{|Y|^q}{q||X||_q^q}$.)

Exercise 4.7 (Minkowski's Inequality) If $X, Y \in L^P (1 \leq p < \infty)$, then $X + Y \in L^P$ and

$$||X + Y||_p \leq ||X||_p + ||Y||_p.$$

(Hint: The claim is obvious when $p = 1$. When $p > 1$, find q such that $1/p + 1/q = 1$. Then $p/q = p - 1$. Consider $||X + Y||_p^p = \int |X + Y|^p dP \leq \int |X||X + Y|^{\frac{p}{q}} dP + \int |Y||X + Y|^{\frac{p}{q}} dP$ and apply the Holder's inequality.)

5 Two Important Inequalities

5.1 Markov Theorem

Theorem 5.1 *(Markov Theorem) Let $X : (\Omega, \mathcal{F}, P) \to (R, \mathfrak{B})$. $g(X)$ is a Borel function $R \to R$ such that $g(x) \geq 0$, and for any x in a set $B \in \mathfrak{B}, g(x) \geq K > 0$, then*

$$P(X \in B) \leq \frac{E(g(X))}{K}.$$

Proof: As $P(X \in B) = P_X(B)$, so

$$E(g(X)) = \int_B g(X) dP_X + \int_{B^c} g(X) dP_X \geq K P_X(B).$$

□

Using the Markov theorem, one can trivially prove the very famous Chebyshev's inequality.

Theorem 5.2 *(Chebyshev's Inequality) If X is a random variable such that$E(X) < \infty$ and $var(X) < \infty$, then for each $u > 0$*

$$P(|X - E(X)| \geq u) \leq \frac{var(X)}{u^2}. \tag{7}$$

Proof: let $\mu = E(X)$ and take $g(x) = (x-\mu)^2$. □

Suppose that a random variable X assumes values $m-a, m, m+a$ with probabilities $p, 1-2p, p$. It can be shown that there is an equality in (7). Hence, Chebyshev's inequality can not be improved without special assumptions on X.

An important application of Chebyshev inequality is in the proof of the *Weak Law of Large Numbers.* Before stating the main theorem, we begin with the concept of *convergence in probability.*

Definition: For a sequence of random variable X_n and a random variable X, we say that X_n converges to X in probability as $n \to \infty$ if and only if for any given $\epsilon > 0$,

$$P(|X_n - X| > \epsilon) \to 0, \qquad n \to \infty.$$

Theorem 5.3 *(Weak Law of Large Numbers) Let $X_1, X_2, \ldots$ be independent random variables (not necessarily with the same distribution), each with finite mean and variance. Assume $var(X_i) < M$ for a fixed $M > 0$ and any $i = 1, 2 \ldots$. Let $S_n = X_1 + \ldots + X_n$. Then $(S_n - E(S_n))/n$ converges in probability to 0, that is, for any given $\epsilon > 0$,*

$$\lim_{n\to\infty} P(|\frac{S_n - E(S_n)}{n}| > \epsilon) = 0.$$

Proof: By Chebyshev's inequality,

$$\begin{aligned} P(|\frac{S_n - E(S_n)}{n}| > \epsilon) &\leq \frac{1}{\epsilon^2} var(\frac{S_n}{n}) \\ &= \frac{1}{\epsilon^2 n^2} var(S_n) \\ &= \frac{1}{\epsilon^2 n^2} \sum_{i=1}^{n} var(X_i) \\ &\leq \frac{M}{\epsilon^2 n} \to 0. \end{aligned}$$

□

Convergence in probability is one of the convergence modes for sequences of random variables. The other modes include convergence almost surely, convergence in distribution and convergence in moment. We briefly talk about the important

concept of convergence almost surely and the other convergence modes and their inter-relationships will be left to discuss later.

Definition of Convergence Almost Surely: Let $(\Omega, \mathcal{F}, \mu)$ be a space, and f a measurable function on $\Omega \to R$, i.e. $f^{-1}(\mathfrak{B}) \subset \mathcal{F}$. If $\{f_n\}$ is a sequence of measurable functions, we let $A = \{\omega : f_n(\omega) \to f(\omega)\}$. We say $f_n \to f$ a.e (with respect to μ), if $\mu(A^c) = 0$.

It should be stressed that the concept of convergence almost surely is always with respect to a particular measure.

Example 5.1 Let $\Omega = R^+$,$\mathcal{F} = \mathfrak{B}$ and $\mu = \mu_L$. If

$$f_n(x) = \begin{cases} e^{-\lambda_n} \frac{\lambda_n^x}{x!} & \text{for } x \in N^+ \\ 0 & \text{otherwise} \end{cases}$$

where $\lambda_n \to \lambda$. Let $f(x) = 0$ for any $x \in R^+$. Then $f_n \to 0$ a.e (with respect to μ_L) but $f_n \nrightarrow 0$ a.e (with respect to μ^*) where μ^* is counting measure of integers. In fact, $f_n \to g$ a.e (with respect to μ^*) where $g(x) = e^{-\lambda} \frac{\lambda^x}{x!}$ (on N^+).

Exercise 5.1 If $X_1, X_2, \ldots, \in L^p(\Omega, \mathcal{F}, \mu)$ $(p > 0)$ and $||X_n - X_{n+1}||_p < (\frac{1}{4})^n, n = 1, 2 \ldots,$, then $\{X_n\}$ converges a.e. (with respect to μ). (Hint: let $A_n = \{\omega : |X_n(\omega) - X_{n+1}(\omega)| \geq 2^{-n}\}$ Use the Markov inequlity to show $\mu(A_n) \leq 2^{-np}$. By the first Borel-Cantelli lemma, $\mu(\limsup_n A_n) = 0$. But if $\omega \notin \limsup_n A_n$, then $|X_k(\omega) - X_{k+1}(\omega)| < 2^{-k}$ for large k, so $X_n(\omega)$ is a Cauchy sequence, and hence converges.)

With little modification in the previous proofs, it follows that the Monotone Convergence theorem and Dominated Convergence theorem in the last section are also valid if the condition of pointwise convergence is relaxed to be that of convergence almost everywhere or almost surely.

Theorem 5.4 *(MCT) If* $0 \leq X_n \uparrow X$ *a.s, then* $E(\lim X_n) = \lim E(X_n)$.

Remark 5.1 A general statement for the MCT, not necessarily under the probability measure, is:

$$\text{If } 0 \leq f_n \uparrow f a.e(\text{with respect to } \mu), \text{ then } \lim_{n\to\infty} \int f_n d\mu = \int f d\mu.$$

Similarly the Dominated Convergence theorem can be restated

Theorem 5.5 *(DCT) If $|f_n| \leq g$ for any n, and $\int g d\mu$ is finite. and $f_n \to f$ a.e (with respect to μ). then $\int f_n d\mu \to \int f d\mu$.*

The following theorem concerning interchanging integral and differentiation is a direct result of the DCT.

Theorem 5.6 *(Differentiation Under Integral) Given any measurable space $(\Omega, \mathcal{F}, \mu)$, let I be an interval in R and $\{f_\theta, \theta \in I\}$ be a class of measurable functions indexed by θ such that $f_\theta : \Omega \to R$. If $\int f_\theta d\mu$ is defined for any θ, then under conditions: (1) $\frac{df_\theta}{d\theta}$ exists at θ_0 and (2) $|\frac{f_\theta - f_{\theta_0}}{\theta - \theta_0}| \leq g_{\theta_0}$ for all θ in a small neighborhood of θ_0 and $\int g_{\theta_0} d\mu$ is finite,*

$$(\frac{d}{d\theta} \int f_\theta d\mu)|_{\theta_0} = \int (\frac{df_\theta}{d\theta})|_{\theta_0} d\mu$$

Proof: homework. (Hint: for any real sequence $\theta_n \to \theta_0$, consider

$$\frac{\int f_{\theta_n} d\mu - \int f_{\theta_0} d\mu}{\theta_n - \theta_0} = \int \frac{f_{\theta_n} - f_{\theta_0}}{\theta_n - \theta_0} d\mu$$

and apply DCT.) □

If the conditions above are replaced by stronger conditions: (1) $\frac{df_\theta}{d\theta}$ exists at θ_0. (2) $|\frac{df_\theta}{d\theta}| \leq g$, independent of θ, and $\int g d\mu$ is finite. Then the above operation is valid at every $\theta \in I$.

5.2 Jensen's Inequality

Definition of Convex function: A Borel function $g : R \to R$ is convex in an interval $I \subset R$ if $g(\lambda x_1 + (1-\lambda)x_2) \leq \lambda g(x_1) + (1-\lambda) g(x_2)$, for any $x_1, x_2 \in I$ and any $0 \leq \lambda \leq 1$.

The basic properties of a convex function g are:

(1) g is continuous on I.

(2) For any $x \in I$, the right and left derivatives exists

(3) If g is twice differentiable, then $g''(x) \geq 0$ for any $x \in I$ and, conversely, a positive second derivative implies the function is convex.

(4) For any $x_0 \in I$, there exists a number $L(x_0)$ such that
$g(x) - g(x_0) \geq (x - x_0)L(x_0)$ for any $x \in I$.

Theorem 5.7 *(Jensen's inequality) Let X be a random variable, $g : R \to R$ is a Borel function which is convex on R and $E(X)$ finite, then $E(g(X)) \geq g(E(X))$.*

Proof: Use property (4), with $x_0 = E(X)$

$$g(X) - g(E(X)) \geq (X - E(X))L(E(X)),$$

and then take expectations on both side. □

Remark 5.2 The condition on g in the theorem can be weakened to be "g is convex on an interval I such that $P_X(I) = 1$".

We give below two simple but important applications of Jensen's inequality.

Example 5.2 Let X be a non-negative random variable. Choose $g(x) = \frac{1}{x}$ and $I = (0, \infty)$. Then $E(\frac{1}{X}) \geq \frac{1}{E(X)}$.

Example 5.3 Denote by $v_r = E(|X|^r)$. Let $g(r) = \log E|X|^r = \log v_r$, $r \geq 0$. Then $g(r)$ is a convex function of r. Therefore, $[v_r]^{\frac{1}{r}}$ is increasing as a function of r, i.e.

$$E|X| \leq v_2^{\frac{1}{2}} \leq v_3^{\frac{1}{3}} \leq \cdots$$

Proof: homework. □

Exercise 5.2 Let $f(x, y)$ be a convex real function on the two-dimensional plane. Show that f is convex if it has continuous second derivatives that satisfy

$$f_{11} \geq 0, f_{22} \geq 0, f_{11}f_{22} \geq f_{12}^2.$$

Then show function $f(x, y) = y^2 - 2xy$ is convex in each variable separately but not convex on the plane.

6 The Radon-Nikodym Derivative (Density Function)

6.1 Abstract Continuity and Probability Density Function

Given a triplet $(\Omega, \mathcal{F}, \mu)$, where μ is a measure on $(\Omega, \mathcal{F})$ (it may not be a probability measure), let $f : \Omega \to R^+ = [0, \infty)$ and be measurable. We define a new set function

$$P(A) = \int_A f d\mu,$$

for any $A \in \mathcal{F}$.

One may easily observe that

(1) P is a measure on $(\Omega, \mathcal{F})$.

Proof: Let $A = \sum A_j$. To show $P(A) = \sum P(A_j)$, we write

$$P(A) = \int f I_A d\mu = \int f I_{\sum A_j} d\mu$$

$$= \int \sum_{j=1}^{\infty} f I_{A_j} d\mu = \sum_{j=1}^{\infty} \int f I_{A_j} d\mu = \sum P(A_j).$$

The last equality holds by MCT. □

Here, f is called the *Radon-Nikodym derivative* of P with respect to μ and write as

$$f = \frac{dP}{d\mu} \text{ or } dP = f d\mu$$

(2) $\mu(A) = 0$ implies $P(A) = 0$.

Proof: homework. □

In particular if P is a probability measure ($\int_\Omega f d\mu = 1$), f is called probability density function of P with respect to μ.

6.2 The Radon-Nikodym Theorem

Now we know if $P(A) = \int_A f d\mu$, then certainly $\mu(A) = 0$ implies $P(A) = 0$. But can we go in the opposite direction? That is, given a probability P and a measure μ

such that $\mu(A) = 0$ implies $P(A) = 0$ for any $A \in \mathcal{F}$, is there an f such that $P(A) = \int_A f d\mu$? The answer is yes as indicated by the Radon-Nikodym theorem.

Before going into the main theorem, we start with some terminologies.

Definition: Let μ and v be two measures on $(\Omega, \mathcal{F})$. v is called *absolutely continuous* with respect to μ if and only if $\mu(A) = 0$ implies $v(A) = 0$ for any $A \in \mathcal{F}$. In this case v is often called *dominated* by μ, denoted by $v << \mu$.

Definition: v and μ are called mutually singular if there exists a set $N \in \mathcal{F}$ such that $\mu(N) = 0, v(N^c) = 0$.

Remark 6.1 These two definitions can not both hold except for the trivial measure.

Example 6.1 Consider the space $(R^+, \mathfrak{B}^+)$. Let μ^* be a counting measure of integers and P a probability measure of binomial distribution $b(n, p)$. That is,

$$P\{x\} = \frac{n!}{x!(n-x)!} p^x q^{n-x}, x = 0, 1, 2, \cdots, n.$$

Then

$$P << \mu^*.$$

But how about the other way around, i.e. $\mu^* << P$? In fact, this is not true. Since $P(\{n+1\}) = 0,$ but $\mu^*(\{n+1\}) = 1.$

Example 6.2 Let P^0 be a probability measure of a Poisson distribution, i.e.

$$P^0(x) = e^{-\lambda} \frac{\lambda^x}{x!}, x = 0, 1, 2, \cdots,$$

then $P^0 << \mu^*$ and $\mu^* << P^0$. Hence, P^0 and μ are mutually absolutely continuous.

Example 6.3 Define P^0 as above and let $\mu = \mu_L$ be the Lebesgue measure. For $N = \{0, 1, 2, \cdots\} \in \mathfrak{B}$, $P^0(N^c) = 0$ and $\mu_L(N) = 0$ imply that P^0 and μ_L are mutually singular.

Theorem 6.1 *(The Radon-Nikodym Theorem) Let μ and v be two $\sigma-$finite measures on $(\Omega, \mathcal{F})$. $v << \mu$ if and only if there exists a finite non-negative measurable function f such that $v(A) = \int_A f d\mu$ for any $A \in \mathcal{F}$. Also, f is uniquely determined except possibly on a $\mu-$null set.*

Proof: see Ash (1972, p.63-65). □

We now consider an important application of the theorem on the real line. Consider $(\Omega, \mathcal{F}) = (R, \mathfrak{B})$. Let $v = P_X$ be a probability measure induced by a random variable and $\mu = \mu_L$ the Lebesgue measure. Suppose that $P_X << \mu_L$, then by the R-N Theorem, there exists an $f \geq 0$ such that $P_X(B) = \int_B f d\mu_L$. Furthermore

$$F_X(x) = P_X(-\infty, x) = \int_{-\infty}^{x-0} f d\mu_L \overset{\text{convention}}{=} \int_{-\infty}^{x} f(t)dt.$$

In this case, we also say P_X or F is *absolutely continuous* with respect to μ_L and $f(x)$ is the probability density function of X.

Example 6.4 Let $(X_1, \cdots, X_K)$ be a random vector on $(R^k, \mathfrak{B}^k)$ and μ_L^k a $k-$dimensional Lebesgue measure. Let $P_X << \mu_L^k$ then

$$F(x_1, \cdots, x_k) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_k} f(x_1, x_2, \cdots, x_k) dx_1 \cdots dx_k.$$

Example 6.5 Consider again $(R, \mathfrak{B})$. Let μ^* be a counting measure of integer and P_X a distribution for binomial $b(n, p)$. As $P_X << \mu^*$, hence

$$F(x) = \int_{-\infty}^{x-0} f(x) d\mu^*$$

by the R-N theorem, where

$$f(x) = \begin{cases} \frac{n!}{x!(n-x)!} p^x q^{n-x}, & x = 0, 1, \cdots, n \\ 0 & \text{otherwise} \end{cases}$$

is the probability density function of $b(n, p)$, with respect to μ^*.

When the counting measure is considered, we call the random variable X a *discrete* type. Another commonly used definition for a discrete random variable is as follows.

Definition: A random variable X is called *discrete* if and only if there is a countable set $B \subset R$ such that $P(X \in B) = 1$.

Remark 6.2 That F is absolutely continuous implies that F is continuous.

Proof:

$$F(x) = \int_{-\infty}^{x-0} f(t)dt = \int_{-\infty}^{x} f(t)dt = F(x+0)$$

since $\mu_L(\{x\}) = 0$. □

The following example shows that the other direction usually is not true!

Example 6.6 Note that $F(x, y) = \min(x, y)$ on the square $[0, 1] \times [0, 1]$ is a continuous function. But F is not absolutely continuous with respect to $\mu_L^{(2)}$. To see this, notice that $F(x, y)$ is the c.d.f for a random variable which uniformly distributed on the diagonal segment of the unit square. Then the $F-$measure for the diagonal segment, a $\mu_L^{(2)}$-null set, is nonzero.

We construct below a random variable which has a continuous distribution, but does not have a density function. This example was given in Feller (V2 Section 1.11).

Example 6.7 Let Y_k be mutually independent random variables assuming 1 and 0 with probability $\frac{1}{2}$. Let $X = 3\sum_{k=1}^{\infty} \frac{Y_k}{4^k}$ (think X be the gain of a gambler receiving the amount of 3×4^{-k} if the k-th toss yields a head or 0 if the k-th toss yields a tail). Hence,

$$0 \le X \le 3\sum_{k=1}^{\infty} \frac{1}{4^k} = 1.$$

If $Y_1 = 1$, then $X = \frac{3}{4} + \cdots \ge \frac{3}{4}$, otherwise if $Y_1 = 0$ then $X \le 3\sum_{k=2}^{\infty} \frac{1}{4^k} = \frac{1}{4}$. Then $P(X \le \frac{1}{4}) = \frac{1}{2}$ and $P(X \ge \frac{3}{4}) = \frac{1}{2}$, which implies that $P(\frac{1}{4} < X < \frac{3}{4}) = 0$. Let $F(x)$ be the c.d.f of X. Hence, $F(x) = \frac{1}{2}$ for any $x \in (\frac{1}{4}, \frac{3}{4}]$ and $F(x)$ has no jump exceeding $\frac{1}{2}$.

We may further calculate that:

$$P(Y_1 = 1, Y_2 = 1) = \frac{1}{4} \Longrightarrow P(X \ge \frac{15}{16}) = \frac{1}{4}$$

$$P(Y_1 = 1, Y_2 = 0) = \frac{1}{4} \Longrightarrow P(\frac{3}{4} \le X \le \frac{13}{16}) = \frac{1}{4}$$

$$P(Y_1 = 0, Y_2 = 1) = \frac{1}{4} \Longrightarrow P(\frac{3}{16} \le X \le \frac{1}{4}) = \frac{1}{4}$$

$$P(Y_1 = 0, Y_2 = 0) = \frac{1}{4} \Longrightarrow P(X \le \frac{1}{16}) = \frac{1}{4}.$$

Consider that

$$F(x) = \begin{cases} \frac{1}{4} & x \in (\frac{1}{16}, \frac{3}{16}], \\ \frac{1}{2} & x \in (\frac{1}{4}, \frac{3}{4}], \\ \frac{3}{4} & x \in (\frac{13}{16}, \frac{15}{16}]. \end{cases}$$

Hence, $F(x)$ has no jump exceeding $\frac{1}{4}$.

