Measure Theory and the Central Limit Theorem

Peter Brown

August 24, 2011

Abstract

It has long been known that the standard Riemann integral, encountered first as one of the most important concepts of calculus, has many shortcomings, and can only be defined over the limited class of sets which have boundaries of Jordan content zero. In this paper I present the measure theory necessary to develop the more robust Lebesgue integral. This is interesting in its own right, and although we dwell little on the consequences of our definition, any calculus student who is familiar with the properties of the Riemann integral will recognize its tremendous advantages. But we will push further: measure theory and the Lebesgue integral provide us with the ideal background to develop a rigorous theory of probability. To this end, we introduce random variables and develop the theory of distribution functions. This paper culminates in the proof of the Central Limit Theorem (CLT), which explains the ubiquitous nature of the normal distribution. This is a deep and fascinating result, but relatively straightforward once one has provided the correct machinery. Indeed, we do most of the hard work studying measure theory. This paper does not provide the most intuitive approach towards the CLT, but it has the advantage of providing an introduction to rigorous probability that may serve as a starting point for future study. We have tried to provide the most expedient treatment of the material possible without sacrificing the intuition or the depth of the theoretical background.

1 Introduction

This paper is designed to provide an exposition of basic measure theory and the Lebesgue integral in preparation for putting probability theory on a rigorous foundation and proving the Central Limit Theorem. It is primarily intended for the reader who has never been exposed to these concepts before, and as such we spend much time giving definitions and proving basic facts. Results such as the Fubini Theorem are proved in great detail so that readers new to this topic may see how the large body of machinery we have developed works in practice. Indeed, all the material we present is necessary to understand the proof of the Central Limit Theorem, which is the final goal of this paper. It is our hope that the reader then recognizes that the primary connection between the first several sections of the paper and the last is not merely philosophical: we need measure

theory to formalize distribution functions, and we need integration theory to formalize characteristic functions.
The first section attempts to motivate the study of the objects of measure theory, particularly σ-algebras, measurable functions, and measures. The second defines the Lebesgue integral and provides proofs of some of its most important properties. We prove the Lebesgue Monotone Convergence Theorem and the Dominated Convergence Theorem. Much of the exposition of this section and some additional remarks are devoted to comparing the Riemann and Lebesgue integrals. The third section is devoted entirely to proving the Fubini Theorem. This may appear to be a strange choice. The proof of Fubini is long and technical–much more so than its equivalent in a calculus course. However, there are important reasons for presenting this material. Reading and understanding the proofs of this section require a strong understanding of the material in the previous sections and, as such, provides the reader with a barometer of his or her understanding of the topic. Further, the Fubini Theorem is a highly important and useful computational result which will frequently be of use to us. The fourth section is devoted to defining the basic functional tools of probability theory. We introduce random variables, expectations, distribution functions, and characteristic functions and prove numerous basic results about each. While the vocabulary is familiar to anyone who has had an introduction to probability, the section is entirely measure-theoretic. Section five proves the Lévy Continuity Theorem, which is the tool needed to prove the Central Limit Theorem, and the final section is devoted to the Central Limit Theorem itself.

2 Measure Theory

Let X be a nonempty set. How are we to measure the *size* of a subset $S \subset X$? Naïvely, one might count the number of elements of S. Alternatively, if X is endowed with a metric, we could try to define the size of a set in terms of the set's diameter. These methods are not particularly subtle, and neither is capable of, for example, determining the dimension of a set. The easiest conclusion is that our question is ill-posed: one needs to first determine what properties the notion of *size* should have. Ideally, we would have some function mapping from $\mathcal{P}(\mathbb{X}) \to [0, \infty]$, which takes a subset of X to a nonnegative number (or infinity). It seems reasonable to demand that this function be additive across finitely many disjoint sets, and assign 0 to the empty set, and after that we might work out additional properties from a more detailed examination of X; for example, in the case of $\mathbb{R}$ we might decide that every nonempty open interval should have positive size. Hence, the function that assigns to an interval its *length* seems a good starting point for our intuition. There is at least one major difficulty with this process. In many cases, including the example above, it is not possible to define a function on the entire power set of X which has all the properties we might want. This is not obvious a priori, and we will not have the time to delve into this topic deeply. Instead, we begin artificially by defining the sorts of subcollections of $\mathcal{P}(X)$ on which more restricted measures of size make sense. Philosophical questions

focusing on why we choose these particular axioms may be raised in the context of probability theory (particularly in generalizing probability theory to certain aspects of Quantum Mechanics — presently an active field of research), but we will not concern ourselves with these either. We begin with a definition.

Definition 2.1. An *algebra* of sets is a collection $\mathcal{F}$ of subsets of some set X, i.e. $\mathcal{F} \subset \mathcal{P}(X)$, such that the following conditions are satisfied:

i. $X, \emptyset \in \mathcal{F}$.

ii. $\forall\ E \in \mathcal{F}, E^c = X \setminus E \in \mathcal{F}$.

iii. $\forall\ \{E_i\}_{i=1}^N \subset \mathcal{F}$, we have $\bigcup_{i=1}^N E_i \in \mathcal{F}$.

We say that $\mathcal{X}$ is a *σ-algebra* if it additionally satisfies:

iv. $\forall\ \{E_i\}_{i\in\mathbb{N}} \subset \mathcal{F}$, we have $\bigcup_{i=1}^\infty \in \mathcal{F}$.

Elements of $\mathcal{F}$ are called *measurable*.

In other words, we say that $\mathcal{F}$ is a σ-algebra if it is an algebra which is closed under countable unions. The axioms for a σ-algebra are relatively strong. Any σ-algebra also satisfies the axioms for a topology. This is an apt comparison because a topology also gives us a sense of size of sets. Intuitively, an open set is big, a closed set is small, and a compact set is a generalization of a finite set. Usually a σ-algebra is too large to be interesting as a topology, but sometimes it will be useful to contrast these two structures.

Remark 2.2. We call $\mathcal{F}$ an algebra because $(\mathcal{F}, \triangle, \cap, \emptyset, X)$ is an algebra over the field with two elements, where $\triangle$ is symmetric difference and $\emptyset$ and X are respectively the additive and multiplicative units of the algebra. Actually, it is also a field. Some texts, for example [3], refer to algebras of sets as fields and σ-algebras of sets as σ-fields, and this terminology is common in probability.

A σ-algebra is also closed under countable intersections. If, again, $\{E_i\}_{i\in\mathbb{N}}$ is a collection of subsets, then $\bigcap_i E_i = \bigcap_i (E_i^c)^c = (\bigcup_i E_i^c)^c$ by DeMorgan's Law, and this last is contained in the σ-algebra since the σ-algebra is closed under complementation and countable union. Furthermore, any σ-algebra $\mathcal{F}$ on X induces a σ-algebra on $E \subset X$. In fact, it is easy to see that $\mathcal{E} = \mathcal{P}(E) \cap \mathcal{F}$ satisfies the axioms for a σ-algebra on E. (i.) and (iv.) follow trivially, and we need only check that for all $A \in \mathcal{E}$ we have $E \setminus A \in \mathcal{E}$, where $E \setminus A = E \cap A^c$. Since $E, A \in \mathcal{F}$ we have $E \setminus A = E \cap A^c \in \mathcal{F}$ by the previous remark. Taking this to its logical conclusion, it is easy to verify that the intersection of two σ-algebras is again a σ-algebra. Hence, as we do in many other cases in mathematics, we can define the σ-algebra generated by some collection of sets $\mathcal{C}$ by $\mathcal{A}_\mathcal{C} = \bigcap_{i\in\mathcal{I}} \mathcal{A}_i$ where $\{\mathcal{A}_i\}$ is the collection of all σ-algebras containing $\mathcal{C}$.

Definition 2.3. We call a pair $(X, \mathcal{F})$, where X is a nonempty set and $\mathcal{F}$ is a σ-algebra on X, a *measurable space*.

In the context of topology, one defines a *continuous* function by requiring that a continuous map $f : X \to Y$ induce a map $f^{-1} : \mathcal{T}(Y) \to \mathcal{T}(X)$. As we have noted, the concept of a measurable space is similar to that of a topological space. Hence it seems reasonable to study functions that preserve the same type of structure. We will start by examining real-valued functions, where the relevant σ-algebra on $\mathbb{R}$ is the so-called *Borel algebra* generated by the euclidean topology on $\mathbb{R}$.

Definition 2.4. We say that a function $f : (X, \mathcal{F}) \to \mathbb{R}$ is *measurable* if for all $\alpha \in \mathbb{R}$, we have $f^{-1}[(\alpha, \infty)] \in \mathcal{F}$. We denote the space of measurable real-valued functions on X by $\mathcal{M}(X, \mathcal{F})$, and the space of non-negative measurable real-valued functions on X by $\mathcal{M}(X, \mathcal{F})^+$.

This is the sort of measurable function we will be most concerned with, but it is not the most general notion we could have given. If Y is another measurable space, or more generally a topological space, we might say that a function $f : X \to Y$ is *measurable* if for every measurable, respectively open set U in Y, that $f^{-1}(U)$ belongs to $\mathcal{F}$. It is also useful to know if measurability is preserved under certain operations, such as composition, addition, and so on. The following proposition answers these questions. We use the more general definition of a measurable function because this serves to simplify our notation.

Proposition 2.5. *Let $(X, \mathcal{F})$ be a measurable space.*

i. If Y and Z are topological spaces, $f : X \to Y$ is measurable, and $g : Y \to Z$ is continuous, then $h := g \circ f$ is measurable.

ii. If Y is a topological space, $u, v : X \to \mathbb{R}$ are measurable functions, and $\Phi : \mathbb{R}^2 \to Y$ is continuous, then $h(x) = \Phi(u(x), v(x))$ is measurable.

iii. If $f : X \to \mathbb{R}$, $g : X \to \mathbb{R}$ are measurable and $\eta \in \mathbb{R}$ then $f + g$, fg, and ηf are also measurable functions.

Proof. To prove i., let $U \subset Z$ be an open set. Then

$$h^{-1}(U) = (f \circ g)^{-1}(U) = f^{-1}(g^{-1}(U)).$$

$V = g^{-1}(U)$ is open by continuity, so $f^{-1}(V)$ is measurable.

To prove ii., it suffices to show that $f(x) = (u(x), v(x))$ is measurable by i. Note that all open sets in $\mathbb{R}^2$ are unions of open rectangles, hence it suffices to show that for any open rectangles $(a, b) \times (c, d)$, $f^{-1}[(a, b) \times (c, d)]$ is a measurable set. But $f^{-1}[(a, b) \times (c, d)] = u^{-1}[(a, b)] \cap v^{-1}[(c, d)]$ which is measurable.

Finally, since $\Phi(x, y) = x + y$, $\Phi(x, y) = xy$, and $\Phi(x, y) = \eta x$ are continuous functions, iii. follows from ii. □

We will also frequently be interested in considering the limit of a sequence of functions, or the *limit supremum* and *limit infimum*. We briefly recall these

notions for general sequences. One way to think of the limit infimum and the limit supremum of a sequence a_n is as the infimum and supremum respectively of the set $S = \{x : x \text{ is an accumulation point of } \{a_n\}\}$. However, it can be hard to work directly from this definition, so we give an equivalent one.

Definition 2.6. The *limit infimum* and *limit supremum* of a sequence $\{a_n\}$, denoted $\liminf a_n$ and $\limsup a_n$, are defined by

$$\liminf a_n = \sup_n \inf_{n \leq k} a_k$$

and

$$\limsup a_n = \inf_n \sup_{n \leq k} a_k.$$

If a sequence has a limit, then $\lim_n a_n = \limsup a_n = \liminf a_n$. We leave this, and the equivalence of the two definitions given, as exercises for the reader.