Inductively, at the $n-$stage, we may show that $F(x)$ has no jump exceeding $\frac{1}{2^n}$. Let

$$A_n = (\frac{1}{4}, \frac{3}{4}] + \{(\frac{1}{16}, \frac{3}{16}] + (\frac{13}{16}, \frac{15}{16}]\} + \cdots + \{(\frac{1}{4^n}, \frac{3}{4^n}] + \cdots + (\frac{4^n-3}{4^n}, \frac{4^n-1}{4^n}]\}.$$

Hence

$$\mu_L(A_n) = \frac{1}{2} + \frac{1}{2^2} + \cdots + \frac{1}{2^n} = 1 - \frac{1}{2^n}.$$

But $P_X(A_n) = P(X \in A_n) = P(X \in (\frac{1}{4}, \frac{3}{4}]) + \ldots + P(\frac{4^n-3}{4^n}, \frac{4^n-1}{4^n}]) = 0$ and $F(x)$ has no jump exceeding $\frac{1}{2^n}$. Let $n \to \infty$. So $F(x)$ is continuous.

As $A_n \to A = \cup_{i=1}^{\infty} A_i$, $P_X(A) = 0$. But $\mu_L(A) = \lim_{n\to\infty} \mu_L(A_n) = 1$. So P_X and μ_L are mutually singular. Hence, P_X can not be absolutely continuous with respect to μ_L or have a density function. □

In fact, $F(x)$ is a continuous function increasing from $F(0) = 1$ to $F(1) = 1$ in such a way that the intervals of constancy add up to length 1. Roughly speaking F increases only in a μ_L-null A^c, which is like a Cantor set.

Cantor Set: The Cantor set is constructed as follows: from $[0, 1]$ remove the open middle third $(\frac{1}{3}, \frac{2}{3})$; from the remainder, a union of two closed intervals, remove the two open middle thirds, $(\frac{1}{9}, \frac{2}{9})$ and $(\frac{7}{9}, \frac{8}{9})$. The Cantor set is what remains when this process is continued infinitely. It can be shown that the Cantor set is uncountable but has (Lebesgue) measure 0.

6.3 Absolute Continuous Cumulative Distribution Function

The previous section stated that $F(x)$ is called an absolute continuous cumulative distribution function for a random variable X if $F(x)$ can be written in a Lebesgue integral form

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

where $f(x)$ is a nonnegative real Borel function, called the probability density function of X. If f is continuous, this integral is a Riemann integral. Since $F(x) \to 1$ as $x \to \infty$, we have $\int_{-\infty}^{\infty} f(t)dt = 1$.

Furthermore, if $F(x)$ is absolutely continuous with respect to the probability density function $f(x)$, then $\frac{dF(x)}{dx}$ exists a.e (with respect to μ_L) and $\frac{dF(x)}{dx} = f(x)$

Let us consider a pair of random variables, (X_1, X_2). Assume its cumulative distribution function $F(x_1, x_2)$ is absolutely continuous with respect to $\mu_L^{(2)}$. Then by the R-N Theorem, there exists a finite (a.e) $f(x_1, x_2) \geq 0$ such that

$$F(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f(y_1, y_2) d\mu_L^{(2)}$$

By convention, this Lebesgue integral can be written

$$\int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f(y_1, y_2) dy_1 dy_2.$$

Also, $\frac{\partial^2 F}{\partial x_1 \partial x_2} = f(y_1, y_2)$ a.e (with respect to $\mu_L^{(2)}$).

We next derive the marginal distribution for X_1. As

$$\begin{aligned} F_1(x_1) &= P(X_1 < x_1) = P(X_1 < x_1, X_2 < \infty) \\ &= \int_{-\infty}^{x_1} \int_{-\infty}^{+\infty} f(y_1, y_2) dy_2 dy_1 = \int_{-\infty}^{x_1} g(y_1) dy_1, \end{aligned}$$

where $g(y_1) = \int_{-\infty}^{+\infty} f(y_1, y_2) dy_2$, hence, $g(y)$ is the probability density function of X_1.

Example 6.8 Consider a mixed distribution of $F_1(x) : N(0, 1) << \mu_L$ and $F_2(x) : b(n, p) << \mu^*$, such that $F(x) = \frac{1}{2}F_1(x) + \frac{1}{2}F_2(x)$. Let $\mu_0 = \mu_L + \mu^*$, i.e. $\mu_0(A) = \mu_L(A) + \mu^*(A)$, then μ_0 is a measure too. One may show

$$F(x) << \mu_0.$$

In fact, $\mu_0(B) = 0$ implies $\mu_L(B) = 0$ and $\mu^*(B) = 0$. Hence $P_1(B) + P_2(B) = 0$.

Then from the N-R theorem, there exists a probability density function $f(x)$ of $F(x)$ with respect to μ_0 such that $F(x) = \int_{-\infty}^{x} f(y)dy$. It turns out

$$f(x) = \frac{1}{2}(\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} I_{A^c} + \binom{n}{x} p^x q^{n-x} I_A)$$

where $A = \{0, 1, 2, \cdots, n\}$. (homework: verify this.)

Exercise 6.1 If F is an absolutely continuous distribution function, show

$$\int_{-\infty}^{\infty} F(x+c) - F(x)dx = c$$

for any $c \in R$.

7 Transformation of Random Variables

Our goal here is to find the distribution of functions of random variables. Let $X = (X_1, \cdots, X_k)$ be a k-dimensional random vector. Suppose $g = (g_1, \ldots, g_l) : R^k \to R^l$. Consider

$$Y_i = g_i(X), i = 1, \ldots, l.$$

A question of particular interest is what the distribution of $Y = (Y_1, \cdots, Y_l)$ is.

Assume that $X = (X_1, \cdots, X_k)$ has a cumulative distribution function $F(x)$ and a probability density function $f(x)$. To find the probability density function (or cumulative distribution function) of $g(X)$, we use the definition of cumulative distribution function, i.e.

$$F_Y(y) = P(g(X) < y) = P_X(g^{-1}(-\infty, y)) = \int_{g^{-1}(-\infty, y)} dF(x).$$

For illustration purposes, some typical cases are stratified below.

Case 1: X has a discrete distribution with probability density function $f_X(x)$ and $Y = g(X)$, where g is a Borel function $R^l \to R$. Here, l is the dimension of X.

As X is discrete, $S = \{x : f_X(x) > 0\}$ is countable. Let $S' = g(S)$ and denote by $A_y = \{x \in S, g(x) = y\}$. Then for any $y \in S'$

$$f_Y(y) = \begin{cases} \sum_{x \in A_y} f_X(x), & y \in S', \\ 0, & y \notin S'. \end{cases}$$

The key step is $[\omega : Y(\omega) = y] = [\omega : X(\omega) \in A_y]$.

Example 7.1 Calculate the distribution of the sum of two independent Poisson random variables. Suppose that $X = (X_1, X_2)$ are independent Poisson random variables with distribution parameters λ_1, λ_2. Consider $Y = g(X) = X_1 + X_2$. So here $g : R^2 \to R$. Let $N^+ = \{1, 2, \cdots, \}$. Then $S = N^+ \times N^+$ and $S' = g(S) = N^+$. Hence, for any $y \in S' = N^+$

$$f_Y(y) = \sum_{\substack{(x_1,x_2)\in N^+\times N^+ \\ x_1+x_2=y}} e^{-\lambda_1}\frac{\lambda_1^{x_1}}{x_1!}e^{-\lambda_2}\frac{\lambda_2^{x_2}}{x_2!} = e^{-\lambda}\frac{\lambda^y}{y!}.$$

where $\lambda = \lambda_1 + \lambda_2$.

Case 2: X is absolutely continuous with probability density function $f_X(x)$ and $Y = g(X)$. Here, g is a discrete function, i.e. a function whose range is at most countable.

Let $S = \{x, f_X(x) > 0\}$. Suppose that $S' = g(S)$ is a countable set, say $\{y_1, y_2, \cdots\}$. Let $S_i = \{x \in S, g(x) = y_i\} = g^{-1}(\{y_i\}) \cap S$. Then

$$f_Y(y_i) = P(Y = y_i) = \int_{S_i} f_X(x)dx, i = 1, 2, \cdots .$$

Example 7.2 Suppose that $X = (X_1, X_2, X_3)$ are independent identically distribution with a common probability density function $e^{-x}, 0 < x < \infty$. Thus, the probability density function of X is

$$f_X(x) = \begin{cases} e^{-(x_1+x_2+x_3)} & , \text{ on } R^{3+} \\ 0 & \text{otherwise.} \end{cases}$$

Now $g(X) = Y =$ the rank of X_1. Let $S = \{(x_1, x_2, x_3) : x_i > 0, i = 1, 2, 3\}$. Then $P_X(S) = 1$.
On S, rank is well-defined $S' = g(S) = \{1, 2, 3\}$. Then one may calculate the distribution of Y on S'. For example,

$$P(Y = 2) = f_Y(2) = \int_{\substack{x_2<x_1<x_3 \\ x_3<x_1<x_2}} e^{-(x_1+x_2+x_3)} = \frac{1}{3}.$$

Case 3:

(i) X is absolutely continuous with a probability density function $f_X(x)$ and g is one to one. Let $S = \{x : f_X(x) > 0\}$ and $S' = g(S)$. We further assume that $g^{-1}(y)$ has a continuous first order derivative on S'. Then $Y = g(X)$ is also absolutely continuous and has a cumulative function

$$F_Y(t) = P(g(X) < t) = \int_{g^{-1}(-\infty,t)\cap S} f_X(x)dx$$
$$= \int_{-\infty}^{t} f_X(g^{-1}(y))|\frac{dg^{-1}(y)}{dy}|dy = \int_{-\infty}^{t} f_Y(y)dy.$$

Therefore, $f_Y(y) = f_X(g^{-1}(y))|\frac{dg^{-1}(y)}{dy}|$ is the probability density function of Y.

(ii) With the same set up as above, except that g is not one to one on $S \to S'$, we assume that: (a) there is a partition of $S = \sum_{l=1}^{L} S_l$ such that g restricted to each S_i is one to one onto a set S_l': $g_l : S_l \overset{\text{one to one}}{\longrightarrow} S_l'$. g_l is restriction of g to the domain of S_l, $S' = \cup_{l=1}^{L} S_l'$ and S_l' need not be disjoint; (b) each g_l^{-1} has a continuous derivative on S_l'. Then the probability density function of Y is

$$f_Y(y) = \sum_{i=1}^{k} I_{S_i'}(y) f_X(g_i^{-1}(y))|\frac{dg_i^{-1}(y)}{dy}|.$$

Proof: Consider $F_Y(t) = P(g(X) < t)$. As $A = \{x \in S : g(x) < t\} = \sum AS_i$, hence

$$F_Y(t) = \int_A f_X(x)dx = \sum_{i=1}^{k} \int_{AS_i} f_X(x)dx$$
$$= \sum_{i=1}^{k} \int_{(-\infty,t)\cap S_i'} f_X(g_i^{-1}(y))|J_i(y)|dy$$
$$= \sum_{i=1}^{k} \int_{-\infty}^{t} I_{S_t'}(y) f_X(g_i^{-1}(y))|J_i(y)|dy$$
$$= \int_{-\infty}^{t} \sum_{i=1}^{k} I_{S_t'}(y) f_X(g_i^{-1}(y))|J_i(y)|dy.$$

□

Exercise 7.1 Let $X \sim N(0,1)$ and $g(x) = (x-2)^2$. Find the probability density function of $g(x)$.

Exercise 7.2 Let X and Y be independent and follow $N(0, \sigma^2)$. Define random variables R and Θ by $X = R\cos(\Theta)$ and $Y = R\sin(\Theta)$. Show R and Θ are independent and find their density functions.

Case 4:

(i) $X = (X_1, X_2, \cdots, X_k)$ has a continuous cumulative distribution function $F_X(x)$ and a probability density function $f_X(x)$. Suppose $g = (g_1, \ldots, g_p) : R^k \to R^p$. Consider $Y_i = g_i(X), i = 1, 2, \cdots, p, (l \leq k)$.

Assume that: (a) we can find $k - p$ other functions, $Y_j = g_j(x), j = p + 1, \cdots, k$, such that, the function $Y = g(X) = (g_1(X), \cdots, g_k(X))^T : R^k \to R^k$ is one to one on $S \to S'$, where S and S' are defined as before; (b) let the inverse function denoted by $h(y) = (h_1(y), \cdots, h_k(y))^T$ and let

$$J(y) = |\frac{\partial h}{\partial y}| = \begin{vmatrix} \frac{\partial h_1}{\partial y_1} & \cdots & \frac{\partial h_1}{\partial y_k} \\ \cdots & \cdots & \cdots \\ \frac{\partial h_k}{\partial y_1} & \cdots & \frac{\partial h_k}{\partial y_k} \end{vmatrix}.$$

Assume that the partial derivatives of h_i exists, continuous and $J(y)$ doesn't vanish except possibly on a set of Lebesgue measure 0. Then the joint probability density function of $Y = (Y_1, \cdots, Y_k)$ is

$$f_Y(y) = f_X(h_1(y), \cdots, h_k(y))|J(y)|$$

on S'. We can integrate out $y_{p+1}, \cdots, y_k$ in $f_Y(y)$ to get the probability density function of $Y_1, \cdots, Y_p$.

(ii) With the similar assumptions as before, instead of assuming $g = (g_1, \ldots, g_k)$ is one to one on $R^k \to R^k$, we suppose that S is a finite disjoint union of S_l, i.e. $S = \bigcup_{l=1}^{L} S_l$, such that on each S_l, $g : S_l \to S_l'$ is one to one. Here, $S' = \cup_{l=1}^{L} S_l'$ and S_l' need not be disjoint. Let h^l be the inverse of g restricted to S_l, i.e. $h^l : S_l' \to S_l$ and $h^l \circ g(x) = x$ for any $x \in S_l'$. Then the probability density function of $Y = g(X) = (g_1(X), \ldots, g_l(X))$ is

$$f_Y(y) = \sum_{l=1}^{L} I_{S_l'}(y) f_X(h_1^l(y), \cdots, h_k^l(y))|J^l(y)|.$$

Example 7.3 Let X_1, X_2 be two independent random variables following $N(0, 1)$. Find the probability density function of $Y_1 = \frac{X_1}{X_2}$.

As X_1, X_2 are independent, the density function for $X = (X_1, X_2)$ is $f_X(x_1, x_2) = \frac{1}{2\pi} e^{-\frac{x_1^2+x_2^2}{2}}$ on R^2. We take $S = R^2 - (\cdot, 0)$ (since $P_X(X_2 = 0) = 0$). Let

$$y_1 = g_1(x_1, x_2) = \frac{x_1}{x_2}, y_2 = g_2(x_1, x_2) = x_2.$$

Then $S' = g(S) = R^2$ and

$$x_1 = y_1 y_2,\ x_2 = y_2.$$

And

$$J(y) = \begin{vmatrix} y_2 & y_1 \\ 0 & 1 \end{vmatrix}.$$

So on S',

$$f_Y(y_1, y_2) = \frac{1}{2\pi} e^{-\frac{y_1^2 y_2^2 + y_2^2}{2}} |y_2|.$$

Hence,

$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-\frac{y_1^2 y_2^2 + y_2^2}{2}} |y_2| dy_2 = \frac{1}{\pi} \frac{1}{1 + y_1^2}$$

for $y_1 \in R$, which is the probability density function of Cauchy distribution.

Note as both $E(X^+)$ and $E(X^-)$ are $+\infty$, the expectation of the random variable following the Cauchy distribution doesn't exist.

Example 7.4 Let $X_1, X_2, \cdots, X_n$ be i.i.d random variables with an absolutely continuous cumulative distribution function $F(x)$ and a probability density function $f(x)$. Suppose that $Y_1 < Y_2 < \cdots < Y_n$ are the order statistics of the X's. That is, $Y_1 = min(X_1, \cdots, X_n)$, $Y_2 =$ second smallest of $X_1, \cdots, X_n$, and so on.

The space of X is $S = R^n$ and the space of Y is $S' = \{y : y_1 \leq y_2 \leq \cdots y_n\}$. We can exclude all equalities from S and S' since the probability of any two X being equal is 0. Then $S' = \{y : y_1 < y_2 < \cdots y_n\}$. $f_X(x) = \prod_{i=1}^n f(x_i)$, Let $S_1, S_2, \cdots, S_{n!}$ be all $n!$ possible permutation of the ordinals of S', $S = \sum_{i=1}^{n!} S_i$.

On S_i, the map is

$$y_1 = x_{i_1}, \cdots, y_n = x_{i_n},$$

where $(i_1, \cdots, i_n)$ is a fixed permutation of $(1, 2, \cdots, n)$ and $g^i : S_i \to S'$ is one to one.

$$\begin{pmatrix} y_1 = x_{i_1} \\ \vdots \\ y_k = x_{i_n} \end{pmatrix} \Longrightarrow \begin{pmatrix} x_1 = y_{j_1} \\ \vdots \\ x_k = y_{j_n} \end{pmatrix} \Longrightarrow |J^{(i)}(y)| = 1$$

Since the determinant of $n \times n$ matrix whose columns are some permutation of the of the columns of $I_{n\times n}$ is 1 or -1,

$$f_Y(y) = \sum_{i=1}^{n!} \prod_{i=1}^{n} f(y_i) = n! \prod_{i=1}^{n} f(y_i)$$

for $-\infty < y_1 < y_2 < \cdots < y_n < \infty$. The last equality follows as on each set S_i, $\prod_{j=1}^{n} f(x_{i_j}) = \prod_{i=1}^{n} f(y_i)$.

Next to find the probability density function of $Y_1, \cdots, Y_r (r < n)$, one may proceed by integrating out $y_{r+1}, \cdots, y_n$ in $f_Y(y)$ over the range $y_r < y_{r+1} < \cdots < y_n$. Note that

$$\begin{aligned} &\int_{y_{n-1}}^{+\infty} f(y_n) dy_n = 1 - F(y_{n-1}), \\ &\int_{y_{n-2}}^{+\infty} (1 - F(y_{n-1})) f(y_{n-1}) dy_{n-1} \\ &= \int_{F(y_{n-2})}^{1} (1 - u) u du = \frac{(1 - F(y_{n-2}))^2}{2}, \\ &\cdots \end{aligned}$$

and so on. Hence,

$$f(y_1, \cdots, y_r) = \frac{n!}{(n-r)!} \prod_{i=1}^{r} f(y_i)(1 - F(y_r))^{n-r}$$

for $-\infty < y_1 < \cdots < y_r < \infty$. □

Exercise 7.3 Suppose A, B, C are positive, independent random variables with distribution function F. Show the quadratic equation $Az^2 + Bz + C = 0$ has real solutions with probability $\int_0^\infty \int_0^\infty F(x^2/4y) dF(x) dF(y)$.

Other commonly used methods to determine the distribution of transformed random variables include:

(1) Handling the cumulative distribution function first then obtaining the probability density function.

Example 7.5 Find probability density function of $Y_1 = \min(X_1, \cdots, X_n)$. Note that

$$1 - F_{Y_1}(y) = Pr(Y_1 \geq y) = Pr(\text{all } X_i \geq y) = (1 - F(y))^n.$$

Therefore, $F_{Y_1}(y) = 1 - (1 - F(y))^n$. Hence,

$$f_{Y_1}(y) = n[1 - F(y)]^{n-1} f(y).$$

(2) Use of characteristic function or moment generating function.

(3) Use of Probability generating function.

(4) Identification of moments to those of a known distribution.

8 Conditional Distribution and Expectation

8.1 Conditional Distribution

Definition: Let $(\Omega, \mathcal{F}, P)$ be a probability space and let a pair of random variables, (X, Y), have a cumulative distribution function $F(x, y)$. Denote by P_X and P_Y the probability measures induced by X and Y respectively. Let $\mathfrak{B}$ be the Borel $\sigma-$field on R. We define the *conditional distribution* as

$$P(Y \in B_2 | X \in B_1) = \frac{P(X^{-1}(B_1) \cap Y^{-1}(B_2))}{P(X^{-1}(B_1))} \stackrel{\text{def}}{=} \frac{V(B_1, B_2)}{P_X(B_1)},$$

provided $P_X(B_1) > 0$.

For example, if $B_1 = \{x\}$ and $P_X(\{x\}) > 0$,

$$P(Y \in B_2 | X \in \{x\}) = \frac{V(\{x\}, B_2)}{P_X(\{x\})}.$$

The left side above also can be denoted as $P_{Y|X=x}(B_2)$.