Proposition 2.7. *Let $\{f_n\}$ be a sequence of measurable functions. Then $\phi(x) = \inf_n f_n(x)$, $\Phi(x) = \sup_n f_n(x)$, $\underline{f}(x) = \liminf f_n(x)$, and $\overline{f}(x) = \limsup f_n(x)$ are measurable functions.*

Proof. First, we claim that $\Phi^{-1}[(\alpha, \infty)] = \bigcup_n f_n^{-1}[(\alpha, \infty)]$. Clearly we must have $\bigcup_n f_n^{-1}[(\alpha, \infty)] \subset \Phi^{-1}[(\alpha, \infty)]$. To see the other inclusion, assume we have x such that $\Phi(x) \in (\alpha, \infty)$ but $f_n(x) \leq \alpha$ for all $n \in \mathbb{N}$. Then clearly $\Phi(x) = \sup_n f_n(x) \leq \alpha$, so that $\Phi(x) \notin (\alpha, \infty)$ and the two sets are equal. Note each element in the union $\bigcup_n f_n^{-1}[(\alpha, \infty)]$ is measurable, and the union is countable; hence it is a measurable set by the definition of σ-algebra. Thus Φ is a measurable function. The argument for ϕ is analogous. Next, recall that we can write $\overline{f}(x) = \inf_n \sup_{n \leq k} f_k(x)$ and $\underline{f}(x) = \sup_n \inf_{n \leq k} f_k(x)$. These are then compositions of the functions Φ and ϕ with countably many members of the sequence $\{f_n\}$, hence also measurable. □

Up until now we have been describing collections of measurable sets and functions, and particularly what combinations of these objects we would hope remain measurable. We have yet to actually define a function that *measures* them. As noted in the exposition, we expect that this function will have certain nice properties. For example, given two disjoint sets A and B, the measure of $A \cup B$ ought to be the measure of A plus the measure of B. (In later sections, when we turn our focus to probability, the property we have just described is very natural. If we identify measurable sets with possible events and the function which measures them with their probability of occurring, this property implies that if A has probability $\mathbb{P}_1$ of occurring and B has probability $\mathbb{P}_2$ and the two events are mutually exclusive, then $A \cup B$ has probability $\mathbb{P}_1 + \mathbb{P}_2$ of occurring). The following definitions clarify this intuition.

Definition 2.8. A *measure space* is a triple $(X, \mathcal{F}, \mu)$ where $(X, \mathcal{F})$ is a measurable space and $\mu : \mathcal{F} \to [-\infty, \infty]$ is a function satisfying the following:

i. $\mu(\emptyset) = 0$.

ii. $\forall\ E \subset X, \mu(E) \geq 0$.

iii. If $\{A_n\}$ is a collection of pairwise disjoint sets, then $\mu(\bigcup_n A_n) = \sum_n \mu(A_n)$.

We call the size function μ a *measure*. If μ has properties i. and iii. but not ii. we call μ a *signed measure*. If μ is a measure that additionally satisfies

iv. $\mu(X) = 1$,

then μ is referred to as a *probability measure* and we refer to $\mu(A)$ as the *probability that A occurs*. In this case we refer to the elements of $\mathcal{F}$ as *events*, and to X as a *probability space*.

Probability measures are special cases of measures where we restrict the values of $\mu(E)$ to be finite; these functions are called *finite measures*. It is an easy corollary of the next proposition that a measure μ is a finite measure if and only if $\mu(X) < \infty$. If there exists a countable covering of our space X of sets in $\mathcal{F}$ such that all sets in the covering have finite measure, then we say μ is *σ-finite*. Our next proposition outlines some basic properties of measure.

Proposition 2.9. *Let $(X, \mathcal{F}, \mu)$ be a measure space. Let $A, B, \{E_n\} \subset \mathcal{F}$; then the following properties hold:*

i. if $A \subset B$ then $\mu(A) \leq \mu(B)$.

ii. if $A \subset B$ and $\mu(A) \neq +\infty$ then $\mu(B \setminus A) = \mu(B) - \mu(A)$.

iii. if $\{E_n\}$ is a sequence of increasing sets, then $\mu(\bigcup_n E_n) = \lim_n \mu(E_n)$.

iv. if $\mu(E_1) \neq +\infty$ and $\{E_n\}$ is decreasing, then $\mu(\bigcap_n E_n) = \lim_n \mu(E_n)$.

Proof. First note that finite additivity of μ follows from countable additivity (property iii. of the definition of measure). Now, if $A \subset B$ the sets $A, B \setminus A$ are disjoint. Since the measure of any set in $\mathcal{F}$ is nonnegative we have

$$\mu(A) \leq \mu(A) + \mu(B \setminus A) = \mu(A \cup (B \setminus A)) = \mu(B);$$

this proves i. Furthermore, ii. follows from the second to last equality.

Let $\{E_n\}$ be an increasing sequence; that is, $E_n \subset E_{n+1}$ for all n. Note that $\{E_{n+1} \setminus E_n\}$ is a sequence of disjoint sets, where we conventionally set $E_1 = \emptyset$. Note also that $\bigcup_n E_n = \bigcup_n E_{n+1} \setminus E_n$. Thus, by countable additivity, $\mu(\bigcup_n E_n) = \mu(\bigcup_n E_{n+1} \setminus E_n) = \sum_n \mu(E_{n+1} \setminus E_n) = \sum_n \mu(E_{n+1}) - \mu(E_n)$ by ii., but this is a telescoping series. Since

$$\mu(\bigcup_{n=1}^{N} E_n) = \sum_{n=1}^{N} [\mu(E_{n+1}) - \mu(E_n)] = \mu(E_N),$$

we have $\lim_N \mu(E_N) = \lim_N \mu(\bigcup_{n=1}^N E_n) = \mu(\bigcup_n E_n)$. Now suppose $\{E_n\}$ is a decreasing sequence. Define $A_n = E_1 \setminus E_n$ for each n. Note that $\{A_n\}$ is an increasing sequence; then

$$\mu(\bigcup_n A_n) = \mu(\bigcup_n E_1 \setminus E_n) = \mu(E_1 \setminus \bigcap_n E_n) = \mu(E_1) - \mu(\bigcap_n E_n),$$

and also

$$\mu(\bigcup_n A_n) = \lim_n \mu(A_n) = \lim_n \mu(E_1 \setminus E_n) = \lim_n \mu(E_1) - \mu(E_n) = \mu(E_1) - \lim_n \mu(E_n)$$

by (iii). Equating these expressions and subtracting $\mu(E_1)$ gives iv. □

Once we have chosen a σ-algebra $\mathcal{F}$ and a measure μ over a space X we can begin to phrase results in terms of the size of subsets of X. Often sets of larger size will be more important than sets of smaller size. For example, sets of measure zero — that is, sets $N \in \mathcal{F}$ such that $\mu(N) = 0$ — are in some sense negligible. In fact, in many cases we may find that a statement holds everywhere except over such a set of measure 0. If this is the case, we say this property holds *almost everywhere*; if X is a probability space we say the property occurs *almost surely*. Many theorems require only that we assume our hypotheses hold almost everywhere. Conversely, if we are working in a finite measure space we regard sets of full measure as extremely important.

3 Integration

The integral is a natural generalization of the concept of a sum. Generally one's first introduction to this generalization — which extends the concept of summation to the continuum — is the fairly natural Riemann integral. If f is a sufficiently nice function defined on some interval $[a, b]$, a Riemann sum consists of a finite set of terms of the form $f(x_i^*)(x_{i+1} - x_i)$, where the x_j are an ordered partition of $[a, b]$ and $x_i^* \in [x_i, x_{i+1}]$. Each term approximates some of the area under the curve of f. Recalling our discussion of what properties a "size function" should have, we note that one reasonable way to define the size of an open interval of $\mathbb{R}$ is as its length; indeed, if one sets $\mu([a, b)) = b - a$, there is a canonical extension of μ from the set of all finite unions and intersections of disjoint half-open intervals to the Borel algebra on $\mathbb{R}$. The resulting measure is called *Lebesgue measure*. The term $(x_{i+1} - x_i)$ above is exactly the Lebesgue measure of the interval $[x_i, x_{i+1})$, and the Riemann sum does not use any of the field properties of the real numbers. Hence we might hope to rephrase the definition of integral to hold for real-valued functions on an arbitrary measure space. We will find that this can be done, and that the resulting object on $\mathbb{R}$, called the *Lebesgue integral*, is more robust than the Riemann integral. It behaves better with respect to limits of sequences of functions and extends the range of the standard Riemann integral significantly. However, these serendipitous consequences are not obvious a priori.

Definition 3.1. Let $(X, \mathcal{F}, \mu)$ be a measure space. A *simple function* is a measurable function $\varphi : X \to \mathbb{R}$ which takes only a finite number of distinct values. That is, φ is a simple function if $A = \{a : \varphi(x) = a \text{ for some } x \in X\}$ is a finite set.

We can always represent a simple function in the form $\varphi = \sum_{i=1}^{n} a_i \mathbb{1}_{E_i}$ where the a_i are constants, the E_i are measurable sets, and $\mathbb{1}_B$ is the *indicator function* of the set B. If we order A as a subset of $\mathbb{R}$, we might set $E_i = \varphi^{-1}(\{a_i\})$ where $A = \{a_i\}_{i=1}^{n}$, and this gives the canonical representation for φ. We will almost always work with this representation, but none of our results depend on this choice. The reader may check that the following definition is independent of the representation chosen for φ.

Definition 3.2. If $\varphi = \sum_{i=1}^{n} a_i \mathbb{1}_{E_i}$ is a nonnegative simple function, then

$$\int \varphi \, \mathrm{d}\mu = \sum_{i=1}^{n} a_i \mu(E_i)$$

is called the *integral of φ with respect to μ.*

The next proposition shows that this "new" integral is linear and shows how we can produce a new measure by taking the integral of a set E as its measure.

Proposition 3.3. *Let $(X, \mathcal{F}, \mu)$ be a measure space. If φ and ψ are nonnegative simple functions and $\eta \in \mathbb{R}^+$ then the following hold:*

i. $\varphi + \psi$ *and* $\eta\phi$ *are nonnegative simple functions.*

ii. $\displaystyle\int \eta\varphi \, \mathrm{d}\mu = \eta \int \varphi \, \mathrm{d}\mu$

iii. $\displaystyle\int \varphi + \psi \, \mathrm{d}\mu = \int \varphi \, \mathrm{d}\mu + \int \psi \, \mathrm{d}\mu$

iv. $\displaystyle\int \varphi\chi_E \, \mathrm{d}\mu = \lambda(E)$ *defines a measure on* $(X, \mathcal{F})$.

Proof. It is clear that $\varphi + \psi$ and $\eta\varphi$ are nonnegative simple functions. Proposition 2.5 tells us they are also measurable. Let $\varphi = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i}$ be the canonical representation for φ; then by definition,

$$\int \eta\varphi \, \mathrm{d}\mu = \sum_{i=1}^{n} \eta a_i \mu(A_i) = \eta \sum_{i=1}^{n} a_i \mu(A_i) = \eta \int \varphi \, \mathrm{d}\mu.$$

It is not difficult to check that if $\psi = \sum_{j=1}^{m} b_j \mathbb{1}_{B_j}$ is the canonical representation for ψ, we then have

$$\varphi + \psi = \sum_{i=1}^{n} \sum_{j=1}^{m} (a_i + b_j) \mathbb{1}_{A_i \cap B_j}.$$