This definition can also be extended to accommodate when $P_X(B_1) = 0$. Again let $B_1 = \{x\}$ and suppose that $P_X(\{x\}) = 0$. Introduce $B_{1\epsilon} = [x, x + \epsilon]$ so that $P_X(B_{1\epsilon}) > 0$ for any $\epsilon > 0$. So

$$P(Y \in B_2 | X \in B_{1\epsilon}) = \frac{V(B_{1\epsilon}, B_2)}{P_X(B_{1\epsilon})}$$

is well defined. Let $\epsilon \to 0^+$. Provided that the limit $v(x, B_2)$ exists and is unique, we call this limit $P_{Y|X=x}(B_2) = v(x, B_2)$.

Example 8.1 Let $F(x, y)$ be absolutely continuous with a probability density function $f(x, y)$. Assume that $f(x, y)$ is also continuous and that $f_X(x)$ is the marginal probability density function for X. Let $f_X(x) > 0$ on $[x_0, x_0 + h]$. It follows

$$\begin{aligned} P(Y < y | X \in [x_0, x_0 + \epsilon]) &= \frac{\frac{1}{\epsilon}\int_{x_0}^{x_0+\epsilon}\int_{-\infty}^{y} f(x,t)dtdx}{\frac{1}{\epsilon}\int_{x_0}^{x_0+\epsilon} f_X(x)dx} \\ &\longrightarrow \frac{\int_{-\infty}^{y} f(x_0,t)dt}{f_X(x_0)} \text{ as } \epsilon \to 0^+. \end{aligned}$$

Now formally, we can also write

$$F_{Y|X=x_0}(y) = \frac{\int_{-\infty}^{y} f(x_0,t)dt}{\int_{-\infty}^{+\infty} f(x_0,t)dt} = \frac{\int_{-\infty}^{y} f(x_0,t)dt}{f_X(x_0)}.$$

Here $F_{Y|X=x_0}(y)$ is a distribution function with the probability density function

$$f_{Y|X=x_0}(y) = \frac{f(x_0, y)}{f_X(x_0)}.$$

For two arbitrary sets B_1 and B_2 in $\mathfrak{B}$, recall that

$$V(B_1, B_2) = P(X^{-1}(B_1) \cap Y^{-1}(B_2)).$$

It is easy to show that

Theorem 8.1 *Keep B_2 fixed, allow B_1 to vary over $\mathfrak{B}$. Then $V(\cdot, B_2)$ is a finite measure on $(R, \mathfrak{B})$ and $V(\cdot, B_2) << P_X$.*

Proof: Notice that $V(R, B_2) = P(Y^{-1}(B_2)) \leq 1$ is finite and $V(\cdot, B_2)$ is non-negative.

In addition,

$$\begin{aligned} V(\sum_j B_{1j}, B_2) &= P(X^{-1}(\sum_j B_{1j}) \cap Y^{-1}(B_2)) = P(\sum_j X^{-1}(B_{1j}) \cap Y^{-1}(B_2)) \\ &= \sum_j P(X^{-1}(B_{1j}) \cap Y^{-1}(B_2)) = \sum_j V(B_{1j}, B_2). \end{aligned}$$

So, $V(\cdot, B_2)$ is a finite measure.

If $P_X(B_1) = 0$, so is $P(X^{-1}(B_1))$. Hence, $V(B_1, B_2) = 0$ as $V(B_1, B_2) \leq P(X^{-1}(B_1))$. □

Thus, by the R-N theorem, there exists a uniquely determined measurable function $g(x, B_2) \geq 0$ such that

$$V(B, B_2) = \int_B g(x, B_2) dP_X \tag{8}$$

for any $B \in \mathfrak{B}$.

We next study this measurable function, $g(x, B_2)$. One may show that

(1) $0 \leq g(x, B_2) \leq 1$, a.e (with respect to P_X).

Proof: if possible, suppose $g(x, B_2) > 1 + \epsilon$ on B_1, $P_X(B_1) > 0$, so $V(B_1, B_2) > P_X(B_1)$ which is impossible. □

(2) if $B = \{x\}$ and $P_X(\{x\}) > 0$, $V(\{x\}, B_2) = g(x, B_2) P_X(\{x\})$ so

$$g(x, B_2) = \frac{V(\{x\}, B_2)}{P_X(\{x\})}.$$

(3) Define $P_{Y|X=x}(B_2)$ to be $g(x, B_2)$. Then $P_{Y|X=x}(\cdot)$ has all the properties of a probability measure for almost all x, $a.e($ with respect to $P_X)$.

Proof: First, the nonnegativity of $P_{Y|X=x}(\cdot)$ is obvious.

Second, we show its regularity, i.e. $P_{Y|X=x}(R) = 1$ a.e. In fact, notice that

$V(B,R) = P_X(B) = \int_B dP_X$ for any $B \in \mathfrak{B}$. Using (8) where $B_2 = R$, we have

$$\int_B P_{Y|X=x}(R) dP_X = \int_B dP_X.$$

Hence, $P_{Y|X=x}(R) = 1, a.e$(with respect to P_X).

To prove countable additivity, we show $P_{Y|X=x}(B_2) = \sum_{j=1}^{\infty} P_{Y|X=x}(B_{2j})$ a.e (with respect to P_X) if $B_2 = \sum_{i=1}^{\infty} B_{2j}$. We consider

$$\begin{aligned} V(B,B_2) &= \sum_{j=1}^{\infty} V(B,B_{2j}) \\ &= \sum_{j=1}^{\infty} \int_B P_{Y|X=x}(B_{2j}) dP_X \\ &= \int_B [\sum_{j=1}^{\infty} P_{Y|X=x}(B_{2j})] dP_X, \end{aligned}$$

where the last equality is by the MCT.

Again compare (8) and obtain that

$$P_{Y|X=x}(B_2) = \sum_{j=1}^{\infty} P_{Y|X=x}(B_{2j}) \text{ a.e.}$$

□

One may notice that equation (8) is actually essential in the development of acquiring the conditional probability of $Y \in B_2$ given X, $P_{Y|X=x}(B_2)$. Now directly starting from equation (8), we give below a much more abstract definition for the conditional probability.

Definition: Given two random variables X, Y on the probability space $(\Omega, \mathcal{F}, P)$, suppose a bivariate function, $g(x, B_2)$, where $x \in R$ and B_2 a Borel set in $\mathfrak{B}$, satisfies (i) for any fixed B_2, $g(x, B_2)$ is a Borel measurable function with respect to x, i.e. $g(X, B_2)$ is a random variable. (ii) for any fixed x, $g(x, \cdot)$ is a probability measure on $(R, \mathfrak{B})$. If, for a fixed $B_2 \in \mathfrak{B}$,

$$P(X \in B, Y \in B_2) = \int_B g(x, B_2) dP_X$$

holds for any $B \in \mathfrak{B}$, then $g(x, B_2)$ is called the conditional probability of $Y \in B_2$ given $X = x$, and the random variable $g(X, B_2)$ is defined to be the conditional probability of $Y \in B_2$ given X, often denoted by $P_{Y|X}(B_2)$.

8.2 Conditional Expectation

In Section 3.4, we introduced the cumulative distribution function $F(x) = P_X(-\infty, x)$ to characterize a random variable X. Similarly, we may define the conditional distribution function of Y given $X = x$ to be

$$F_{Y|X=x}(y) = P_{Y|X=x}(-\infty, y),$$

based on which we may further consider its conditional expectation.

Definition of Conditional Expectation: The conditional expectation of Y given $X = x$ is defined to be $E(Y|X = x) = \int_R y dF_{Y|X=x}(y)$, or if $h(\cdot)$ is a measurable function on R, define $E[h(Y)|X = x] = \int_R h(y) dF_{Y|X=x}(y)$. Let $g(x) = E[h(Y)|X = x])$. Then define $E[h(Y)|X]) = g(X)$.

Notice that there is a subtle difference between $E[h(Y)|X = x]$ and $E[h(Y)|X]$. Simply speaking, the former is a real value of a measurable function evaluated at x, while the latter is a random variable. For instance, if $E[h(Y)|X = x] = x^2$, then $E[h(Y)|X] = X^2$.

Theorem 8.2

$$E(E(h(Y)|X)) = E(h(Y)).$$

Proof: First take $h(\cdot)$ to be an indicator function, i.e. $h(y) = I_B(y)$ where B is a Borel set. Then

$$\begin{aligned} E[h(Y)] &= P(Y \in B) = P(X \in R, Y \in B) \\ &= \int_R [P_{Y|X=x}(B)] dP_X = \int_R \int_R (I_B(y) dP_{Y|X=x}) dP_X \\ &= E_X[E(h(Y)|X)]. \end{aligned}$$

Next, use simple functions $\uparrow h(y)$(nonnegative) and use the fact that $h(y) = h^+(y) - h^-(y)$. □

If taking $h(\cdot)$ to be an identity function, we have

$$E(Y) = E(E(Y|X)) \tag{9}$$

Exercise 8.1 Follow the proof of Theorem 8.2 and show that, for any Borel set B on the real line,

$$E(I(X \in B)Y) = E(I(X \in B)E(Y|X)). \tag{10}$$

Using conditional expectation, we have an alternative way to define conditional probability when Y, X are random variables and $E(Y)$ is finite. We first consider nonnegative random variables and extend to general random variables easily.

Nonnegative Random Variable: We will use the R-N theorem to define the conditional expectation for a nonnegative random variable.

Let Y be a nonnegative random variable with a finite expectation. Define a set function v on $(R, \mathfrak{B})$ by $v(B) = E[Y I_B(X)]$. We show that $v(\cdot)$ is a finite measure dominated by P_X. Thus the conditions in the R-N theorem are satisfied and we may define the resulting Radon-Nikodym derivative to be the conditional expectation.

To proceed we first show

Theorem 8.3 *$v(\cdot)$ is a finite measure dominated by P_X on $(R, \mathfrak{B})$.*

Proof: Obviously $v(R) = E(Y)$ is finite and v is non-negative. In addition,

$$v(\sum_j B_j) = E(Y I_{\sum_j B_j}(X)) = E[\sum_{j=1}^{\infty} Y I_{B_j}(X)] = \sum_{j=1}^{\infty} v(B_j),$$

where the last equality is by the MCT.

As $P_X(B) = 0$ implies $P_{X,Y}(B \times R) = 0$, hence

$$v(B) = \int_{R \times R} y I_B(x) dP_{X,Y} = \int_{B \times R} y dP_{X,Y} = 0.$$

□

By the R-N theorem, there exists a non-negative, finite, measurable function $h(x) : R \to R^+$, determined uniquely a.e (with respect to P_X), such that

$$v(B) = \int_B h(x) dP_X, \text{ for any } B \in \mathfrak{B}.$$

We can thus define $E(Y|X=x)=h(x)$, a measurable function of x (instead of y!). Hence,

$$E_{X,Y}(YI_B(X))=\int_B E(Y|X=x)dP_X=E_X[I_B(X)E(Y|X)].$$

General Random Variable: If Y is any random variable with finite $E(Y)$, let $Y=Y^+-Y^-$. With $E(Y^+|X), E(Y^-|X)$ defined, we readily define $E(Y|X)=E(Y^+|X)-E(Y^-|X)$.

8.3 Abstract Conditional Expectation

We have defined the conditional expectation in terms of a conditional probability function, and this is adequate as long as we only deal with a fixed pair of random variables. In the later sections, we might be interested in stochastic processes or a whole family of random variables. It turns out a more flexible definition can be established independent of conditional distributions. That is, we may use identity (10) to define $E(Y|X=x)$ and $E(Y|X)$.

Definition: Let X,Y be a pair of random variables on $(\Omega,\mathcal{F},P)$. If $E(Y)$ is finite and there exists a Borel measurable function g such that that

$$E(YI_B(X))=\int_B g(x)dP_X(x)$$

holds for any $B\in\mathfrak{B}$. Furthermore, if the integrand function g is unique with respect to P_X. Then we define $E(Y|X=x)=g(x)$ and $E(Y|X)=g(X)$.

From this definition, it is noticeable that the conditional expectation $E(Y|X=x)$ of Y given $X=x$ is a finite measurable function of x, whose value at a point x is denoted by $E(Y|X=x)$ and $E(Y|X)$ is defined in such a way that (10) holds for any $B\in\mathfrak{B}$.

Exercise 8.2 If Y is a constant c a.e. then prove $E(Y|X=x)=c$ a.e. (with respect to P_X).

Exercise 8.3 Using the fact "if $\int_B f d\mu\le\int_B g d\mu$ for any $B\in\mathfrak{B}$, where μ is a $\sigma-$finite measure on R, then $f\le g$ a.e. (with respect to μ)," prove that $Y_1\le Y_2$ a.e. implies $E(Y_1|X=x)\le E(Y_2|X=x)$ a.e. (with respect to P_X).

The conditional expectation $E(Y|X)$ has the following properties.

Property 8.1 $E(Y|X=x)$ is a measurable function of x alone. Hence, $E(Y|X)$ is a random variable and $E_X(E(Y|X)) = E_Y(Y)$.

Property 8.2 Let $t(Y)$ be a measurable function of Y, with finite expectation. Then $E(t(Y)|X=x)$ is well defined. In particular, we may redefine the conditional probability of $Y \in B_2$, given X, to be $P_{Y|X}(B_2) = E[I_{B_2}(Y)|X]$ for any Borel set B_2.

Property 8.3 One may extend to the case with a $k-$ vector $X = (x_1, \cdots, x_k)$, $v(B) = E[Y I_B(X)], B \subset R^k$ so $E(Y|X=x) : R^k \to R$ is a measurable function of x.

Property 8.4 Conditional expectation has all the properties of expectation or integral a.e (with respect to P_X). In particular, (1) If $Y \geq 0$, the $E(Y|X) \geq 0$ a.e (with respect to P_X); (2) $E(Y_1|X) + E(Y_2|X) = E(Y_1 + Y_2|X)$ a.e (with respect to P_X).

Proof of (2): Consider for any $B \in \mathfrak{B}$,

$$\begin{aligned} & \int_B (E(Y_1|X=x) + E(Y_2|X=x)) dP_X \\ = & \int_B E(Y_1|X=x) dP_X + \int_B E(Y_2|X=x) dP_X \\ = & E(I_B(X)Y_1) + E(I_B(X)Y_2) \\ = & E(I_B(X)(Y_1+Y_2)). \end{aligned}$$

The first and the third equalities are by linearity of integrals, while the second is due to the definition of conditional expectation. Hence, again by definition, $E(Y_1|X) + E(Y_2|X) = E(Y_1 + Y_2|X)$ a.e (with respect to P_X). □

Property 8.5 If $E(Y)$ is finite and $E(Yg(X))$ is finite, then

$$E(Yg(X)|X=x) = g(x)E[Y|X=x] \text{ a.e (with respect to } P_X).$$

Proof: First, take $g(x) = I_C(x)$ where $C \in \mathfrak{B}$. Then for any $B \in \mathfrak{B}$, by definition,

$$E(Y I_C(X) I_B(X)) = \int_B E(Y I_C(x)|X=x) dP_X.$$

But, we also have

$$E(YI_C(X)I_B(X)) = E(YI_{BC}(X)) = \int_{BC} E(Y|X=x)dP_X = \int_B I_C(x)E(Y|X=x)dP_X.$$

So

$$E(YI_C(X)|X=x) = I_C(x)E(Y|X=x).$$

At the second step, we use a sequence of simple random variables to approximate a non-negative random variable and generalize to general cases. □

Property 8.6 That X and Y are independent implies
$E(Y|X) = E(Y)$ a.e (with respect to P_X).

Property 8.7 Conditional variance, with all the following moments assumed finite,

$$var(Y|X) = E(Y^2|X) - (E(Y|X))^2,$$

is a measurable function of the random variable of X.

It then follows that

$$\begin{aligned} E_X(var(Y|X)) &= E(Y^2) - E_X(E(Y|X)^2) \\ &= var(Y) + E(Y)^2 - E_X E(Y|X)^2 \\ &= var(Y) - (E_X(E(Y|X)^2) - (E_X E(Y|X))^2) \\ &= var(Y) - var(E(Y|X)). \end{aligned}$$

Hence,

$$var(Y) = E_X(var(Y|X)) + var(E(Y|X)).$$

So far we have defined the conditional expectation for a pair of random variables. A little more abstraction and generalization allow us to to define the conditional expectation of an arbitrary variable Y with respect to an arbitrary σ-field of sets.

Definition of Abstract Conditional Expectation: In an abstract probability space $(\Omega, \mathcal{F}, P)$, let $\mathcal{F}_0$ be an arbitrary σ-field of sets contained in $\mathcal{F}$. Let Y be a random variable with finite expectation. A random variable U is called an *abstract*

conditional expectation of Y relative to $\mathcal{F}_0$ if and only if U is $\mathcal{F}_0$-measurable (that is, $(U \leq a) \in \mathcal{F}_0$ for any real a) and

$$E(I_B Y) = E(I_B U) \text{ or } \int_B Y dP = \int_B U dP$$

holds for every $B \in \mathcal{F}_0$. In this case, we write $U = E(Y|\mathcal{F}_0)$.

Generally, the conditional expectation $E(Y|X = x)$ is more intuitive than the conditional expectation on a σ-field. However, the latter is often easier to handle in formal arguments and, hence, is almost universally preferred in the probability literature. In fact, denoting by $\sigma(X) = \{X^{-1}(B) : B \in \mathfrak{B}\}$, the σ-field generated by X, and letting $\mathcal{F}_0 = \sigma(X)$, the present definition concurs with (10). In other words, the conditional expectation of $E(Y|X)$ should be understood to be $E(Y|\sigma(X))$.

By definition, if $\mathcal{F}_0 = \mathcal{F}$, we may take $E(Y|\mathcal{F}_0) = Y$; if $\mathcal{F}_0$ is a trivial σ-field, containing only Ω and the empty set, then $E(Y|\mathcal{F}_0) = E(Y)$; if Y is itself $\mathcal{F}_0$ measurable, then $E(Y|\mathcal{F}_0) = Y$.

The existence of abstract conditional expectation can be proved by using the R-N theorem. Additionally, one may have a general formula

$$E(YZ|\mathcal{F}_0) = Z \cdot E(Y|\mathcal{F}_0)$$

for any $\mathcal{F}_0$-measurable function Z. The following, an extension of Theorem (8.2) lists another important property of the abstract conditional expectation.

Theorem 8.4 *If $\mathcal{F}_0$, $\mathcal{F}_1$ are two σ-fields and $\mathcal{F}_0 \subset \mathcal{F}_1$. Then*

$$E(Y|\mathcal{F}_0) = E(E(Y|\mathcal{F}_1)|\mathcal{F}_0)$$

and

$$E(Y|\mathcal{F}_0) = E(E(Y|\mathcal{F}_0)|\mathcal{F}_1).$$

Proof: Let $U_0 = E(Y|\mathcal{F}_0)$ and $U_1 = E(Y|\mathcal{F}_1)$. It follows that

$$E(Y \cdot I_B) = E(U_1 \cdot I_B) = E(U_0 \cdot I_B)$$

for any $B \in \mathcal{F}_0$. By definition, if $\mathcal{F}_0 \subset \mathcal{F}_1$, then $E(Y|\mathcal{F}_0) = E(E(Y|\mathcal{F}_1)|\mathcal{F}_0)$.

From U_0 is $\mathcal{F}_0$-measurable, hence, $\mathcal{F}_1$-measurable, follows
$E(Y|\mathcal{F}_0) = E(E(Y|\mathcal{F}_0)|\mathcal{F}_1)$. □

Exercise 8.4 If Y is a constant c a.e. then prove $E(Y|\mathcal{F}_0) = c$ a.e. (with respect to P).

Exercise 8.5 If $Y_1 \leq Y_2$ a.e, then $E(Y_1|\mathcal{F}_0) \leq E(Y_2|\mathcal{F}_0)$ a.e. (with respect to P).

Parallel to the unconditional expectations, Jensen's inequality holds for abstract conditional expectations.

Theorem 8.5 *(Jensen's inequality) Let X be a random variable on $(\Omega, \mathcal{F}, P)$ and $\mathcal{F}_0$ a sub-σ-algebra. If $g : R \to R$ is a Borel function which is convex on R and $E(X|\mathcal{F}_0)$ is finite a.e (with respect to P), then $E(g(X)|\mathcal{F}_0) \geq g(E(X|\mathcal{F}_0))$ a.s with respect to P.*

8.4 Martingales

For a sequence of random variables $X_1, X_2, \ldots$, we may think of X_n as the price for a stock at time t_n. Having observed the first n prices, the expected price for time t_{n+1} is $E(X_{n+1}|X_1, \ldots, X_n)$. If this is equal to X_n, the market is 'fair' as the expected gain at time t_{n+1} is $E(X_{n+1} - X_n|X_1, \ldots, X_n) = 0$. If $E(X_{n+1}|X_1, \ldots, X_n) \geq X_n$, the market is 'favorable'; otherwise, if $E(X_{n+1}|X_1, \ldots, X_n) \leq X_n$, the market is 'unfavorable'.

Study of this type of sequences motivates an important concept of martingale in modern probability literature.

Definition: Let $X_1, X_2, \ldots$ be a sequence of random variables on a probability space $(\Omega, \mathcal{F}, P)$ and $\mathcal{F}_1, \mathcal{F}_2, \ldots$ be a sequence of σ-fields in $\mathcal{F}$ such that $\mathcal{F}_i \subset \mathcal{F}_{i+1}$. The sequence $(X_i, \mathcal{F}_i), i = 1, 2, \ldots$ is called a *martingale* if and only if the following conditions hold:

(i) X_i is $\mathcal{F}_i$-measurable for $i = 1, 2, \ldots$.

(ii) $E(|X_i|) < \infty$.

(iii) $E(X_{i+1}|\mathcal{F}_i) = X_i$ a.e (with respect to P).

If condition (i) holds, we often say $\mathcal{F}_i$ form a filtration and if condition (ii) is true, we call X_i adapted to $\mathcal{F}_i$. More explicitly speaking, if X_i represents the total gain of

a gambler after the i-th play and $\mathcal{F}_i$ represents his information about the gam at that time, condition (iv) means that his expected fortune after the next play is the same as his present fortune. Thus a martingale represents a fair game. If " $=$ " in condition (iv) is changed to " $\leq$ " (or " $\geq$ ") we call $(X_i, \mathcal{F}_i)$ a supermartingale (or submartingale).

Exercise 8.6 If $(X_i, \mathcal{F}_i)$ is a martingale, show for any $k \leq n$, with probability 1,

$$E(X_n|\mathcal{F}_k) = X_k.$$

Exercise 8.7 If $(X_i, \mathcal{F}_i)$ is a martingale, show for any positive integer k and n,

$$\text{cov}(X_{n+k} - X_n, X_n) = 0.$$

By Jensen's inequality, convex functions of martingales are submartingales.

Theorem 8.6 *If $X_1, X_2, \ldots$ is a martingale on a probability space $(\Omega, \mathcal{F}, P)$ with respect to $\mathcal{F}_1 \subset \mathcal{F}_2, \ldots$, a sequence of σ-fields in $\mathcal{F}$, and if $g(\cdot)$ is convex and $g(X_n)$ are integrable, then $g(X_1), g(X_2), \ldots$ is a submartingale with respect to $\mathcal{F}_1, \mathcal{F}_2, \ldots$.*

Proof: As $X_n = E(X_{n+1}|\mathcal{F}_n)$ and so $g(X_n) = g(E(X_{n+1}|\mathcal{F}_n))$. If g is convex, by Jensen's inequality, it follows that

$$g(X_n) = g(E(X_{n+1}|\mathcal{F}_n)) \leq E(g(X_{n+1})|\mathcal{F}_n) \text{a.s.}$$

A more interesting generalization of the discrete martingale is in the framework of continuous processes, e.g. survival or death processes. Before proceeding further, we first introduce the concept of a stochastic process.

Definition of Stochastic Process: On a common probability space $(\Omega, \mathcal{F}, P)$, a *stochastic process* is a family of random variables $X = \{X(t) : t \in \Gamma\}$ indexed by a set Γ.

In the definition, Γ usually indexes time, and is often either $\{0, 1, 2 \ldots, \}$ (discrete process) or $[0, \infty)$ (continuous process). The random functions $X(\cdot, \omega) : R^+ \to R$ for each $\omega \in \Omega$ are called the *sample paths* or *trajectories* of X. A process is called $right-$ or $left-$ continuous or said to have limits from left or right if the sample paths have such a property almost surely with respect to P.

For a continuous process, i.e $\Gamma = [0, \infty)$, the following is a rigorous formulation of information accruing over time.

Definition of Filtration: On $(\Omega, \mathcal{F}, P)$, a family of sub-σ-fields $\{\mathcal{F}_t, t > 0\}$ is called a $filtration$ if $s \leq t$ implies $\mathcal{F}_s \subset \mathcal{F}_t$.

For example, we may define $\mathcal{F}_t = \sigma\{X(s) : 0 \leq s \leq t\}$, the smallest σ-algebra with respect to which each of the random variables $X(s), 0 \leq s \leq t$ is measurable. In plain words, $\mathcal{F}_t$ contains the information generated by the process X on $[0, t]$.

Definition of Predictable Process: A process is called *predictable* with respect to a σ-field if it is measurable with respect to that σ-field.

We are now able to define the continuous version of a martingale, which has fundamental implications in many diverse areas such as stochastic differential equations, queuing theory and survival analysis.

Definition: On a common probability space $(\Omega, \mathcal{F}, P)$, let $M = \{M(t) : t \geq 0\}$ be a right continuous stochastic process with left-hand limits and $\{\mathcal{F}_t : t \geq 0\}$ a filtration. M is called a *martingale* with respect to $\{\mathcal{F}_t : t \geq 0\}$ if

(i) For each t, $M(t)$ is measurable with respect to $\mathcal{F}_t$-measurable.

(ii) $E(|M(t)|) < \infty$.

(iii) $E(M_{t+s}|\mathcal{F}_t) = M(s)$ a.s (with respect to P) for all $s, t \geq 0$.

M is called a *submartingale* if (iii) is replace by $E(M_{t+s}|\mathcal{F}_t) \geq M(s)$ a.s (with respect to P). and a *supermartingale* if (iii) is replace by $E(M_{t+s}|\mathcal{F}_t) \leq M(s)$ a.s (with respect to P).

We conclude this section with a theorem which gives the theoretical foundation to the survival analysis.

Doob-Meyer Decomposition Let N be a right-continuous nonnegative submartingale with respect to a stochastic basis $(\Omega, \mathcal{F}, \{\mathcal{F}_t : t \geq 0\}, P)$. Then there exists, uniquely (with respect to P) a right-continuous martingale M and an increasing right-continuous predictable process A such that $E(A(t)) < \infty)$ and

$$N(t) = M(t) + A(t)$$

a.s for any $t \geq 0$.

9 Product Measure, Iterated Integral and Convolution

Let $(\Omega_1, \mathcal{F}_1, \mu_1), (\Omega_2, \mathcal{F}_2, \mu_2)$ be two measurable sample spaces. The Cartesian product of these two spaces, $\Omega_1 \times \Omega_2$, is the set of all ordered pairs (ω_1, ω_2), where $\omega_i \in \Omega_i, i = 1, 2$. Among all the sets in this product space, we consider rectangles $A_1 \times A_2$, where $A_i \in \Omega_i$. With sets of this form, we wish to construct on $\Omega_1 \times \Omega_2$ a product measure μ such that $\mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$. In the case where μ_1 and μ_2 are Lebesgue measure on the real line, μ will be Lebesgue measure in the plane. The main result is Fubini's theorem, by which double integrals can be calculated as iterated integrals.

9.1 Product Measure

Definition: Let $(\Omega_1, \mathcal{F}_1, \mu_1), (\Omega_2, \mathcal{F}_2, \mu_2)$ be two sample spaces. The *product space* $\Omega_1 \times \Omega_2$ is defined to be

$$\Omega = \{(w_1, w_2) : w_1 \in \Omega_1, w_2 \in \Omega_2\}.$$

A set of the form $A_1 \times A_2 = \{(w_1, w_2) : w_1 \in A_1, w_2 \in A_2\}$, is called a *rectangle* (*measurable rectangle*: $\mathcal{F} = \{A_1 \times A_2 : A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2\}$).

Remark 9.1 $\mathcal{F}$ is neither a field nor a $\sigma-$ field. The $\sigma-$field generated by $\mathcal{F}$, $\sigma(\mathcal{F})$, is called the product $\sigma-$field of $\mathcal{F}_1$ and $\mathcal{F}_2$ and is denoted by $\mathcal{F}_1 \times \mathcal{F}_2$.

In this way, we can define the *Borel* σ-field on a two-dimensional plane. Take $\Omega_1 = \Omega_2 = R$, $\mathcal{F}_1 = \mathcal{F}_1 = \mathfrak{B}$. Then the Euclidean Borel σ-field is defined to be $\mathfrak{B}^2 = \mathfrak{B} \times \mathfrak{B}$, the smallest $\sigma-$ field generated by the Borel rectangle of the form $B_1 \times B_2 = \{(x, y) : x \in B_1 \in \mathfrak{B}, y \in B_2 \in \mathfrak{B}\}$. It, in fact, is also generated by rectangles of the form $\{(x, y) : a < x \leq b, c < y \leq d\}$; see Chung (2001, p.38). We are then able to define a Borel function on a two-dimensional plan.

Definition: If $f : R^2 \to R$ such that $f^{-1}(B) \subset \mathfrak{B}^2$ for any $B \in \mathfrak{B}$, we call f a *Borel function* on $R^2 \to R$.

Returning to introduce a measure on the product measurable space $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2)$, we define, on the class $\mathcal{F}$, $\mu(A_1 \times A_2) = \mu_1(A_1)\mu_2(A_2)$. If μ_1 and

μ_2 are $\sigma-$finite, by the extension theorem, this can be extended uniquely to a measure on $\mathcal{F}_1 \times \mathcal{F}_2$ and this extended measure is denoted by $\mu_1 \times \mu_2$.

Example 9.1 Suppose $\Omega_1 = R_1, \mathcal{F}_1 = \mathfrak{B}_1, \mu_1 = \mu_L$ and $\Omega_2 = R_2, \mathcal{F}_2 = \mathfrak{B}_2, \mu_2 = \mu_L$ as in the real space. Consider two components of an experiment:
$(\Omega_1, \mathcal{F}_1, P_1) \stackrel{X_1}{\to} (R, \mathfrak{B})$ and $(\Omega_2, \mathcal{F}_2, P_2) \stackrel{X_2}{\to} (R, \mathfrak{B})$. Now given $Y = f(X_1, X_2)$, what is the domain of Y? In fact, it is the product space
$(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2, \mu_1 \times \mu_2 = \text{product measure})$, as (X_1, X_2) is a pair of independent random variables $(\Omega_1, \Omega_2) \to R^2$.

9.2 Iterated Integral Theorem (Fubini's Theorem)

Let $Y = f(X_1, X_2)$, where $f(\cdot, \cdot)$ is a bivariate Borel function and X_1, X_2 are two random variables. We first prove that Y is indeed a random variable.

Theorem 9.1 *Suppose X_i is a random variable with respect to $(\Omega_i, \mathcal{F}_i), i = 1, 2$ and $f(\cdot, \cdot)$ is a bivariate Borel function. Then $Y = f(X_1, X_2)$ is a random variable with respect to the product space $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2)$.*

Proof: For any 2-dimensional set $A \subset R^2$, we write
$(X_1, X_2)^{-1}A = \{(w_1, w_2) : (X_1(w_1), X_2(w_2)) \in A)\}$. Then

$$f(X_1, X_2)^{-1}\mathfrak{B} = (X_1, X_2)^{-1} f^{-1}(\mathfrak{B}) \subset (X_1, X_2)^{-1}\mathfrak{B}^2 \subset \mathcal{F}_1 \times \mathcal{F}_2.$$

Only the last inclusion needs a proof. If $A = B_1 \times B_2$, where $B_1 \in \mathfrak{B}$ and $B_2 \in \mathfrak{B}$, then

$$(X_1, X_2)^{-1}(A) = X_1^{-1}(B_1) \times X_2^{-1}(B_2) \in \mathcal{F}_1 \times \mathcal{F}_2.$$

Now the collection of sets A such that $(X_1, X_2)^{-1}(A) \in \mathcal{F}_1 \times \mathcal{F}_2$ form a σ-field (similar to the proof in Theorem (3.4). As this σ-field contains all the *Borel rectangles*, hence, it must contain $\mathfrak{B}^2$. Therefore, each set in $\mathfrak{B}^2$ belongs to this collection, which completes the proof. □

Exercise 9.1 (Two-dimensional Jensen's Inequality) Let $f(x, y)$ be a real convex Borel function on the plane and X, Y are two random variables with finite expectations. Prove

$$f(E(X), E(Y)) \leq E(f(X, Y)).$$

Now we consider $f_{w_1}(w_2) : \Omega_2 \to R$ defined by $f_{w_1}(w_2) = f(w_1, w_2)$. This is called w_1-section of the function f and it can be shown that it is measurable respect to $\mathcal{F}_2$. We have the following integrals

$$\int_{\Omega_1 \times \Omega_2} f(w_1, w_2) d(\mu_1 \times \mu_2), \tag{11}$$

$$\int_{\Omega_1} (\int_{\Omega_2} f_{w_1}(w_2) d\mu_2) d\mu_1 \tag{12}$$

and

$$\int_{\Omega_2} (\int_{\Omega_1} f_{w_2}(w_1) d\mu_1) d\mu_2. \tag{13}$$

Integral (11) is called the *double integral*, while integrals (12) and (13) iterated integrals. Under some general conditions, these integrals exist and are equal, as indicated by the following Fubini's theorem for product measures.

Theorem 9.2 *(Fubini's Theorem or Iterated integrals theorem) Let* $(\Omega_1, \mathcal{F}_1, \mu_1)$ *and* $(\Omega_2, \mathcal{F}_2, \mu_2)$ *be two measure spaces where both* μ_1 *and* μ_2 *are* $\sigma-$*finite. If* $f(w_1, w_2)$ *is an* $(\mathcal{F}_1 \times \mathcal{F}_2)$ *measurable function* $(\Omega_1, \Omega_2) \to R$ *and one of the following conditions holds (1)* $f \geq 0$ *or (2)* $\int_{\Omega_1 \times \Omega_2} f d(\mu_1 \times \mu_2)$ *is finite, then integrals (11), (12) and (13) are equal.*

Proof: see Billingsley (1995, p.234). □

A simple application of Fubini's Theorem can give the following well-known result.

Theorem 9.3 *Let* X, Y *be two independent random variables. Then*

$$E(XY) = E(X)E(Y).$$

Proof: Here, $\Omega_1 = R_1 \to P_X$, $\Omega_2 = R_2 \to P_Y$ and $P_X \times P_Y = P_{X,Y}$. By independence $F_X F_Y = F_{X,Y}$. Hence,

$$E(XY) = \int_{R_1 \times R_2} xy d(P_X \times P_Y) = \int_{R_1} x \int_{R_2} y dP_Y dP_X = E(Y)E(X).$$

□

Products of more than two probability spaces can be treated similarly; for detailed discussion, see Billingsley (1995, p.238).

9.3 Convolution

Convolution, in addition to moment generating functions or characteristic functions (which will be covered later). is an important technique in quantifying the finite summation of independent random variables. To see this, we consider a pair of independent random variables X and Y with distribution functions F and G respectively. Consider the probability of $X + Y$ falling into B, a Borel set on the real line, i.e. $B \in \mathfrak{B}$. As we will show later,

$$P(X + Y \in B) = \int_{-\infty}^{\infty} G(B - x)dF(x).$$

This motivates us to define *convolution* as follows.

Definition: Let F and G be two univariate cumulative distribution functions. Define a function on $R \to [0, 1]$ by

$$F * G(u) = \int_{-\infty}^{\infty} G(u - x)dF(x), \text{ for any } u \in R.$$

The function $F * G$ is called the *convolution* of F with G.

Remark 9.2 $F * G$ itself is a proper cumulative distribution function .

Proof:

(Non-decreasing) If $u < u'$, then $G(u - x) \leq G(u' - x)$ for any x. Hence, $\int G(u - x)dF \leq \int G(u' - x)dF$.

(Left continuous) Let $u_n \uparrow u$ we want to show $F * G(u_n) \uparrow F * G(u)$.

Let $u_n \uparrow u$, then by left continuity of G, $|G(u_n - x)| \leq 1$ and $G(u_n - x) \uparrow G(u_x)$. Then by DCT,

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} G(u_n - x)dF(x) = \int_{-\infty}^{\infty} G(u - x)dF(x)$$

Lastly, $F * G(\infty) = 1, F * G(-\infty) = 0$. □

Remark 9.3 If X and Y are two independent random variables with cumulative distribution function F and G, the $F * G$ is the cumulative distribution function of $X + Y$.

Proof:

$$P(X+Y<u) = \int_{R_1 \times R_2} I_{[x+y<u]}(x,y) d(F \times G)$$
$$\int_{R_1} (\int I_{[y<u-x]}(y) dG) dF = \int_{-\infty}^{\infty} G(u-x) dF$$
$$= F * G(u).$$

□

But the converse is not necessarily true. That is, the fact that $F * G$ may be the cumulative distribution function of $X + Y$ for random variables X, Y with cumulative distribution function F and G doesn't imply X and Y are independent.

Remark 9.4 (commutative) As $F * G(u)$ or $G * F(u)$ is the cumulative distribution function for $X + Y$ if X and Y are two random variables with cumulative distribution function F and G, therefore, $F * G(u) = G * F(u)$.

Remark 9.5 (associative)

$$(F_1 * F_2) * F_3 = F_1 * (F_2 * F_3).$$

If $X_1, \cdots, X_n$ are i.i.d with common cumulative distribution function F and F^{n*} is F convoluted with itself n times, i.e. $F^{n*} = F^{(n-1)*} * F$, then F^{n*} is the cumulative distribution function of $\sum_{i=1}^{n} X_i$.

In addition, the convolution $*$ also has two important properties.

Property 9.1 If either F or G is continuous then $F * G$ is continuous.

Property 9.2 If either F or G is absolutely continuous (with respect to μ_L) then $F * G$ is also absolutely continuous (with respect to μ_L).

Proof: Let $G << \mu_L$ and has $g(y) = \frac{dG}{d\mu_L}$. Consider $v(u) = \int_{-\infty}^{\infty} g(u-x) dF(x)$ and let $F * G$ be denoted by V, we will show that $V << \mu_L$ and $v(u)$ is its probability

density function.

$$\int_a^b v(u)d\mu_L = \int_a^b (\int_{-\infty}^{\infty} g(u-x)dF(x))d\mu_L$$
$$= \int_{-\infty}^{\infty} (\int_a^b g(u-x)du)dF(x)$$
$$= \int_{-\infty}^{\infty} (\int_{a-x}^{b-x} g(y)dy)dF(x)$$
$$= \int_{-\infty}^{\infty} (G(b-x) - G(a-x))dF(x) = V(b) - V(a).$$

If this is true for arbitrary intervals, then $V << \mu_L$ and $\frac{dV}{d\mu_L} = v(u)$. Additionally, if $F << \mu_L$ and $\frac{dF}{d\mu_L} = f(x)$, then $v(u) = \int_{-\infty}^{\infty} g(u-x)f(x)dx$ is called the convolution of two densities of f and g. □

Example 9.2 Let $X \sim b(n,p)$ and $Y \sim U(0,1)$ be two independent random variables. Find the distribution function of $U = X + Y$.

We know that the cumulative distribution function for $U(0,1)$ is absolutely continuous (with respect to μ_L) with density $g(u) = I_{[0,1]}(u)$. Then form the convolution of densities of $b(n,p)$ and $U(0,1)$:

$$v(u) = \int_{-\infty}^{\infty} g(u-x)dF(x)$$
$$= \int_0^n I_{[0,1]}(u-x)dF(x)$$
$$= \binom{n}{[u]} p^{[u]} q^{n-[u]},$$

where $[u]$ is the greatest integer $\leq u$.