Further, note that the A_i's are pairwise disjoint as are the B_j's since we are dealing with the canonical representation. Then

$$\begin{aligned}
\int \varphi + \psi \, \mathrm{d}\mu &= \sum_{i=1}^{n} \sum_{j=1}^{m} (a_i + b_j) \mu(A_i \cap B_j) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} a_i \mu(A_i \cap B_j) + \sum_{i=1}^{n} \sum_{j=1}^{m} b_j \mu(A_i \cap B_j) \\
&= \sum_{i=1}^{n} a_i \sum_{j=1}^{m} \mu(A_i \cap B_j) + \sum_{j=1}^{m} b_j \sum_{i=1}^{n} \mu(A_i \cap B_j) \\
&= \sum_{i=1}^{n} a_i \mu(A_i) + \sum_{j=1}^{m} b_j \mu(B_j) \\
&= \int \varphi \, \mathrm{d}\mu + \int \psi \, \mathrm{d}\mu,
\end{aligned}$$

where we have used only basic properties of measures and finite sums. That $\lambda(\emptyset) = 0$ and $\lambda(E) \geq 0$ follow from the definition of the integral of a simple function. It remains to show countable additivity. Suppose E_i is a countable family of pairwise disjoint sets. Then

$$\int_{\bigcup_i E_i} \varphi \, \mathrm{d}\mu = \int \varphi \mathbb{1}_{\bigcup_i E_i} \, \mathrm{d}\mu$$

If we allow

$$\varphi = \sum_{j=1}^{n} \alpha_j \mathbb{1}_{A_j}$$

we have

$$\int \varphi \mathbb{1}_{\bigcup_i E_i} \, \mathrm{d}\mu = \sum_{j=1}^{n} \alpha_j \mu(A_i \cap \bigcup_i E_i) = \sum_{j=1}^{n} \alpha_j \mu(\bigcup_i A_j \cap E_i) = \sum_{j=1}^{n} \alpha_j \sum_i \mu(A_j \cap E_i)$$

Where we have applied the countable additivity of μ. Note that the first sum is finite so we can certainly change the order of summation. Then we have

$$\sum_i \sum_j \alpha_j \mu(A_j \cap E_i) = \sum_i \int_{E_i} \varphi \, \mathrm{d}\mu$$

□

Of course in some sense this result is not very interesting because we have not yet extended the integral to general measurable functions. Let us first define the more general integral for positive functions. We will then extend the integral to any measurable function by making use of the fact that $f = f^+ - f^-$ where f^+ and f^- are defined by

$$f^+(x) = \begin{cases} f(x) & \text{if } f(x) > 0, \\ 0 & \text{otherwise;} \end{cases} \qquad f^-(x) = \begin{cases} -f(x) & \text{if } f(x) < 0, \\ 0 & \text{otherwise} \end{cases},$$

and each of the above is a positive function.

Definition 3.4. Let $(X, \mathcal{F}, \mu)$ be a measure space, and let

$$\mathcal{M}(X, \mathcal{F})^+ = \{f : X \to \mathbb{R} : f \text{ is } \mathcal{F}\text{-measurable and } f(x) \geq 0 \ \forall \ x \in X\}$$

and

$$\Phi_f^+ = \{\varphi \in \mathcal{M}(X, \mathcal{F})^+ : \varphi \text{ is simple and } 0 \leq \varphi(x) \leq f(x) \ \forall \ x \in X\}.$$

If $f \in \mathcal{M}(X, \mathcal{F})^+$, we define the *integral of f with respect to μ* to be

$$\int f \, \mathrm{d}\mu = \sup_{\varphi \in \Phi_f^+} \int \varphi \, \mathrm{d}\mu$$

Moreover, we define the integral of f over any measurable set E by

$$\int_E f \, \mathrm{d}\mu = \int f \mathbb{1}_E \, \mathrm{d}\mu$$

Now we turn our attention to the interaction of the integral with limits. It is well known that given a sequence of functions $\{f_n\}$ which are Riemann integrable with limit f, the limit of the integrals need only converge to the integral of f provided the convergence of f_n to f is uniform. But we shall see that there are several weaker criteria which imply this type of convergence for Lebesgue integrals. These results are some of the most important improvements of the Lebesgue integral over the standard Riemann integral. The first criterion is monotonicity of the sequence of measurable functions $\{f_n\}$.

Theorem 3.5. (Monotone Convergence Theorem) *Let $(X, \mathcal{F}, \mu)$ be a measure space, and let $\{f_n\}$ be an increasing sequence of functions in $\mathcal{M}(X, \mathcal{F})^+$ converging pointwise to a function $f : X \to \mathbb{R}$. Then $f \in \mathcal{M}(X, \mathcal{F})^+$ and we have:*

$$\int f \, \mathrm{d}\mu = \lim_n \int f_n \, \mathrm{d}\mu$$

Proof. Note that since $\lim_n f_n(x) = f(x)$ exists, we have

$$\limsup f_n(x) = \lim_n f_n(x) = f(x),$$

which is measurable by proposition. Further, by hypothesis $f_n(x) \geq 0$ for all x, so that $f(x) \geq 0$ for all x, and $f \in \mathcal{M}(X, \mathcal{F})^+$. Since $f_n(x) \leq f(x)$, it follows that $\int f_n \, \mathrm{d}\mu$ is a bounded monotonically increasing sequence and in particular has a limit since $\int f \, \mathrm{d}\mu$ (the bound) exists; this shows half of what we want to prove. It remains to show that we also have the opposite inequality. Fix a nonnegative simple function $\varphi(x)$ where $\varphi(x) \leq f(x)$ for all $x \in X$. We first consider the function $\alpha\varphi$ where $\alpha \in (0, 1)$ and define $E_n = E_n(\alpha) = \{x \in X : \alpha\varphi(x) \leq f_n(x)\}$.

Now, since $f_n(x)$ converges pointwise to $f(x)$ and $\varphi(x) \leq f(x)$ by assumption, we have, for any x, either $\alpha\varphi(x) \leq f(x)$ or $\alpha\varphi(x) = f(x) = 0$, in which case we also have $f_n(x) = 0$ for all n because f_n is nonnegative and bounded above by f. Hence in either case we can find some N such that $\varphi(x) \leq f_N(x) \leq f_{n+1}(x)$, so that $\bigcup_n E_n = X$. (This was our motivation for introducting the constant α; we cannot make this claim without strict inequality). Then, for any n,

$$\int_{E_n} \alpha\varphi \, \mathrm{d}\mu = \alpha \int_{E_n} \varphi \, \mathrm{d}\mu \leq \int_{E_n} f_n \, \mathrm{d}\mu \leq \int f_n \, \mathrm{d}\mu$$

This tells us that

$$\alpha \lim_n \int_{E_n} \varphi \, \mathrm{d}\mu \leq \lim_n \int f_n \, \mathrm{d}\mu$$

Proposition 2.14 and the final part of Proposition 3.3 tell us that

$$\lim_n \int_{E_n} \varphi \, \mathrm{d}\mu = \int \varphi \, \mathrm{d}\mu$$

In particular, we have

$$\sup_{\alpha \in (0,1)} \alpha \lim_n \int_{E_n} \varphi \, \mathrm{d}\mu = \sup_{\alpha \in (0,1)} \alpha \int \varphi \, \mathrm{d}\mu \leq \lim_n \int f_n \, \mathrm{d}\mu$$

But then, since this procedure works for any simple function such that $\varphi \leq f$ we have

$$\int f \, \mathrm{d}\mu = \sup_{\phi \in \Phi_f^+} \int \varphi \, \mathrm{d}\mu \leq \lim_n \int f_n \, \mathrm{d}\mu$$

□

Remark 3.6. In our proof of the Monotone Convergence Theorem we used the fact that if φ is a nonnegative function and $E \subset F$ then $\int_E \varphi \, \mathrm{d}\mu \leq \int_F \varphi \, \mathrm{d}\mu$.

Theorem 3.7. *Let $(X, \mathcal{F}, \mu)$ be a measure space. If φ and ψ are nonnegative measurable functions and $\eta \in \mathbb{R}^+$ then the following hold:*

i. $\varphi + \psi$ *and* $\eta\phi$ *are nonnegative measurable functions.*

ii. $\int \eta\varphi \, \mathrm{d}\mu = \eta \int \varphi \, \mathrm{d}\mu$.

iii. $\int (\varphi + \psi) \, \mathrm{d}\mu = \int \varphi \, \mathrm{d}\mu + \int \psi \, \mathrm{d}\mu$.

iv. If $\varphi \geq \psi$ *then* $\int \varphi \, \mathrm{d}\mu \geq \int \psi \, \mathrm{d}\mu$.

v. $\lambda(E) = \int_E \varphi \, \mathrm{d}\mu$ *defines a measure on* $(X, \mathcal{F})$.

Proof. As in the proof of 3.3, i. is clear. Next, choose a monotonically increasing sequence of nonnegative simple functions $\{\varphi_n\}$ converging to φ and a similar sequence $\{\psi_n\}$ converging to ψ and note that $\{\eta\varphi_n\}$ and $\{\varphi_n + \psi_n\}$ are monotonically increasing sequences of nonnegative simple functions converging to $\eta\varphi$ and $\varphi + \psi$ respectively. Applying Proposition 3.3, we then invoke the Monotone Convergence Theorem to see

$$\lim_n \int \eta\varphi_n \, \mathrm{d}\mu = \eta \lim_n \int \varphi_n \, \mathrm{d}\mu = \eta \int \varphi \, \mathrm{d}\mu$$

and

$$\lim_n \int \varphi_n + \psi_n \, \mathrm{d}\mu = \lim_n \int \varphi_n \, \mathrm{d}\mu + \lim_n \int \psi_n \, \mathrm{d}\mu = \int \varphi_n \, \mathrm{d}\mu + \int \psi \, \mathrm{d}\mu,$$

proving ii. If we now suppose that $\varphi \geq \psi$ then by assumption $\psi_n \leq \psi \leq \varphi$. Then using the notation of Definition 3.4 we have

$$\int \psi_n \, \mathrm{d}\mu \leq \sup_{g \in \Phi_f^+} \int g \, \mathrm{d}\mu$$

since $\psi_n \in \Phi_f^+$. iii. then follows from the Monotone Convergence Theorem.

Finally, certainly $\lambda(\emptyset) = 0$. For all $E \in \mathcal{F}$, we have $\varphi \mathbb{1}_E \geq 0$, hence $\lambda(E) \geq 0$. Let $\{E_i\}$ be a countable collection of pairwise disjoint sets. Let $\mathbb{1}_k = \sum_{i=1}^k \mathbb{1}_{E_i}$. Clearly $\lim_k \mathbb{1}_k = \mathbb{1}_{\bigcup_i E_i}$ and $\{\mathbb{1}_k\}$ is a monotonically increasing sequence. Note that $\{\varphi\mathbb{1}_k\}$ is also a monotonically increasing sequence converging to $\varphi\mathbb{1}_{\bigcup_i E_i}$. Hence,

$$\lambda(\bigcup_i E_i) = \int \varphi \mathbb{1}_{\bigcup_i E_i} \, \mathrm{d}\mu = \lim_k \int \varphi \mathbb{1}_k \, \mathrm{d}\mu,$$

and this last is exactly

$$\lim_k \int \varphi \sum_{i=1}^n \mathbb{1}_{E_i} \, \mathrm{d}\mu = \lim_k \sum_{i=1}^k \int \varphi \mathbb{1}_{E_i} \, \mathrm{d}\mu = \sum_i \int \varphi \mathbb{1}_{E_i} \, \mathrm{d}\mu = \sum_i \lambda(E_i),$$

where we have again applied the Monotone Convergence Theorem. □

This together with our preceding remark establishes the linearity of the general integral. Note that this methodology removes the messy and unenlightening manipulation of Riemann sums required to show the Riemann integral is linear. In this context, linearity appears as a consequence of the way in which the Lebesgue integral behaves with respect to limits, whereas in some sense the proof of this fact for Riemann integrals obfuscates what is actually going on. In fact, taking the measure-theoretic point of view has no downside: it is trivial to show that the Lebesgue integral is an extension of the Riemann integral. That is, any Riemann integrable function is Lebesgue integrable and their integrals coincide. We do not prove this result here because we have not actually constructed the Lebesgue measure on $\mathbb{R}$, only argued that such a measure — if existent — meets the intuitive requirements for a size function on the real line.