Therefore, the probability density function of $X + Y$ is
$v(u) = \binom{n}{[u]} p^{[u]} q^{n-[u]}, 0 \leq u < n+1.$ □

Example 9.3 Suppose that $X_1, \cdots, X_n$ are i.i.d random variables with the common probability density function $f(x) = e^{-x} 0 < x < \infty$. Find the cumulative distribution function of $U_n = \sum_{i=1}^n X_i$.

We proceed by using induction. First consider when $n = 2$,

$$\begin{aligned} v_2(u) &= \int_{-\infty}^{\infty} f(u-x)f(x)dx \\ &= \int_0^u f(u-x)f(x)dx = e^{-u}u \end{aligned}$$

Now assume for $n = k-1$,

$$v_{k-1}(u) = \frac{e^{-u}u^{k-2}}{(k-2)!}.$$

Then for $n = k$,

$$\begin{aligned} v_k(u) &= \int_{-\infty}^{\infty} q_{k-1}(u-x)f(x)dx \\ &= \int_0^u \frac{e^{-u}(u-x)^{k-2}}{(k-2)!}dx = \frac{e^{-u}}{(k-2)!}\int_0^u (u-x)^{k-2}dx = \frac{e^{-u}u^{k-1}}{(k-1)!} \end{aligned}$$

for any $u > 0$. □

Exercise 9.2 Show that the family of normal distributions is closed with respect to convolution in the sense that the convolution of any two in the family with arbitrary parameters is another in the family with some parameters.

10 Characteristic Function

10.1 Complex Numbers

Complex numbers are abstract quantities that turn out to have many useful applications. (Actually, the same can be said about real numbers.) A complex number x is usually written in the form $x = a + bi$, where a and b are real numbers called the *real* and *imaginary* parts, and i is an abstract quantity defined by the property that $i^2 = -1$. The '+' in $a + bi$ should be interpreted similar to the '+' in algebraic expressions such as $x + \exp(y)$. It does not have an interpretation in terms of addition of (real) numbers, though, and complex numbers could just as easily be thought of as being values (a, b) lying in a 2-dimensional vector space. The plane formed by plotting a on the horizontal (real) axis and b on the vertical (imaginary) axis is called the complex plane. It will be convenient to define 2 functions,

$\Re(a+bi) = a$ and $\Im(a+bi) = b$, which give the real and imaginary parts of a complex number. Also, let $\mathcal{C}$ be the set of complex numbers.

Two complex numbers $a+bi$ and $c+di$ are equal if $a=c$ and $b=d$. The basic arithmetic operations are

$$(a+bi)+(c+di) = (a+c)+(b+d)i,$$

$$(a+bi)-(c+di) = (a-c)+(b-d)i,$$

$$(a+bi)(c+di) = (ac-bd)+(ad+bc)i,$$

and

$$(a+bi)/(c+di) = \frac{ac+bd}{c^2+d^2} + \frac{bc-ad}{c^2+d^2}i, \quad (c^2+d^2 \neq 0).$$

Note that the definition of multiplication can be obtained by multiplying term by term, and using the definition $i^2 = -1$. Also, the definition of division is such that $(c+di)\{(a+bi)/(c+di)\} = a+bi$, if $c^2+d^2 > 0$. Using the definition of multiplication with $d=0$,

$$c(a+bi) = (ac)+(bc)i.$$

From these definitions, the use of '+' in $a+bi$ is justified, since it satisfies the usual properties of '+' in manipulating algebraic formulas.

Exercise 10.1 Find an expression of the form $a+bi$ for $1/i$.

The *complex conjugate* of $a+bi$ is $\overline{a+bi} = a-bi$. The *norm* (or absolute value or modulus) of a complex number is

$$|a+bi| = \{(a+bi)(\overline{a+bi})\}^{1/2} = (a^2+b^2)^{1/2}.$$

In general, a norm on a vector space V is a function $\rho : V \to R^1$ such that (a) $\rho(v) \geq 0$ for all $v \in V$, with $\rho(v) = 0$ if and only if $v = 0$, (b) $\rho(\alpha v) = |\alpha|\rho(v)$ for all $\alpha \in R^1$ and $v \in V$, and (c) $\rho(v+w) \leq \rho(v) + \rho(\omega)$. Here properties (a) and (b) are immediate from the definition, while property (c), called the triangle inequality, will be established below. (In general it is assumed here that vector spaces are defined over the field of real numbers.) It is easily verified directly from the definition that

$$\max(|a|,|b|) \leq |a+bi| \leq |a|+|b|. \tag{14}$$

Exercise 10.2 Show that the product of complex conjugates is the complex conjugate of the product. That is, if $x = a + bi$ and $y = c + di$, then $\overline{xy} = (\overline{x})(\overline{y})$.

If the point $a + bi$ is plotted in complex the plane, then it can also be represented in polar coordinates, using the standard transformation $a = r\cos(\theta)$ and $b = r\sin(\theta)$. Then $r = |a + bi|$, θ (the counter-clockwise angle in radians from the positive horizontal axis) is given by $\theta = \pm \arccos(a/r)$, with the sign of θ chosen so that both $\cos(\theta)$ and $\sin(\theta)$ have the correct sign, and $a + bi = r\{\cos(\theta) + i\sin(\theta)\}$ (θ can equivalently be defined in terms of extended versions of the arcsin or arctan functions). The angle θ is called the *argument* of $a + bi$, which is often written $\arg(a + bi)$. The angle is not unique, since adding any integer multiple of 2π gives the same point in the complex plane. The value of θ lying in $(-\pi, \pi]$ is called the *principal value*. (This definition for the interval of the principal value is common, but may not be universal.)

From the polar coordinate representation and the properties of the sine and cosine functions, if $a + bi = r_1\{\cos(\theta_1) + i\sin(\theta_1)\}$ and $c + di = r_2\{\cos(\theta_2) + i\sin(\theta_2)\}$, then

$$\begin{aligned}(a + bi)(c + di) &= r_1 r_2[\cos(\theta_1)\cos(\theta_2) - \sin(\theta_1)\sin(\theta_2) + i\{\cos(\theta_1)\sin(\theta_2) + \cos(\theta_2)\sin(\theta_1)\}] \\ &= r_1 r_2\{\cos(\theta_1 + \theta_2) + i\sin(\theta_1 + \theta_2)\}.\end{aligned}$$

This is a polar coordinate representation with radius $r_1 r_2$ and angle $\theta_1 + \theta_2$. Thus

$$|(a + bi)(c + di)| = r_1 r_2 = |a + bi||c + di|,$$

so the norm of a product is the product of the norms.

Exercise 10.3 If $a + bi = r\{\cos(\theta) + i\sin(\theta)\}$, show $(a + bi)^k = r^k\{\cos(k\theta) + i\sin(k\theta)\}$.

An inner product (or scalar product) on a vector space V is a function $\langle \cdot, \cdot \rangle : V \times V \to R^1$, such that (a) $\langle v, v \rangle \geq 0$, with equality if and only if $v = 0$, (b) $\langle v, w \rangle = \langle w, v \rangle$, (c) $\langle \alpha v, w \rangle = \alpha \langle v, w \rangle$ for all $\alpha \in R^1$, and (d) $\langle u + v, w \rangle = \langle u, w \rangle + \langle v, w \rangle$. An inner product for complex numbers can be defined by

$$\langle a + bi, c + di \rangle = \Re\{(a + bi)(\overline{c + di})\} = ac + bd. \tag{15}$$

(If (a, b) and (c, d) are points in R^2, then the standard Euclidean inner product gives the same formula.) This inner product is consistent with the definition of the

norm above, since

$$\langle a+bi, a+bi\rangle = a^2+b^2 = |a+bi|^2.$$

To verify that (15) is an inner product, from the previous expression

$$\langle a+bi, a+bi\rangle \geq 0,$$

with equality only if $a=b=0$. Also

$$\langle a+bi, c+di\rangle = ac+bd = ca+db = \langle c+di, a+bi\rangle,$$

$$\langle \alpha(a+bi), c+di\rangle = (\alpha a)c+(\alpha b)d = \alpha(ac+bd) = \alpha\langle a+bi, c+di\rangle$$

for any real α, and

$$\langle (a+bi)+(e+fi), c+di\rangle = (a+e)c+(b+f)d = (ac+bd)+(ec+fd) = \langle a+bi, c+di\rangle+\langle e+fi, c+di\rangle.$$

It can be shown that any inner product satisfying properties (a) through (d) above will satisfy the Cauchy-Schwarz inequality, so

$$|\langle a+bi, c+di\rangle| \leq |a+bi||c+di|.$$

This is also easily verified here directly from the definition of the inner product.

The triangle inequality,

$$|a+bi+c+di| \leq |a+bi|+|c+di|,$$

can be established using the Cauchy-Schwarz inequality. Setting $x=a+bi$ and $y=c+di$, it follows that

$$|x+y|^2 = \langle x+y, x+y\rangle = \langle x, x\rangle + 2\langle x, y\rangle + \langle y, y\rangle \leq |x|^2+2|x||y|+|y|^2 = (|x|+|y|)^2.$$

A sequence of complex numbers $x_n = a_n + b_n i$ has a limit $x_0 = a_0 + b_0 i$, if for any $\epsilon > 0$, there is a value $N(\epsilon)$ such that $|x_n - x_0| < \epsilon$ for all $n > N(\epsilon)$. Equivalently, $x_n \to x_0$ if $|x_n - x_0| \to 0$.

Theorem 10.1 $\lim_{n\to\infty} x_n = x_0$ *if and only if* $\lim_{n\to\infty} a_n = a_0$ *and* $\lim_{n\to\infty} b_n = b_0$.

Proof: Suppose $a_n \to a_0$ and $b_n \to b_0$. From (14),
$|(a_n + b_n i) - (a_0 + b_0 i)| \leq |a_n - a_0| + |b_n - b_0| \to 0$, so $(a_n + b_n i) \to (a_0 + b_0 i)$.
Conversely, from (14), $|a_n - a_0| \leq |x_n - x_0|$ and $|b_n - b_0| \leq |x_n - x_0|$, so if
$|x_n - x_0| \to 0$, so must $|a_n - a_0|$ and $|b_n - b_0|$. □

A set $A \subset \mathcal{C}$ of complex numbers is open if for every $x \in A$ there is an $\epsilon_x > 0$ such that $\{y : |y - x| < \epsilon_x\} \subset A$. A set B is closed if B^c is open.

Exercise 10.4 Prove that if B is closed, then it contains all its limit points. That is, show that if $x_1, x_2, \ldots$ is a sequence of points with $x_n \in B$ for all n, and if $x_n \to x_0$, then $x_0 \in B$.

A complex infinite series $\sum_{n=1}^{\infty}(a_n + b_n i)$ is *absolutely convergent* if $\sum_n |a_n + b_n i| < \infty$. From (14), $\sum_n |a_n + b_n i| < \infty$ if and only if both $\sum_n |a_n| < \infty$ and $\sum_n |b_n| < \infty$. If $\sum_{n=1}^{\infty}(a_n + b_n i)$ is absolutely convergent, then its value is defined to be $\lim_{m\to\infty} \sum_{n=1}^{m}(a_n + b_n i)$. Clearly,

$$\lim_{m\to\infty} \sum_{n=1}^{m}(a_n + b_n i) = \sum_{n=1}^{\infty} a_n + i \sum_{n=1}^{\infty} b_n. \tag{16}$$

It can be shown in general that the value of an absolutely convergent series is not affected by arbitrary rearrangements of the terms in the series.

10.2 Complex-valued Functions

A complex-valued function f is a rule assigning a unique complex number to each point in the function's domain. Regardless of the domain of f, since every complex number can be written in the form $a + bi$, it follows that f can be represented

$$f(\cdot) = g(\cdot) + h(\cdot)i, \tag{17}$$

for real valued functions $g(\cdot) = \Re\{f(\cdot)\}$ and $h(\cdot) = \Im\{f(\cdot)\}$. In particular, if the domain of f is a subset of the complex numbers, then

$$f(a + bi) = g(a, b) + h(a, b)i, \tag{18}$$

for real valued functions g and h. Important properties of f, such as continuity and differentiability, could be investigated through these real component functions.

Generalizing familiar real functions to complex variables is nontrivial, and generally not unique. Two examples follow.

Example 10.1 The exponential function. How should $\exp(x)$ be defined when $x = a + bi$ is complex? Two minimal criteria are first, the definition should reduce to the standard exponential function for reals when $b = 0$, and second, familiar properties of the exponential function, such as $\exp(x + y) = \exp(x)\exp(y)$, should hold for complex arguments. If in fact we require $\exp(a + bi) = \exp(a)\exp(bi)$, where $\exp(a)$ has the usual definition for real a, then the remaining problem is to define $\exp(bi)$.

Another property that $\exp(a)$ has for real a is the Taylor series representation

$$\exp(a) = \sum_{k=0}^{\infty} a^k/k!.$$

Formally substituting bi for a, the resulting complex series is absolutely convergent for all b. Thus this gives a well defined function, whose value is defined through (16). Since the series is absolutely convergent, the terms may be rearranged. Thus

$$\begin{aligned}
\sum_{k=0}^{\infty}(bi)^k/k! &= \sum_{k=0}^{\infty} i^{2k}b^{2k}/(2k)! + \sum_{k=0}^{\infty} i^{2k+1}b^{2k+1}/(2k+1)! \\
&= \sum_{k=0}^{\infty}(-1)^{k+2}b^{2k}/(2k)! + i\sum_{k=0}^{\infty}(-1)^{k+2}b^{2k+1}/(2k+1)! \\
&= \cos(b) + i\sin(b),
\end{aligned}$$

from the Taylor series representations of the sine and cosine functions. This suggests defining $\exp(ib) = \cos(b) + i\sin(b)$, and

$$\exp(a + ib) = \exp(a)\{\cos(b) + i\sin(b)\}.$$

This is the standard definition. As the Taylor series representation suggests, with this definition many of the properties of the real $\exp(\cdot)$ function carry over to the

complex setting. For example,

$$\begin{aligned}
&\exp\{(a+bi)+(c+di)\} \\
&\quad = \exp\{(a+c)+(b+d)i\} \\
&\quad = \exp(a)\exp(c)\{\cos(b+d)+i\sin(b+d)\} \\
&\quad = \exp(a)\exp(c)[\cos(b)\cos(d)-\sin(b)\sin(d)+i\{\sin(b)\cos(d)+\cos(b)\sin(d)\}] \\
&\quad = \exp(a)\exp(c)\{\cos(b)+i\sin(b)\}\{\cos(d)+i\sin(d)\} \\
&\quad = \exp(a+bi)\exp(c+id).
\end{aligned}$$

□

Example 10.2 Natural logarithms. The $\log(\cdot)$ function is defined as the inverse of $\exp(\cdot)$. That is, $\log(a+bi)$ is the value $c+di$ such that $\exp(c+di) = a+bi$. Defining r and θ to be the norm and argument of $a+bi$, so $a+bi = r\{\cos(\theta)+i\sin(\theta)\}$, and noting that $\exp(c+di) = \exp(c)\{\cos(d)+i\sin(d)\}$, gives that $c = \log(r)$ and $d = \theta$, so

$$\log(a+bi) = \log(|a+bi|) + i\arg(a+bi) = \log(a^2+b^2)/2 \pm i\arccos\{a/(a^2+b^2)^{1/2}\},$$

where again $\arg(x)$ is the argument of the complex value x. (Note that $\pm\arccos\{a/(a^2+b^2)^{1/2}\} = \pm\arcsin\{b/(a^2+b^2)^{1/2}\} = \pm\tan^{-1}(b/a)$, with the signs in each case chosen to put the point in the proper quadrant, so there are several alternate forms.) Since $\arg(x)$ is not unique, neither is the complex logarithm, but the principal value of the logarithm can be defined as the value obtained from the principal value of the $\arg(\cdot)$ function. From this definition of the logarithm, arbitrary powers of complex numbers can be defined by

$$(a+bi)^{c+di} = \exp\{(c+di)\log(a+bi)\}.$$

□

In both examples, a formal definition of the function has been given of the form (18). While it is true that complex functions can always be represented in this form, it may not always be trivial to give explicit formulas.

Exercise 10.5 Using the complex form of the natural logarithm, give an expression for the principal value of $\log(-a)$ for any positive real a.

A complex function $f : \mathcal{C} \to \mathcal{C}$ is continuous at x_0 if for any $\epsilon > 0$ there is a $\delta_\epsilon > 0$ such that $|f(x) - f(x_0)| < \epsilon$ whenever $|x - x_0| < \delta_\epsilon$. Here $|\cdot|$ is again the complex norm function.

Exercise 10.6 Show that if $f(a + bi) = g(a, b) + h(a, b)i$, then f is continuous at $x_0 = a_0 + b_0 i$ if and only if g and h (as functions from $R^2 \to R^1$) are continuous at (a_0, b_0).

Clearly if f is continuous at x_0, then $f(x_n) \to f(x_0)$ whenever $x_n \to x_0$. The converse is also true, as given in the next exercise.

Exercise 10.7 Show that if $f(x_n) \to f(x_0)$ for every sequence $x_n \to x_0$, then f is continuous at x_0.

A continuous limit can also be defined for complex functions. $\lim_{x \to x_0} f(x) = f_0$ if for any $\epsilon > 0$, there is a $\delta_\epsilon > 0$ such that $|f(x) - f_0| < \epsilon$ for $|x - x_0| < \delta_\epsilon$. If f is continuous at x_0, then $\lim_{x \to x_0} f(x) = f(x_0)$.

The derivative of a complex function f at x can be defined by

$$f'(x) = \lim_{z \to 0} \{f(x + z) - f(x)\}/z, \tag{19}$$

if the limit exists. Here $z \to 0$ in the set of complex numbers, and the division by z is a complex division. Because $z = c + di$ is two-dimensional, there are many paths by which z can approach 0. For the derivative to exist, the above limit must exist and give the same value along any such path.

A function f is *analytic* (or holomorphic or regular) at x, if it is defined and is differentiable at every point in some neighborhood of x.

Suppose $f(a + bi) = g(a, b) + h(a, b)i$. Representing $z = c + di$, consider taking the limit in (19) by first taking the limit as $d \to 0$, and then the limit as $c \to 0$. If g and h are continuous and have partial derivatives, then

$$\lim_{d \to 0} \frac{g(a + c, b + d) + h(a + c, b + d)i - g(a, b) - h(a, b)i}{c + di} = \frac{g(a + c, b) + h(a + c, b)i - g(a, b) - h(a, b)}{c}$$

and

$$\lim_{c \to 0} \frac{g(a + c, b) + h(a + c, b)i - g(a, b) - h(a, b)i}{c} = \frac{\partial g(a, b)}{\partial a} + \frac{\partial h(a, b)}{\partial a} i. \tag{20}$$

Similarly, taking $\lim_{c\to 0}$ first, and then $\lim_{d\to 0}$, gives

$$-i\frac{\partial g(a,b)}{\partial b} + \frac{\partial h(a,b)}{\partial b}, \tag{21}$$

since $1/i = -i$. If the derivative of f exists at $a+bi$, then (20) and (21) must be equal, since they are limits as $z \to 0$ along two different paths. These expressions are equal if

$$-\frac{\partial g(a,b)}{\partial b} = \frac{\partial h(a,b)}{\partial a} \quad \text{and} \quad \frac{\partial g(a,b)}{\partial a} = \frac{\partial h(a,b)}{\partial b}. \tag{22}$$

These relationships are known as the Cauchy-Riemann equations. If these four partial derivatives are continuous at (a,b) and satisfy (22), then it can be shown that f is analytic at $a+bi$, and in particular, f is differentiable, and the derivative is given by either of (20) or (21).