Definition 3.8. Let $(X, \mathcal{F}, \mu)$ be a measure space and let $\mathcal{L}(X, \mathcal{F}, \mu)$ be the set

$\{f \in \mathcal{M}(X, \mathcal{F})$ such that f^+ and f^- have a finite integral with respect to $\mu\}$,

where (as we mentioned before) f^+ and f^- are defined by

$$f^+(x) = \begin{cases} f(x) & \text{if } f(x) > 0, \\ 0 & \text{otherwise;} \end{cases} \qquad f^-(x) = \begin{cases} -f(x) & \text{if } f(x) < 0, \\ 0 & \text{otherwise} \end{cases}.$$

For all $f \in \mathcal{L}(X, \mathcal{F}, \mu)$ the *integral* of f with respect to μ is defined to be

$$\int f \, \mathrm{d}\mu = \int f^+ \, \mathrm{d}\mu - \int f^- \, \mathrm{d}\mu$$

We finish our introduction to measure theory by examining some other ways in which the integral interacts nicely with limits.

Proposition 3.9. (Fatou's Lemma) *Let $(X, \mathcal{F}, \mu)$ be a measure space. Let $\{f_n\}$ be a sequence of functions in $\mathcal{M}(X, \mathcal{X})^+$; then*

$$\int \underline{f} \, \mathrm{d}\mu \leq \liminf_n \int f_n \, \mathrm{d}\mu.$$

Proof. The function $\inf_{k \geq n} f_k(x)$ is measurable by Proposition 2.7. Clearly $\inf_{k \geq n} f_k(x) \leq f_k(x)$ for $k \geq n$. This implies

$$\int \inf_{k \geq n} f_k \, \mathrm{d}\mu \leq \int f_k \, \mathrm{d}\mu$$

for $k \geq n$. In particular this remains true when we take the infimum over $k \geq n$. Thus

$$\lim_n \int \inf_{k \geq n} f_k \, \mathrm{d}\mu \leq \lim_n \inf_{k \geq n} \int f_k \, \mathrm{d}\mu;$$

but note that $\inf_{k \geq n} f_k(x)$ is a sequence of monotonically increasing positive functions indexed by n. Hence by the Monotone Convergence Theorem

$$\int \lim_n \inf_{k \geq n} f_k \, \mathrm{d}\mu = \lim_n \int \inf_{k \geq n} f_k \, \mathrm{d}\mu \leq \lim_n \inf_{k \geq n} \int f_k \, \mathrm{d}\mu.$$

But by the definition of the limit infimum this is what we wanted to show. □

Proposition 3.10. *If $f \in \mathcal{M}(X, \mathcal{F})$, then*

i. $f \in \mathcal{L}(X, \mathcal{F}, \mu)$ *if and only if* $|f| \in \mathcal{L}(X, \mathcal{F}, \mu)$.

ii. $\left| \int f \, \mathrm{d}\mu \right| \leq \int |f| \, \mathrm{d}\mu.$

Proof. Assume $f \in \mathcal{L}(X, \mathcal{F}, \mu)$. Then $|f|^+ = |f| = f^+ + f^-$ which are measurable and have finite integrals by assumption, and $|f|^- = 0$. Thus clearly $|f| \in \mathcal{L}(X, \mathcal{F}, \mu)$. On the other hand, if $|f| \in \mathcal{L}(X, \mathcal{F}, \mu)$ note that $f^+ \leq |f|$ and $f^- \leq |f|$ since $-|f| \leq f \leq |f|$. Then clearly these integrals are finite by comparison; hence $f \in \mathcal{L}(X, \mathcal{F}, \mu)$. This proves i. ii. follows since $-|f| \leq f \leq |f|$ implies

$$-\int |f| \, \mathrm{d}\mu \leq \int f \, \mathrm{d}\mu \leq \int |f| \, \mathrm{d}\mu.$$

□

Remark 3.11. Note that the previous Proposition is not true for the Riemann integral. Consider, for example

$$f(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q} \\ -1 & \text{if } x \in \mathbb{R} \setminus \mathbb{Q} \end{cases}$$

on the interval $[0, 1]$. $|f|$, being a constant, is Riemann integrable, but f, being discontinuous everywhere, is not.

Theorem 3.12. (Dominated Convergence Theorem) *Let $(X, \mathcal{F}, \mu)$ be a measure space and $\{f_n\}$ a sequence of real functions in $\mathcal{M}(X, \mathcal{F})$ converging to a real valued function $f \in \mathcal{M}(X, \mathcal{F})$. Let $g \in \mathcal{L}(X, \mathcal{F}, \mu)$ be a nonegative function. If $|f_n| \leq g$, then $\forall n f_n, f \in \mathcal{L}(X, \mathcal{F}, \mu)$ and*

$$\int f \, \mathrm{d}\mu = \lim_n \int f_n \, \mathrm{d}\mu$$

Proof. First note that $|f_n| \leq g$ so the integral of $|f_n|$ is finite, hence $|f_n| \in \mathcal{L}(X, \mathcal{F}, \mu)$ since g is. Then by Proposition 1.6 we have $f_n \in \mathcal{L}(X, \mathcal{F}, \mu)$. It follows that $f \in \mathcal{L}(X, \mathcal{F}, \mu)$ also since

$$\int \underline{f} \, \mathrm{d}\mu = \int \lim_n f_n \, \mathrm{d}\mu = \int f \, \mathrm{d}\mu \leq \liminf \int f_n \, \mathrm{d}\mu$$

which is clearly finite since each f_n has a finite integral. Since $g \pm f_n \geq 0$ we apply Fatou's lemma to $\{g \pm f_n\}$ This yields

$$\begin{aligned} \int \liminf(g \pm f_n) \, \mathrm{d}\mu &= \int g \, \mathrm{d}\mu \pm \int \liminf f_n \, \mathrm{d}\mu \\ &= \int g \, \mathrm{d}\mu \pm \int f \, \mathrm{d}\mu \\ &\leq \liminf \int g \pm f_n \, \mathrm{d}\mu \\ &= \liminf \left(\int g \, \mathrm{d}\mu \pm \int f_n \, \mathrm{d}\mu \right) \end{aligned}$$