Example 10.3 For $f(a+bi) = \exp(a+bi)$, $g(a,b) = \exp(a)\cos(b)$ and $h(a,b) = \exp(a)\sin(b)$. Thus

$$\frac{\partial g(a,b)}{\partial a} = \exp(a)\cos(b), \quad \frac{\partial g(a,b)}{\partial b} = -\exp(a)\sin(b), \quad \frac{\partial h(a,b)}{\partial a} = \exp(a)\sin(b),$$

and

$$\frac{\partial h(a,b)}{\partial b} = \exp(a)\cos(b),$$

so (22) is satisfied, and from (20), the derivative of $\exp(x)$ at $x = a+bi$ is $\exp(a)\cos(b) + \exp(a)\sin(b)i = \exp(a+bi)$. Thus the exponential function is its own derivative, just as it is for real numbers. □

Exercise 10.8 Show that the principal value of $\log(x)$ is differentiable for $x \neq 0$ and $-\pi < \arg(x) < \pi$, with $d\log(x)/dx = 1/x$.

It can be shown that usual properties of derivatives, such as formulas for derivatives of products and the chain rule for differentiating composite functions, apply to derivatives of complex functions (at points where they are differentiable).

10.3 Measurability and Integration

Recall that the σ-field of Borel sets in R^1 (or R^k) is the σ-field generated by the open sets; that is, the smallest σ-field containing the open sets. Since we have a

well-defined system of open sets for complex numbers, we can define a σ-field of measurable sets in a similar way. Let $\mathcal{B}_c$ be the smallest σ-field containing the open sets in $\mathcal{C}$. (This is not the only useful definition of measurable complex sets. Recall in particular that the Lebesgue measurable sets in real spaces is a larger collection than that generated by the open sets.)

Let $\mathcal{F}$ be a σ-field of measurable subsets of a space Ω. A function $f : \Omega \to \mathcal{C}$ is measurable with respect to $\mathcal{F}$ if the pre-image of any measurable complex set is in $\mathcal{F}$. That is, f is measurable if $f^{-1}(A) = \{\omega \in \Omega : f(\omega) \in A\} \in \mathcal{F}$ for every measurable $A \subset \mathcal{C}$.

As is the case in real Euclidean spaces, the σ-field $\mathcal{B}_c$ generated by the open sets can be generated by various other collections of sets. In particular, $\mathcal{B}_c$ is generated by the collection of open rectangles of the form

$$A_r = \{a + bi : r_1 < a < r_2, r_3 < b < r_4\}.$$

In fact, any open set can be expressed as a countable union of such sets (this is related to the fact that the set of rational numbers is everywhere dense, so only rectangles with rational vertices need to be considered, and the set of rational numbers is countable). For $f : \Omega \to \mathcal{C}$, setting $g = \Re(f)$ and $h = \Im(f)$, clearly $f(\omega) = g(\omega) + h(\omega)i \in A_r$ if and only if $r_1 < g(\omega) < r_2$ and $r_3 < h(\omega) < r_4$. Thus $f^{-1}(A_r) = g^{-1}\{(r_1, r_2)\} \cap h^{-1}\{(r_3, r_4)\}$. Since the collection of open rectangles of the form A_r generates $\mathcal{B}_c$, f is measurable if $f^{-1}(A_r) \in \mathcal{F}$ for all such A_r. This is equivalent to $g^{-1}(I_r) \in \mathcal{F}$ and $h^{-1}(I_r) \in \mathcal{F}$ for all open intervals $I_r = (r_1, r_2)$, which is equivalent to the Borel measurability of g and h (see eg Theorem 13.1 (i), Billingsley, 1995). Thus a complex valued function f is measurable if and only if the two real valued functions $g(\cdot) = \Re(f(\cdot))$ and $h(\cdot) = \Im(f(\cdot))$ are measurable.

With concepts of measurability defined, it would be possible to directly build up a general theory of integration for complex valued functions similar to that for real valued functions. However, the representation (17) provides a simpler approach. Suppose $(\Omega, \mathcal{F}, P)$ is a probability space, and that $f : \Omega \to \mathcal{C}$ is a measurable complex valued function. Then $f(\omega) = g(\omega) + h(\omega)i$ for some real valued measurable functions g, h. As with real valued functions (also called random variables), f induces a probability measure on $\mathcal{C}$. That is, $P(f(\omega) \in A) = P(\omega \in f^{-1}(A))$ for measurable A. The integral of f with respect to the measure P is defined to be

$$\int f(\omega)\, dP(\omega) = \int g(\omega)\, dP(\omega) + i \int h(\omega)\, dP(\omega).$$

f is defined to be integrable if both g and h are; that is, if $\int |g|\, dP < \infty$ and $\int |h|\, dP < \infty$. Since

$$\max\{|g|, |h|\} \le |f| \le |g| + |h|$$

(by (14)), f is integrable if $\int |f|\, dP < \infty$, and integrability of f implies integrability of g and h. Also,

$$\left| \int f\, dP \right| \le \int |f|\, dP. \tag{23}$$

To see that this is true, first note that if g and h are simple functions (step functions) taking values g_j and h_j on a measurable partition of Ω, this result follows directly from the triangle inequality, since then $\int f\, dP = \sum_j \alpha_j (g_j + h_j i)$ for some constants $\alpha_j \ge 0$ (the α_j are the probabilities of the corresponding components of the partition), and $|\sum_j \alpha_j (g_j + h_j i)| \le \sum_j \alpha_j |g_j + h_j i| = \int |f|\, dP$. Since the integrals in the general case are limits of sequences of integrals of simple functions, (23) must hold in general, too.

Since the complex integral is defined in terms of two real integrals, the usual properties of integration, such as $\int (f_1 + f_2)\, dP = \int f_1\, dP + \int f_2\, dP$, are immediate. The convergence theorems of Lebesgue integration also can be generalized to complex functions by applying them separately to the real and imaginary parts of f.

In the special case where $\Omega = R^1$ and P is absolutely continuous with respect to Lebesgue measure, with density $p(\cdot)$, it follows that

$$\int f(u)\, dP(u) = \int g(u) p(u)\, du + i \int h(u) p(u)\, du.$$

10.4 Characteristic Functions

The characteristic function of a real valued random variable X with distribution P is

$$\phi(t) = \mathrm{E}\{\exp(itX)\} = \int \exp(itu)\, dP(u) = \int \cos(tu)\, dP(u) + i \int \sin(tu)\, dP(u).$$

Since $\int |\cos(tu)|\, dP(u) \leq 1$ and $\int |\sin(tu)|\, dP(u) \leq 1$, $\exp(itu)$ is integrable, and the characteristic function always exists. Also,

$$\begin{aligned} |\phi(t+h) - \phi(t)| &\leq \int |\exp(itx)||\exp(ihx) - 1|\, dP(x) \\ &= \int [\{\cos(hx) - 1\}^2 + \sin(hx)^2]^{1/2}\, dP(x) \\ &= \int \{2 - 2\cos(hx)\}^{1/2}\, dP(x), \end{aligned}$$

and this last expression $\to 0$ as $h \to 0$, so $\phi(t)$ is uniformly continuous in t. Also note $\phi(0) = 1$, and $|\phi(t)| \leq \int |\exp(itx)|\, dP(x) = 1$ for all t, by (23).

Example 10.4 Suppose X has support on the countable set $\{x_1, x_2, \ldots\}$, with $P(X = x_j) = p_j$. Then

$$\phi(t) = \sum_{j=1}^{\infty} \cos(tx_j)p_j + i \sum_{j=1}^{\infty} \sin(tx_j)p_j,$$

which can be written more compactly as $\sum_j \exp(itx_j)p_j$ (rearrangement of the terms in the series can be justified by the fact that $\sum_j |\exp(itx_j)p_j| < \infty$). □

Example 10.5 If $X \sim U(0,1)$, then

$$\phi(t) = \int_0^1 \cos(tu)\, du + i \int_0^1 \sin(tu)\, du = t^{-1}[\sin(t) - i\{\cos(t) - 1\}] = i\{1 - \exp(it)\}/t.$$

□

For symmetric distributions with $P(X < -x) = P(X > x)$ for all x, the contributions from the sine term at positive and negative x cancel out, and

$$\phi(t) = \int \cos(tu)\, dP(u),$$

and hence is real.

In general, evaluating the trigonometric integrals in the definition of the characteristic function is not trivial. Often it is better to approach the problem through the theory of line integration in the complex plain. These techniques are

beyond the scope of this course, however. The following result provides a somewhat indirect means of obtaining the characteristic function of many common distributions.

Theorem 10.2 *Let $\phi(t)$ be the characteristic function of the distribution P. If $\psi(t) = \int \exp(tu)\, dP(u) < \infty$ for all t in a neighborhood of 0, then $\phi(t) = \psi(it)$, where $\psi(it)$ is defined as given in the proof below.*

Proof: Since $\psi(t) = \int_{-\infty}^{0} \exp(tu)\, dP(u) + \int_{0}^{\infty} \exp(tu)\, dP(u)$ both integrals must be finite for $\psi(t)$ to be finite, and since this must hold for both positive and negative values of t in a neighborhood of 0, $\int \exp(|tu|)\, dP(u) < \infty$ for t in some neighborhood of 0. Thus by (47),

$$\infty > \int \exp(|tu|)\, dP(u) = \int \sum_k \frac{|tu|^k}{k!}\, dP(u) = \sum_k \int \frac{|tu|^k}{k!}\, dP(u),$$

so by (48),

$$\psi(t) = \int \sum_k \frac{(tu)^k}{k!}\, dP(u) = \sum_k \int \frac{(tu)^k}{k!}\, dP(u) = \sum_k \frac{t^k}{k!} \int u^k\, dP(u).$$

Then $\psi(it)$ is defined by substituting it for t in the final series in this expression. From the previous argument, this series is absolutely convergent for t in a neighborhood of 0, so this function is well defined there, and

$$\begin{aligned}
\psi(it) &= \sum_{k=0}^{\infty} \frac{(it)^k}{k!} \int u^k\, dP(u) \\
&= \sum_{k=0}^{\infty} (-1)^{k+2} \frac{t^{2k}}{(2k)!} \int u^{2k}\, dP(u) + i \sum_{k=0}^{\infty} (-1)^{k+2} \frac{t^{2k+1}}{(2k+1)!} \int u^{2k+1}\, dP(u) \\
&= \int \sum_{k=0}^{\infty} (-1)^{k+2} \frac{(tu)^{2k}}{(2k)!}\, dP(u) + i \int \sum_{k=0}^{\infty} (-1)^{k+2} \frac{(tu)^{2k+1}}{(2k+1)!}\, dP(u) \\
&= \int \cos(tu)\, dP(u) + i \int \sin(tu)\, dP(u) \\
&= \phi(t),
\end{aligned}$$

again using (48), the fact that the series are absolutely convergent, and from the Taylor series representations of the sine and cosine functions. □

Example 10.6 For the standard normal distribution, $\psi(t) = \exp(t^2/2)$. Thus $\phi(t) = \exp(-t^2/2)$.

Example 10.7 For the gamma distribution with shape α and scale=1, $\psi(t) = \int x^{\alpha-1} \exp\{-(1-t)x\}\, dx/\Gamma(\alpha) = (1-t)^{-\alpha}$ (for $t < 1$). Thus $\phi(t) = (1-it)^{-\alpha} = \exp\{-\alpha \log(1-it)\}$.

More generally, if $\psi(t)$ involves functions that do not have complex analogs defined, then it may be necessary to work directly with the series representation.

If $Y = \sigma X + \mu$, then

$$\phi_Y(t) = \int \exp\{it(\sigma u + \mu)\}\, dP_X(u) = \exp(it\mu) \int \exp\{i(t\sigma)u\}\, dP_X(u) = \exp(it\mu)\phi_X(\sigma t).$$

10.5 Convolutions

One of the most important properties of characteristic functions is that if X and Y are independent, then $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$; that is, the characteristic function of the sum of independent random variables is the product of their characteristic functions. This follows because

$$\begin{aligned}
\phi_X(t)\phi_Y(t) &= [\mathrm{E}\{\cos(tX)\} + i\mathrm{E}\{\sin(tX)\}][\mathrm{E}\{\cos(tY)\} + i\mathrm{E}\{\sin(tY)\}] \\
&= \mathrm{E}\{\cos(tX)\}\mathrm{E}\{\cos(tY)\} - \mathrm{E}\{\sin(tX)\}\mathrm{E}\{\sin(tY)\} + i[\mathrm{E}\{\cos(tX)\}\mathrm{E}\{\sin(tY)\} \\
&\quad + \mathrm{E}\{\sin(tX)\}\mathrm{E}\{\cos(tY)\}] \\
&= \mathrm{E}\{\cos(tX)\cos(tY) - \sin(tX)\sin(tY)\} + i[\mathrm{E}\{\cos(tX)\sin(tY) + \sin(tX)\cos(tY)\}] \\
&= \mathrm{E}\{\cos(tX + tY)\} + i\mathrm{E}\{\sin(tX + tY)\} \\
&= \phi_{X+Y}(t). \qquad (24)
\end{aligned}$$

Of course, this extends to the sum of any number independent random variables.

10.6 Taylor Series and Derivatives

The convolution formula above makes it easy to give the characteristic function for a sum of independent random variables. To study the asymptotic properties of such sums, formulas for series expansions, with error bounds, are useful.

To derive appropriate formulas, we can start with the expansions for the sine and cosine functions. Since $d^k \cos(x)/dx^k|_{x=0} = \cos(k\pi/2)$, Taylor's theorem gives

$$\cos(x) = \sum_{k=0}^{n} \cos(k\pi/2)\frac{x^k}{k!} + \frac{1}{n!}\int_0^x \frac{d^{n+1}\cos(t)}{dt^{n+1}}(x-t)^n\,dt.$$

By separately considering positive and negative x, it is easily verified that in either case,

$$\left|\int_0^x \frac{d^{n+1}\cos(t)}{dt^{n+1}}(x-t)^n\,dt\right| \le \int_0^{|x|}(|x|-t)^n\,dt = \frac{|x|^{n+1}}{n+1}.$$

Also, using integration by parts,

$$\left|\int_0^x \frac{d^{n+1}\cos(t)}{dt^{n+1}}(x-t)^n\,dt\right| = \left|-\cos(n\pi/2)x^n + n\int_0^x \frac{d^n\cos(t)}{dt^n}(x-t)^{n-1}\,dt\right| \le 2|x|^n.$$

Thus

$$\left|\cos(x) - \sum_{k=0}^{n}\cos(k\pi/2)\frac{x^k}{k!}\right| \le \min\left(\frac{|x|^{n+1}}{(n+1)!}, \frac{2|x|^n}{n!}\right).$$

Similarly,

$$\left|\sin(x) - \sum_{k=0}^{n}\sin(k\pi/2)\frac{x^k}{k!}\right| \le \min\left(\frac{|x|^{n+1}}{(n+1)!}, \frac{2|x|^n}{n!}\right).$$

Since $i^k = \cos(k\pi/2) + i\sin(k\pi/2)$, where $i^0 = 1$ by definition, combining the previous 2 results gives that

$$\begin{aligned}\left|\exp(ix) - \sum_{k=0}^{n}\frac{(ix)^k}{k!}\right| &= \left|\cos(x) + i\sin(x) - \sum_{k=0}^{n}\cos(k\pi/2)\frac{x^k}{k!} - i\sum_{k=0}^{n}\sin(k\pi/2)\frac{x^k}{k!}\right| \\ &\le 2\min\left(\frac{|x|^{n+1}}{(n+1)!}, \frac{2|x|^n}{n!}\right). \qquad (25)\end{aligned}$$

Using this expression, it then follows that if $\phi(t)$ is the characteristic function of X, and $\mathrm{E}(|X|^n) < \infty$, then

$$\left|\phi(t) - \sum_{k=0}^{n}\frac{(it)^k}{k!}\mathrm{E}(X^k)\right| \le 2\mathrm{E}\left\{\min\left(\frac{|tX|^{n+1}}{(n+1)!}, \frac{2|tX|^n}{n!}\right)\right\}. \qquad (26)$$

(With a bit more work, it can be shown that the leading factor of 2 on the error bound is not needed, but that is not important in most applications.) Note that the right hand side of (26) is finite.

Suppose $\mathrm{E}(|X|) < \infty$. Consider $\phi'(0)$. From the definitions,

$$\begin{aligned}\phi'(0) &= \lim_{h\to 0} \frac{\phi(h) - \phi(0)}{h} \\ &= \lim_{h\to 0} \int \frac{\exp(ihu) - 1}{h}\, dP(u).\end{aligned}$$

By (25), $|\{\exp(ihu) - 1\}/h| \le 2|u|$, and by assumption, $\int |u|\, dP(u) < \infty$. Thus by the dominated convergence theorem, the limit may be taken inside the integral, and then

$$\lim_{h\to 0} \frac{\exp(ihu) - 1}{h} = \lim_{h\to 0} \frac{\cos(hu) + i\sin(hu) - 1}{h} = iu,$$

so

$$\phi'(0) = i \int u\, dP(u) = i\mathrm{E}(X).$$

Using an induction argument, it can be shown that if $\mathrm{E}(|X|^n) < \infty$, then $d^n\phi(t)/dt^n$ exists on a neighborhood of 0, and

$$i^{-n} d^n \phi(0)/dt^n = \mathrm{E}(X^n). \tag{27}$$

It can also be shown that if $d^n\phi(t)/dt^n$ exists on a neighborhood of 0, and n is even, then $\mathrm{E}(X^n) < \infty$, so the nth moment exists. For odd n this can fail, since there are distributions where (for example) $\phi'(0)$ exists, but $\mathrm{E}(|X|) = \infty$. What happens in this case is that $\lim_{a\to\infty} \int_{-a}^{a} u\, dP(u)$ exists, but that is not the same as existence of $\mathrm{E}(X)$.

10.7 Cumulants

The cumulant generating function of a random variable X with characteristic function $\phi(t)$ is

$$\kappa(t) = \log \phi(t).$$

If $\mathrm{E}(|X|^j) < \infty$, the jth cumulant of X is defined to be

$$\kappa_j = i^{-j} d^j \kappa(0)/dt^j.$$

From Exercise 10.8, the chain rule, (27), and the fact that $\phi(0) = 1$,

$$\kappa_1 = i^{-1} d\kappa(0)/dt = i^{-1}\phi'(0)/\phi(0) = \mathrm{E}(X),$$

and

$$\kappa_2 = -d^2\kappa(0)/dt^2 = -\phi''(0)/\phi(0) + \{\phi'(0)/\phi(0)\}^2 = \mathrm{Var}(X).$$

Exercise 10.9 Suppose $\mathrm{E}(X^4) < \infty$. Show $\kappa_3 = \mathrm{E}(X-\mu)^3$ and $\kappa_4 = \mathrm{E}(X-\mu)^4 - 3\mathrm{Var}(X)^2$, where $\mu = \mathrm{E}(X)$.

Example 10.8 If $X \sim N(0,1)$, then $\phi(t) = \exp(-t^2/2)$. Thus $\kappa(t) = -t^2/2$, so $\kappa_2 = 1$, and $\kappa_j = 0$ for $j \neq 2$.

In general, if X has characteristic function $\phi(t)$ and cumulant generating function $\kappa(t)$, with cumulants κ_1, $\kappa_2, \ldots$, and $Y = bX + a$, then $\phi_Y(t) = \phi(bt)\exp(ita)$, so $\kappa_Y(t) = ita + \kappa(bt)$. The first cumulant of Y is therefore $a + b\kappa_1$, and the jth cumulant of Y is $b^j\kappa_j$ for $j \geq 2$.

10.8 Uniqueness and Inversion

One of the most important properties of characteristic functions is that they uniquely determine the corresponding probability distribution. That is, different probability distributions will always have different characteristic functions. (It is interesting to note, though, that different distributions can have characteristic functions that agree for t on some interval, but not everywhere; see Chung, 1974, Theorem 6.5.5.) This uniqueness property is a consequence of Theorem 10.3, below, which indicates how to recover the distribution from the characteristic function. This theorem is usually called the inversion theorem.

First 2 preliminary results, needed in the proof, will be established. The first is that if $T > 0$ and c is real, then

$$\int_{-T}^{T} \cos(tc)/t \, dt = 0. \tag{28}$$

This follows because from the change of variables $u = -t$, it follows that $\int_{-T}^{0} \cos(tc)/t \, dt = -\int_{0}^{T} \cos(uc)/u \, du$. The second result is that

$$\lim_{T\to\infty} \int_{0}^{T} \sin(tc)/t \, dt = \mathrm{sign}(c)\pi/2, \tag{29}$$

where $\text{sign}(c) = -1, 0, +1$ for $c < 0$, $c = 0$, and $c > 0$. To see this, first note that the case $c = 0$ is clear. If $c < 0$, then since $\sin(tc) = -\sin(-tc)$, it follows that $\int_0^T \sin(tc)/t\,dt = -\int_0^T \sin(t|c|)/t\,dt$. Thus it remains to show that

$$\lim_{T\to\infty} \int_0^T \sin(tc)/t\,dt = \pi/2,$$

for $c > 0$. From the change of variables $u = ct$,

$$\begin{aligned}
\lim_{T\to\infty} \int_0^T \sin(tc)/t\,dt &= \lim_{T\to\infty} \int_0^{Tc} \sin(t)/t\,dt \\
&= \lim_{T\to\infty} \int_0^{Tc} \sin(t) \int_0^\infty \exp(-tu)\,du\,dt \\
&= \lim_{T\to\infty} \int_0^\infty \int_0^{Tc} \sin(t)\exp(-tu)\,dt\,du \\
&= \lim_{T\to\infty} \int_0^\infty \frac{\{-u\sin(t) - \cos(t)\}\exp(-tu)}{1+u^2}\Big|_{t=0}^{Tc} du \\
&= \int_0^\infty \frac{du}{1+u^2} - \lim_{T\to\infty} \int_0^\infty \frac{\exp(-Tcu)\{u\sin(Tc) + \cos(Tc)\}\,du}{1+u^2} \\
&= \tan^{-1}(\infty) - \tan^{-1}(0) - 0 \\
&= \pi/2.
\end{aligned}$$

(Several steps in this argument require verification of regularity conditions. These are left as exercises.)

Theorem 10.3 *Let $F(x) = P(X \le x)$, and let $\phi(t)$ be the characteristic function of this distribution. If $P(X = a) = P(X = b) = 0$, then*

$$F(b) - F(a) = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{\exp(-ita) - \exp(-itb)}{it} \phi(t)\,dt. \tag{30}$$

Proof: By Fubini's Theorem, which can be applied because the integrand is

bounded,

$$\begin{aligned}
&\int_{-T}^{T} \frac{\exp(-ita) - \exp(-itb)}{2\pi it}\phi(t)\,dt \\
&= \int_{-T}^{T} \frac{\exp(-ita) - \exp(-itb)}{2\pi it} \int_{-\infty}^{\infty} \exp(itu)\,dF(u)\,dt \\
&= \int_{-\infty}^{\infty}\int_{-T}^{T} \frac{\exp\{it(u-a)\} - \exp\{it(u-b)\}}{2\pi it}\,dt\,dF(u) \\
&= \int_{-\infty}^{\infty}\int_{-T}^{T} \frac{\cos\{t(u-a)\} - \cos\{t(u-b)\}}{2\pi it} + i\frac{\sin\{t(u-a)\} - \sin\{t(u-b)\}}{2\pi it}\,dt\,dF(u) \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-T}^{T} \frac{\sin\{t(u-a)\} - \sin\{t(u-b)\}}{t}\,dt\,dF(u) \\
&= \frac{1}{\pi}\int_{-\infty}^{\infty}\int_{0}^{T} \frac{\sin\{t(u-a)\} - \sin\{t(u-b)\}}{t}\,dt\,dF(u),
\end{aligned}$$

from (28), and the fact that $\sin(t)/t = \sin(-t)/(-t)$. Taking the limit as $T \to \infty$, using the dominated convergence theorem to bring the limit inside the integral (which can be justified by showing that the integrand is bounded), and using (29), it follows that

$$\begin{aligned}
&\lim_{T\to\infty} \frac{1}{2\pi}\int_{-T}^{T} \frac{\exp(-ita) - \exp(-itb)}{it}\phi(t)\,dt \\
&= \frac{1}{\pi}\int_{-\infty}^{\infty} \lim_{T\to\infty}\int_{0}^{T} \frac{\sin\{t(u-a)\} - \sin\{t(u-b)\}}{t}\,dt\,dF(u) \\
&= \frac{1}{\pi}\int_{-\infty}^{\infty} [\mathrm{sign}(u-a)\pi/2 - \mathrm{sign}(u-b)\pi/2]\,dF(u) \\
&= \frac{1}{\pi}\int_{-\infty}^{\infty} \pi\{I(a<u<b) + I(u=a)/2 + I(u=b)/2\}\,dF(u) \\
&= \int_{a}^{b} dF(u) \\
&= F(b) - F(a),
\end{aligned}$$

since $P(X = a) = P(X = b) = 0$. □

Corollary 1. *If $\phi_X(t) = \phi_Y(t)$ for all t, then X and Y have the same distribution.*

Proof: By the previous theorem, the characteristic function uniquely determines $P(a < X < b)$ for all a, b with $P(X = a) = P(X = b) = 0$. There are at most a countable number of points c_j with $P(X = c_j) > 0$ (see eg Billingsley, 1995, Theorem 10.2). For any mass point c_j, $P(X = c_j) = \lim_{\delta \to 0} P(c_j + \delta > X > c_j - \delta)$ (the limit must exist because $P(c_j + \delta > X > c_j - \delta)$ is monotone and bounded as a function of δ). Similarly, $P(Y = c_j) = \lim_{\delta \to 0} P(c_j + \delta > Y > c_j - \delta)$. Since $P(c_j + \delta > X > c_j - \delta) = P(c_j + \delta > Y > c_j - \delta)$ except for at most a countable number of δ,

$$P(X = c_j) - P(Y = c_j) = \lim_{\delta \to 0} \{P(c_j + \delta > X > c_j - \delta) - P(c_j + \delta > Y > c_j - \delta)\} = 0.$$

Thus all mass points must have the same mass, and the two distributions must be the same. □

Corollary 2. *If* $\int_{-\infty}^{\infty} |\phi(t)|\, dt < \infty$, *then* F *is absolutely continuous with density*

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-itx)\phi(t)\, dt. \tag{31}$$

Proof: Since $|\cos(x) - 1| \leq |x|$ and $|\sin(x)| \leq |x|$,
$|\exp(ix) - 1| \leq |\cos(x) - 1| + |i \sin(x)| \leq 2|x|$. Thus

$$\left| \frac{\exp(-itb) - \exp(-ita)}{it} \right| = \frac{|\exp\{-it(b-a)\} - 1|}{|t|} \leq 2|b - a|.$$

Applying the previous Theorem,

$$F(b) - F(a) \leq \frac{b-a}{\pi} \int_{-\infty}^{\infty} |\phi(t)|\, dt \to 0$$

as $(b - a) \to 0$, so there can be no mass points, and the distribution is absolutely continuous. Also,

$$\begin{aligned} \lim_{h \to 0} \frac{F(x+h) - F(x)}{h} &= \lim_{h \to 0} \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{\exp(-itx) - \exp\{-it(x+h)\}}{ith} \phi(t)\, dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-itx)\phi(t)\, dt, \end{aligned}$$

since the limit can be taken inside the integral, because the integrand is dominated by $2|\phi(t)|$ (for h small), which by assumption is integrable, and

$$\lim_{h \to 0} \frac{1 - \exp(-ith)}{ith} = \frac{1 - \cos(-th) - i \sin(-th)}{ith} = 1,$$

since $\{1-\cos(x)\}/x \to 0$ and $\sin(x)/x \to 1$ as $x \to 0$. Thus $f(x) = F'(x)$ exists and has the required form. □

Corollary 3. *The cumulative distribution function is given by*

$$F(x) = \frac{1}{2} + \frac{1}{2\pi} \lim_{T\to\infty} \int_0^T r(t,x)\,dt = \frac{1}{2} - \frac{1}{\pi} \lim_{T\to\infty} \int_0^T \frac{\Im\{\exp(-itx)\phi(t)\}}{t}\,dt, \qquad (32)$$

at points x where F is continuous, where

$$r(t,x) = \frac{\exp(itx)\phi(-t) - \exp(-itx)\phi(t)}{it}.$$

Proof (outline): First note that $\Im(x) = (x - \overline{x})/(2i)$ for any complex x. The equivalence of the two integrals in (32) follows from this, and because $\overline{\phi(t)} = \phi(-t)$ and $\overline{\exp(-itx)} = \exp(itx)$ (see also Exercise 10.2).

Now suppose $x > 0$ (the case $x < 0$ requires only minor changes in notation), and x and 0 are continuity points of F. (A similar argument can be used when 0 is a mass point, taking into account that (30) gives the average of the right and left hand limits of F at mass points a or b.) Set $b = x$ and $a = 0$ in (30), and add this to the formula obtained from (30) by making the change of variables $t = -t$ in the integral. This gives

$$2\{F(x) - F(0)\} = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^T r(t,x)\,dt - \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^T r(t,0)\,dt.$$

Also, since $\Im\{\exp(-itx)\phi(t)\}/t$ is always an even function of t, $\int_{-T}^T r(t,x)\,dt = 2\int_0^T r(t,x)\,dt$. Thus

$$F(x) = \frac{1}{2\pi} \lim_{T\to\infty} \int_0^T r(t)\,dt + F(0) - \lim_{T\to\infty} \frac{1}{2\pi} \int_0^T r(t,0)\,dt. \qquad (33)$$

Now $r(t,0) = -2\int \sin(tx)/t\,dF(x)$, and similar to the derivation of (30),

$$\begin{aligned}
\lim_{T\to\infty} \frac{1}{2\pi} \int_0^T r(t,0)\,dt &= -\frac{1}{\pi} \int \lim_{T\to\infty} \int_0^T \frac{\sin(tx)}{t}\,dt\,dF(x) \\
&= -\frac{1}{2} \int \text{ sign } (x)\,dF(x) \\
&= -\frac{1}{2}\left\{\int_0^\infty dF(x) - \int_{-\infty}^0 dF(x)\right\} \\
&= -\{1 - F(0) - F(0) + 0\}/2 \\
&= \{2F(0) - 1\}/2.
\end{aligned}$$

Substituting this expression in (33) completes the proof. □

The inversion formula (30) is expressed in terms of a limit because generally the integrand is not integrable over the entire real line. The limit always exists, though, provided $\phi(t)$ is the characteristic function of a probability distribution or sub-distribution.

Example 10.9 From Example 10.5, the characteristic function of the $U(0,1)$ distribution is $\phi(t) = i\{1 - \exp(it)\}/t$. Using (30) with $a = 0$ and $b = x$ gives that the cumulative distribution function is

$$\begin{aligned} F(x) &= \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{1-\exp(-itx)}{t} \frac{1-\exp(it)}{t}\, dt \\ &= \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{1-\cos(t)-\cos(-tx)+\cos(t-tx)}{t^2}\, dt \\ &= \frac{1}{2}\{1 - x\ \mathrm{sign}(-x) - (1-x)\ \mathrm{sign}(1-x)\} \\ &= xI(0 < x < 1) + I(x \geq 1), \end{aligned}$$

since the sine terms implicit in the first line integrate to 0, and

$$\int_{-T}^{T} \frac{\cos(bt)}{t^2}\, dt = -\frac{\cos(bt)}{t}\bigg|_{-T}^{T} - \int_{-T}^{T} \frac{b\sin(bt)}{t}\, dt \to -\pi b\ \mathrm{sign}\ (b),$$

as $T \to \infty$. □

The inversion theorem provides an indirect way of deriving characteristic functions for some distributions, as illustrated in the following example.

Example 10.10 The double exponential (Laplace) distribution has density $\exp(-|x|)/2$. Since this is symmetric, its characteristic function is given by

$$\phi(t) = \int_0^\infty \cos(tx)\exp(-x)\, dx = \frac{\exp(-x)}{1+t^2}\{t\sin(tx) - \cos(tx)\}\bigg|_0^\infty = \frac{1}{1+t^2},$$

where the antiderivative can be obtained by integrating by parts twice (and can be verified by differentiation). From (31), it then follows that

$$\exp(-|x|)/2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp(-itx)\frac{1}{1+t^2}\, dt.$$

Making the change of variable $u = -t$ and multiplying by 2 gives that

$$\frac{1}{\pi}\int_{-\infty}^{\infty} \exp(iux)\frac{1}{1+u^2}\,du = \exp(-|x|).$$

Since the Cauchy distribution has density $1/\{\pi(1+u^2)\}$, this last expression shows that the characteristic function of the Cauchy distribution is $\phi(t) = \exp(-|t|)$. Note that the Cauchy distribution does not have any finite moments, and its characteristic function is not differentiable at 0. □

Exercise 10.10 Suppose $X \sim$Bernoulli(p).
(a) Show $\phi_X(t) = (1-p) + \exp(it)p$.
(b) Calculate the inversion formula (32) at the mass points $x = 0$ and $x = 1$. Show that the result is the average of the left- and right- hand limits of $F(x)$ at these points.

More interesting applications involve settings where the density or the cumulative distribution function is difficult to derive or compute directly, but the characteristic function can easily be given. Then the inversion formulas provide a means of computing or approximating the distribution. As will be seen later, this is the basis of the central limit theorem and various other asymptotic approximations. In the following example, the inversion formula gives a one-dimensional integral for the exact cumulative distribution function, which can then be evaluated using numerical quadrature methods.

Example 10.11 Consider computing $F(x) = P(X'AX \le x)$, where $A_{k\times k}$ is an arbitrary (known) symmetric matrix, and $X \sim N(\mu, V)$, the k-variate normal distribution with mean vector μ and covariance matrix V (assumed positive definite). Quadratic forms in normal variables arise in a variety of statistical problems, especially in models for variance components. Let B be such that $BB' = V$ (eg the Choleski factor). Then $W = B^{-1}(X-\mu) \sim N(0, I)$, and $X = BW + \mu$. Also, let $Q'DQ$ be the spectral decomposition of $B'AB$. That is, $Q'DQ = B'AB$, D is diagonal with diagonal elements d_j (the eigenvalues of $B'AB$),

and Q is orthogonal ($Q'Q = QQ' = I$; Q consists of the eigenvectors of $B'AB$). Then

$$\begin{aligned} X'AX &= (W + B^{-1}\mu)'B'AB(W + B^{-1}\mu) \\ &= (W + B^{-1}\mu)'Q'DQ(W + B^{-1}\mu) \\ &= (QW + QB^{-1}\mu)'D(QW + QB^{-1}\mu) \\ &= (Z + \nu)'D(Z + \nu) \\ &= \sum_{j=1}^{k} d_j(Z_j + \nu_j)^2, \end{aligned}$$

since D is diagonal, where $Z = QW \sim N(0, I)$, so the Z_j are iid $N(0, 1)$, and $\nu = QB^{-1}\mu$.

Since $(Z_j + \nu_j)^2$ has a noncentral chi-square distribution with 1 degree of freedom, its characteristic function has a known form. It is also easily derived, using Theorem 10.2. For $|t| < 1/2$, the moment generating function of $(Z_j + \nu_j)^2$ is

$$\begin{aligned} \psi_j(t) &= \int_{-\infty}^{\infty} \frac{1}{(2\pi)^{1/2}} \exp\{t(z + \nu_j)^2\} \exp(-z^2/2)\, dz \\ &= \frac{\exp\{t\nu_j^2 + 2(t\nu_j)^2/(1-2t)\}}{(1-2t)^{1/2}} \int_{-\infty}^{\infty} \frac{(1-2t)^{1/2}}{(2\pi)^{1/2}} \exp[-(1-2t)/2\{z - 2t\nu_j/(1-2t)\}^2]\, dz \\ &= \exp\{t\nu_j^2/(1-2t)\}/(1-2t)^{1/2}, \end{aligned}$$

since the last integrand is the density of the $N\{2t\nu_j/(1-2t), 1/(1-2t)\}$ distribution. Then the characteristic function of $(Z_j + \nu_j)^2$ is $\phi_j(t) = \psi(it)$, and because the terms in the sum are independent, the characteristic function of $\sum_j d_j(Z_j + \nu_j)^2$ is

$$\phi(t) = \prod_j \psi_j(d_j it) = \prod_j \exp\{d_j it\nu_j^2/(1-2d_j it)\}/(1-2d_j it)^{1/2}$$

and by (32),

$$F(x) = \frac{1}{2} - \frac{1}{\pi}\int_0^{\infty} \frac{\Im\{\exp(-itx)\phi(t)\}}{t}\, dt. \tag{34}$$

Thus the next step is to derive an expression for

$$\Im\{\exp(-itx)\phi(t)\} = \Im\left\{\exp\left(-itx + \sum_j \frac{d_j it\nu_j^2}{1-2d_j it} - \frac{1}{2}\sum_j \log(1-2d_j it)\right)\right\}.$$

Using the definition of complex division,

$$\frac{i}{1-2d_jit} = \frac{-2td_j+i}{1+4t^2d_j^2},$$

and by Example 10.2,

$$\log(1-2d_jit) = \frac{1}{2}\log(1+4t^2d_j^2) + i\tan^{-1}(-2td_j) = \frac{1}{2}\log(1+4t^2d_j^2) - i\tan^{-1}(2td_j),$$

so

$$\begin{aligned}
\Im\{\exp(-itx)\phi(t)\} &= \frac{\exp\{-\sum_j 2(td_j\nu_j)^2/(1+4t^2d_j^2)\}}{\prod_j(1+4t^2d_j^2)^{1/4}} \times \\
&\quad \Im\left\{\exp\left(-itx + i\sum_j\{t\nu_j^2 d_j/(1+4t^2d_j^2) + (1/2)\tan^{-1}(2td_j)\}\right)\right\} \\
&= \frac{\exp\{-\sum_j 2(td_j\nu_j)^2/(1+4t^2d_j^2)\}}{\prod_j(1+4t^2d_j^2)^{1/4}} \\
&\quad \times \sin\left(-tx + \sum_j\{t\nu_j^2 d_j/(1+4t^2d_j^2) + (1/2)\tan^{-1}(2td_j)\}\right).
\end{aligned}$$

Substituting this formula in (**??**) then gives an expression for $F(x)$. The integral can be evaluated by standard methods, such as the trapezoidal rule (because of the oscillating nature of the integrand, more sophisticated approaches may not perform well). This method for evaluating the distribution of quadratic forms was given by Imhoff (1961, *Biometrika*, 48:419–426). □

Exercise 10.11 Suppose $X_1, \ldots, X_n$ are iid Cauchy random variables. Find the distribution of $\sum_{i=1}^n X_i/n$ (this distribution can be recognized directly from its characteristic function, so a formal inversion is not needed).

Exercise 10.12 Suppose $U_1, \ldots, U_n$ are iid $U(0,1)$, and set $X = U_1 + \cdots + U_n$. Use (32) to give an expression for $P(X \le x)$.

11 Distributions on Spaces of Sequences

Let $X_1, X_2, \ldots$ be independent Bernoulli random variables, with $P(X_j = 1) = p$. Let Ω be the set of all possible outcomes $x = (x_1, x_2, \ldots)$,

$$\Omega = \{x = (x_1, x_2, \ldots) : x_j = 0, 1, j = 1, 2, \ldots\}$$

The cardinality of Ω is the same as that of the real numbers. This follows because every real number $x \in [0,1]$ has a unique representation as $x = \sum_{j=1}^{\infty} x_j/2^j$, where each $x_j = 0, 1$, so there is a 1 to 1 correspondence between the real numbers in $[0,1]$ and the elements of Ω.

Let $X = (X_1, X_2, \ldots)$, and

$$s(x,n) = \sum_{i=1}^{n} x_i.$$

In this space, events on the limiting behavior of the partial sums $s(X,n) = \sum_{i=1}^{n} X_i$, such as

$$P\{\lim s(X,n)/n = p\} \tag{35}$$

and $P\{\lim n^{-1/2}|s(X,n) - np| > \alpha\}$, are of interest. In interpreting these quantities, it is important to realize that each point in the sample space is a binary sequence $x = (x_1, x_2, \ldots)$. For each such point, the event either is or is not satisfied. For example, if $x_j = 1$ for all j (and $p < 1$), then $\lim s(x,n)/n = 1 \neq p$, while $\lim n^{-1/2}|s(x,n) - np| = \lim n^{1/2}(1-p) = \infty > \alpha$ for any finite α. On the other hand, if $p = 1/2$ and $x_{2j} = 1$ and $x_{2j-1} = 0$, $j = 1, 2, \ldots$, then $\lim s(x,n)/n = p$ and $\lim n^{-1/2}|s(x,n) - np| = 0$. The probability that the limit satisfies the stated criterion is the probability of all sequences $x \in \Omega$ which satisfy the event. Thus (35) can equivalently be expressed $P(X \in R)$, where $R = \{x \in \Omega : \lim_n s(x,n)/n = p\}$. Note that there is nothing random in the definition of R; it just consists of all binary sequences with the stated property.

This section is concerned with making the concept of probability distributions on spaces of sequences more precise. Since the cardinality of this space is the same as for real numbers, it turns out that it is not possible to define a probability distribution consistently on all possible subsets of Ω. Thus issues of measurability will also need to be considered.

The statement (35) should not be confused with

$$\lim_n P(|s(X,n)/n - p| < \epsilon) = 1 \tag{36}$$

for all $\epsilon > 0$. (36) only requires considering the distribution of $s(X,n)$ at each finite n, where standard results on measures on finite-dimensional product spaces give measurability and distribution results. It is not necessary to consider the space of sequences to make sense of this. It will be seen later that (36) is weaker than (35).

For a sequence $x = (x_1, x_2, \ldots) \in \Omega$, let $a_n(x) = (x_1, \ldots, x_n)$ be the first n components of x. For any $H \subset \Omega$, define

$$A_n(H) = \{x \in \Omega : a_n(x) = a_n(y) \text{ for some } y \in H\}.$$

Then $A_n(H)$ is the set of all points in Ω whose first n components agree with some element of H. There are 2^n distinct binary sequences $\{x_1, \ldots, x_n\}$. Let $K = K(H, n)$ be the number of distinct such sequences that occur in elements of H, $K \leq 2^n$. Then H can be divided into K distinct subsets H_k, with each $x \in H_k$ having the same values for its first n elements. If $x^{(k)}$ is an arbitrary element of H_k, then $A_n(\{x^{(k)}\}) = A_n(H_k)$, and $A_n(H) = \bigcup_k A_n(\{x^{(k)}\})$. Although $A_n(H)$ is a subset of the space of infinite sequences, only the first n components of a sequence need to be examined to determine whether a point lies in this set. Since $X_1, \ldots, X_n$ are iid Bernoulli, it thus is appropriate to define

$$P\{X \in A_n(\{x^{(k)}\})\} = \prod_{j=1}^{n} p^{x_j^{(k)}}(1-p)^{1-x_j^{(k)}} = p^{s(x^{(k)},n)}(1-p)^{n-s(x^{(k)},n)},$$

for each $x^{(k)}$, and since the sets $A_n(\{x^{(k)}\})$ are disjoint, set

$$P\{X \in A_n(H)\} = \sum_{k=1}^{K} p^{s(x^{(k)},n)}(1-p)^{n-s(x^{(k)},n)} \tag{37}$$

Note that $A_m(H) \subset A_n(H)$ for any $m \geq n$. Also, $A_m\{A_n(H)\} = A_n(H)$ for $n \leq m$, since $A_n(H)$ already contains all possible values for elements in positions m and higher.

Consider the collection $\mathcal{C}_0$ of all such sets $A_n(H)$ for all finite n and all $H \subset \Omega$. This collection is a field, since (1) it contains Ω and ϕ (consider $H = \Omega$ and $H = \phi$), (2) if $m \geq n$, then

$$A_n(H) \bigcup A_m(G) = A_m\{A_n(H)\} \bigcup A_m(G) = A_m\{A_n(H) \bigcup G\} \in \mathcal{C}_0,$$

so $\mathcal{C}_0$ is closed under finite unions, and (3) $A_n(H)^c = A_n\{A_n(H)^c\}$. (This last relationship just states that $A_n(H)^c$ is the set of all sequences whose first n elements do not match those of any element in H.)

If P as defined above is a probability measure on the field $\mathcal{C}_0$, then by the extension theorem (see Section **??**), P will have a unique extension to the σ-field generated by

$\mathcal{C}_0$. To see that P is a probability measure on $\mathcal{C}_0$, first note that from (37), $0 \leq P(X \in A) \leq 1$ for any $A \in \mathcal{C}_0$. To show finite additivity of P on $\mathcal{C}_0$, suppose $A_n(H)$ and $A_m(G)$ are disjoint, and $m \geq n$. Defining K and $x^{(1)}, \ldots, x^{(K)}$ as above, for the set $A_n(H) \bigcup A_m(G)$. Then

$$P[X \in A_m\{A_n(H) \bigcup G\}] = \sum_{k=1}^{K} p^{s(x^{(k)},m)}(1-p)^{m-s(x^{(k)},m)},$$

and

$$\begin{aligned}
P\{X \in A_n(H) \bigcup A_m(G)\} &= P[X \in A_m\{A_n(H) \bigcup G\}] \\
&= \sum_{k=1}^{K} p^{s(x^{(k)},m)}(1-p)^{m-s(x^{(k)},m)} \\
&= \sum_{x^{(k)} \in A_m\{A_n(H)\}} p^{s(x^{(k)},m)}(1-p)^{m-s(x^{(k)},m)} \\
&\quad + \sum_{x^{(k)} \in A_m(G)} p^{s(x^{(k)},m)}(1-p)^{m-s(x^{(k)},m)} \\
&= P[X \in A_m\{A_n(H)\}] + P\{X \in A_m(G)\} \\
&= P\{X \in A_n(H)\} + P\{X \in A_m(G)\},
\end{aligned}$$

so P is finitely additive. To show that P is countably additive, recall that this follows from finite additivity plus the continuity condition (42). Here (42) is easily established by showing that if $E_1 \supset E_2 \supset \cdots$, with $E_n = A_{m_n}(H_n)$ and $P(X \in E_n) \geq \delta > 0$ for all n, then $\bigcap E_n$ cannot be empty. First note that if $P(X \in E_n) \geq \delta$, then E_n cannot be empty, so by assumption, all of the E_n are non-empty. If all the E_n are non-empty, then their intersection cannot be empty, as given in the following lemma. Since $P(X \in E_n) \geq \delta$ implies $\bigcap E_n \neq \phi$, if $\bigcap E_n = \phi$, it must follow that $\lim_n P(X \in E_n) = 0$, so P is countably additive on $\mathcal{C}_0$.

Lemma. *If $E_1 \supset E_2 \supset \cdots$, with $E_n = A_{m_n}(H_n)$, and $E_n \neq \phi$ for all n, then $\bigcap_n E_n \neq \phi$.*

Proof: Since the E_n are non-empty, select an element $x^{(n)} = (x_1^{(n)}, x_2^{(n)}, \ldots)$ from E_n for each n. In the sequence $x_1^{(1)}, x_1^{(2)}, \ldots$, there must be a value (0 or 1) that occurs infinitely often. Let this value be u_1, and let $l_{11} < l_{12} < \cdots$ be all values of n where $x_1^{(n)} = u_1$. Now consider $x_2^{(l_{12})}, x_2^{(l_{13})}, \ldots$. Again, there must a value that occurs infinitely often in this sequence. Let u_2 equal this value, and let $l_{21} < l_{22} < \cdots$ be

the subset of $l_{12}, l_{13}, \ldots$ where $x_2^{(l_{1j})} = u_2$. Continue in this fashion, obtaining a sequence $u_1, u_2, \ldots$, and an increasing sequence of values $l_{11}, l_{21}, \ldots$ such that $x_j^{(l_{n1})} = u_j$, $j = 1, \ldots, n$. Set $x_0 = (u_1, u_2, \ldots)$, and consider an arbitrary $E_k = A_m(H)$ for some m and H. Choose an $l^* \in \{l_{j1} : j \geq m\}$ with $l^* \geq k$. Then there is a $w \in E_{l^*} \subset E_k$, with $w_j = u_j$ for $j \leq m$. Since E_k includes all possible extensions of the first m elements of the sequences in H, it follows that $x_0 \in E_k$. Since k was arbitrary, $x_0 \in E_n$ for all n, and thus $x_0 \in \bigcap_n E_n$, so $\bigcap_n E_n \neq \phi$. □

Since P is a probability measure on $\mathcal{C}_0$, it therefore has a unique extension to $\sigma(\mathcal{C}_0)$. By the continuity properties of measures on σ-fields, this extension must satisfy

$$P(X \in B) = \lim_{n \to \infty} P\{X \in A_n(B)\}$$

for any $B \in \sigma(\mathcal{C}_0)$. Clearly for this distribution, $P(X = x_0) = 0$ for any $x_0 \in \Omega$ (if $0 < p < 1$).

This development thus leads to a well-defined collection of measurable sets and a probability measure on the space of sequences of iid Bernoulli random variables.

Now consider the measurability of the set in (35). Define $B(n, \delta) = \{x \in \Omega : |s(X, n)/n - p| < \delta\}$. Since $B(n, \delta) \in \mathcal{C}_0$ for any n and δ, and since $s(X, n)/n \to p$ if and only if for each l there is an $m(l)$ such that $|s(X, n)/n - p| \leq 1/l$ for $n \geq m(l)$, clearly

$$\{\lim s(X, n)/n = p\} = \bigcap_l \liminf_n B(n, 1/l) = \bigcap_l \bigcup_m \bigcap_{n \geq m} B(n, 1/l) \in \sigma(\mathcal{C}_0). \tag{38}$$

Exercise 11.1 Show that $\{x \in \Omega : \lim n^{-1/2} |s(X, n) - np| > \alpha\} \in \sigma(\mathcal{C}_0)$.

Note that (36) is equivalent to

$$\lim_{\epsilon \downarrow 0} \lim_{m \to \infty} P\{B(m, \epsilon)\} = 1.$$

Using (4) and the representation (38), it follows that

$$P\{\lim s(X, m)/m = p\} = \lim_{\epsilon \downarrow 0} \lim_{m \to \infty} P\left(\bigcap_{k \geq m} B(k, \epsilon) \right).$$

Since $B(m, \epsilon) \supset \bigcap_{k \geq m} B(k, \epsilon)$ for each m, it follows that if $P\{\lim s(X, m)/m = p\} = 1$, then $\lim P(|s(X, m)/m - p| < \epsilon) = 1$ as well. This is a

special case of the result that convergence almost everywhere implies convergence in probability.

A development similar to that above for iid Bernoulli variables can be given fairly generally for probability measure spaces for sequences of independent random variables. Suppose that the X_j, $j = 1, 2, \ldots$ are independent, with $P(X_j \in B) = P_j(B)$ defined for all Borel sets B. Let Ω be the set of all sequences $(x_1, x_2, \ldots)$, $x_j \in R^1$. For Borel sets $B_1, \ldots, B_n$, define

$$A_n(B_1, \ldots, B_n) = \{x \in \Omega : x_j \in B_j, j \leq n; x_j \in R^1, j > n\}.$$

The collection of all such sets for all n and all Borel sets B_j is easily shown to be a field $\mathcal{F}$. On this field, the function

$$P\{A_n(B_1, \ldots, B_n)\} = \prod_{j=1}^{n} P_j(B_j) \tag{39}$$

is probability measure. Thus by the extension theorem, P has a unique extension to $\sigma(\mathcal{F})$, which then defines a probability measure on a σ-field of subsets of Ω. Details of this argument are given in Chung (1974), Theorem 3.3.4.

It also is not necessary to limit attention to sequences of independent random variables. For example, if $s(X, n)$ is defined as before as the sum of the first n terms of X, then $(Y_1, Y_2, \ldots)$, with $Y_n = s(X, n)/n$, has a joint distribution that can be obtained from the distribution of X. A general treatment in the dependent case is similar to that above, with the product in (39) replaced by an appropriate joint probability $P\{(X_1, \ldots, X_n) \in B_1 \times \cdots \times B_n\}$.

In the following, the space of real valued sequences will be denoted by R^∞, and the σ-field of Borel sets as defined above by $\mathcal{B}_\infty$. When equipped with a probability measure P, the resulting probability measure space is denoted

$$(R^\infty, \mathcal{B}_\infty, P). \tag{40}$$

12 Some Useful Theorems

The following results, which are often cited in the probability literature, are summarized here for easy reference.

Continuity Property of Probability Measures. If $(\Omega, \mathcal{F}, P)$ is a probability measure space, and $A_1 \subset A_2, \subset \cdots$ is an increasing sequence of sets in $\mathcal{F}$, and $B_1 \supset B_2, \supset \cdots$ is a decreasing sequence of sets in $\mathcal{F}$, then

$$P(A_n) \to P\left(\bigcup_n A_n\right) \quad \text{and} \quad P(B_n) \to P\left(\bigcap_n B_n\right). \tag{41}$$

Note that $\bigcup_n A_n$ in a sense is the limit of the sequence of sets A_n, and $\bigcap_n B_n$ can similarly be thought of as the limit of B_n.

Suppose P is a *finitely* additive probability measure defined on $(\Omega, \mathcal{F})$. If

$$P(E_n) \to 0 \tag{42}$$

for any decreasing sequence $E_1 \supset E_2 \supset \cdots$ of sets in $\mathcal{F}$ such that $\bigcap E_n = \phi$, then P is countably additive; see eg Chung (1974), Theorem 2.2.1, or Billingsley (1995), Example 2.10. Note that in R^1, the sets $E_n = \{x : 0 < x < 1/n\}$, for example, are a decreasing sequence with $\bigcap_n E_n = \phi$.

Extension Theorem. Let $\mathcal{F}_0$ be a field of subsets of a set Ω, and P a probability measure defined on $\mathcal{F}_0$. Then P has a unique extension to $\mathcal{F} = \sigma(\mathcal{F}_0)$, the σ-field generated by $\mathcal{F}_0$. (Billingsley, 1995, Theorem 3.1)

Dominated Convergence Theorem. For measurable functions, if $|f_n| \leq g$ almost everywhere (ae), where $\int g\, dP < \infty$, and if $f_n \to f$ ae, then

$$\lim \int f_n\, dP = \int f\, dP. \tag{43}$$

In terms of random variables, if $X_1, X_2, \ldots$ are defined on a common probability space and $X_n \to X$ ae, and if $|X_n| \leq Y$ ae for some Y with $\mathrm{E}(Y) < \infty$, then

$$\mathrm{E}(X_n) \to \mathrm{E}(X).$$

Note that the condition $|X_n| \leq Y$ ae can also be expressed $P(|X_n| < Y) = 1$.

On probability spaces, sequences of uniformly bounded random variables are always dominated. That is, if there is an $M < \infty$ such that $P(|X_n| < M) = 1$ for all n, then defining Y by $P(Y = M) = 1$, it follows that $P(|X_n| < Y) = 1$ and $\mathrm{E}(Y) = M < \infty$.

Thus if $X_1, X_2, \ldots$ are uniformly bounded, and $X_n \to X$ ae, then $\mathrm{E}(X_n) \to \mathrm{E}(X)$. In particular, if $X_1, X_2, \ldots$ are uniformly bounded, and $X_n \to 0$ ae, then

$$\mathrm{E}(X_n) \to 0. \tag{44}$$

The following two results are used frequently.

Theorem 12.1 *Suppose $\int |X|\, dP < \infty$ and $a_1 < a_2 < \cdots$ is a sequence of constants with $a_n \to \infty$. Then*

$$\int_{|X|>a_n} |X|\, dP \to 0.$$

Proof: Since X is integrable,

$$\infty > \int |X|\, dP \geq \int_{|X|>a_n} |X|\, dP \geq a_n P(|X| > a_n),$$

so $P(|X| > a_n) \leq \int |X|\, dP/a_n \to 0$. Define $I(|X| > a_n) = 1$ if $|X| > a_n$ and $I(|X| > a_n) = 0$ otherwise. Now

$$P\{\limsup_{m\to\infty} |X| I(|X| > a_m) > 0\} \leq P(|X| > a_n) \to 0,$$

so $|X|I(|X| > a_n) \to 0$ ae. Since $|X|I(|X| > a_n) \leq |X|$, and $|X|$ is integrable, the result follows from the dominated convergence theorem. □

Theorem 12.2 *Suppose $\int |X|\, dP < \infty$ and $A_1, A_2, \ldots$ is a sequence of sets with $P(A_n) \to 0$. Then*

$$\int_{A_n} X\, dP \to 0. \tag{45}$$

Proof: For any $\alpha > 0$,

$$\left| \int_{A_n} X\, dP \right| \leq \alpha P(A_n) + \int_{|X|>\alpha} |X|\, dP. \tag{46}$$

By the previous theorem, $\lim_{\alpha\to\infty} \int_{|X|>\alpha} |X|\, dP = 0$, so given any $\epsilon > 0$, there is an α such that $\int_{|X|>\alpha} |X|\, dP < \epsilon/2$. Since $P(A_n) \to 0$, there is then a value N such

that $P(A_n) < \epsilon/(2\alpha)$ for $n > N$. Thus from (46), given any $\epsilon > 0$, there is an N such that $|\int_{A_n} X\,dP| < \alpha\epsilon/(2\alpha) + \epsilon/2 = \epsilon$. Hence, $\int_{A_n} X\,dP \to 0$ as $n \to \infty$. □

Integration of Series. For measurable functions,
(a) if $f_n \geq 0$, then

$$\int \sum_n f_n\,dP = \sum_n \int f_n\,dP, \tag{47}$$

where both sides may be infinite;
(b) if $\sum_n f_n$ converges ae, and $|\sum_{k=1}^n f_n| \leq g$ ae, where $\int g\,dP < \infty$, then $\sum_n f_n$ and the f_n are integrable, and $\int \sum_n f_n\,dP = \sum_n \int f_n\,dP$;
(c) if $\sum_n \int |f_n|\,dP < \infty$, then $h = \sum_n f_n$ is absolutely convergent ae, is integrable, and

$$\int h\,dP = \sum_n \int f_n\,dP. \tag{48}$$

(see Billingsley, 1995, page 211)

The following result is useful for investigating convergence of infinite series.

Theorem 12.3 *Suppose $f(t)$ is a decreasing function. Then*

$$\int_k^{n+1} f(t)\,dt \leq \sum_{j=k}^n f(j) \leq \int_{k-1}^n f(t)\,dt.$$

Proof:

$$\int_{k-1}^n f(t)\,dt = \sum_{j=k}^n \int_{j-1}^j f(t)\,dt \geq \sum_{j=k}^n \int_{j-1}^j f(j)\,dt = \sum_{j=k}^n f(j)$$

$$= \sum_{j=k}^n \int_j^{j+1} f(j)\,dt \geq \sum_{j=k}^n \int_j^{j+1} f(t)\,dt = \int_k^{n+1} f(t)\,dt.$$

□

Independent Sets. Let $\mathcal{A}_1, \ldots, \mathcal{A}_n$ be collections of measurable subsets of Ω. The collections are independent if $P(\bigcap_{j=1}^n A_j) = \prod_j P(A_j)$ for any sets $A_j \in \mathcal{A}_j$, $j = 1, \ldots, n$.

Theorem 12.4 *Suppose $\mathcal{A}_1, \ldots, \mathcal{A}_n$ are independent, and each $\mathcal{A}_j$ is closed under finite intersections. Then the generated σ-fields $\sigma(\mathcal{A}_1), \ldots, \sigma(\mathcal{A}_n)$ are independent.*

(see Billingsley, 1995, Theorem 4.2). "Closed under finite intersections" means that if $A, B \in \mathcal{A}_j$, then $A \bigcap B \in \mathcal{A}_j$.