Consequently,

$$\int g \, \mathrm{d}\mu + \int f \, \mathrm{d}\mu \leq \int g \, \mathrm{d}\mu + \liminf \int f_n \, \mathrm{d}\mu$$

and

$$\int g\,\mathrm{d}\mu - \int f\,\mathrm{d}\mu \leq \int g\,\mathrm{d}\mu + \liminf(-\int f_n\,\mathrm{d}\mu) \leq \int g\,\mathrm{d}\mu - \limsup \int f_n \mathrm{d}\mu$$

Subtracting the common term from all of these inequalities and multiplying the second by negative one yields

$$\limsup \int f_n\,\mathrm{d}\mu \leq \int f\,\mathrm{d}\mu \leq \liminf \int f_n\,\mathrm{d}\mu$$

Hence

$$\lim_n \int f_n\,\mathrm{d}\mu = \int f\,\mathrm{d}\mu$$

and in particular exists. □

4 The Fubini Theorem

We have developed the machinery for defining and calculating integrals over abstract spaces, but the theory so far does not take into account any additional structure on those spaces. For example, though our theory allows us to integrate over $\mathbb{R}^n$ we must, so to speak, integrate all at once. We would like to treat $\mathbb{R}^n$ as $\mathbb{R} \times \mathbb{R} \times \ldots \mathbb{R}$, and then integrate over $\mathbb{R}^n$ by separately integrating over each copy of $\mathbb{R}$. Towards this end we will now describe how the cartesian product of two measure spaces may be regarded as a measure space.

First we assume that we have two measure spaces $(X, \mathcal{X}, \mu)$ and $(Y, \mathcal{Y}, \nu)$. If $A \in \mathcal{X}$ and $B \in \mathcal{Y}$ are measurable we call $A \times B$ a *measurable rectangle*. We are interested in the σ-algebra generated by the set of all measurable rectangles.

Definition 4.1. We denote the σ-algebra generated by the set of measurable rectangles by $\mathcal{X} \times \mathcal{Y}$. The associated measurable space is $(X \times Y, \mathcal{X} \times \mathcal{Y})$. We refer to this as the *cartesian product* of the measurable spaces $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$.

We would like to define a measure π on $(X \times Y, \mathcal{X} \times \mathcal{Y})$ which is in some way compatible with the measures on $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$. One compatibility condition we might expect is that, for any measurable rectangle $A \times B$, we should have $\pi(A \times B) = \mu(A)\nu(B)$. This would be in accordance with the usual way we compute areas of geometric objects in $\mathbb{R}^2$. Indeed, *defining* π this way gives a measure on the set of finite unions of measurable rectangles, which is an algebra, but this set is not in general a σ-algebra. It is not obvious a priori that we can extend π to be a measure on $\mathcal{X} \times \mathcal{Y}$, or that there should be a unique such extension, and proving this requires some additional machinery, in particular the Carathéodory Extension Theorem, the Hahn Extension Theorem, and the construction of outer measure. These are interesting topics, but not of vital importance to our development. In lieu of developing this machinery, we will simply state the theorem that we will require in the sequel.

Theorem 4.2. (Product Measure) *If $(X, \mathcal{X}, \mu)$ and $(Y, \mathcal{Y}, \nu)$ are measure spaces, then there exists a measure $\pi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ such that $\pi(A \times B) = \mu(A)\nu(B)$ for every measurable rectangle $A \times B$. If μ, ν are σ-finite, then π is unique and σ-finite.*

For a more complete development of these topics see [1]. The reader may recall from multivariable calculus the Fubini Theorem, which is introduced as a means of computing double integrals by iteration; this is exactly the treatment we described in the introduction to this chapter. In preparation for stating the more general measure-theoretic Fubini Theorem we need to consider objects that are in some sense one dimensional slices of the sets in our product measure space. This motivates the following definition.

Definition 4.3. Let $(X \times Y, \mathcal{X} \times \mathcal{Y})$ be as in Definition 4.1. Let $S \subset X \times Y$. Then if $x \in X$, $y \in Y$ we define the *x-section* of S by

$$S_x = \{y \in Y : (x, y) \in S\}$$

and the *y-section* of S by

$$S^y = \{x \in X : (x, y) \in S\}$$

In a similar manner, we can define sections of functions.

Definition 4.4. Let $(X \times Y, \mathcal{X} \times \mathcal{Y})$ be as in Definition 4.1, and let f be a real-valued measurable function on $X \times Y$. Then if $x \in X$, $y \in Y$ we define $f_x : Y \to \mathbb{R}$ and $f^y : X \to \mathbb{R}$ by

$$f_x(y) = f(x, y) \text{ and } f^y(x) = f(x, y)$$

and call f_x the *x-section* of f and f^y the *y-section* of f.

Our goal is to develop an analogue of the Fubini theorem by considering sections of a product set S. We will integrate over all the, say, x-sections of a set with respect to the measure ν associated to Y. One would hope that with suitable measurability conditions that this process would recover the full measure of the set S.

Proposition 4.5. *Let $(X \times Y, \mathcal{X} \times \mathcal{Y})$ be as in Definition 4.1. Let $S \subset X \times Y$ be measurable. Then:*

i. $S^y \in \mathcal{X}$ and $S_x \in \mathcal{Y}$ for every $x \in X$ and $y \in Y$.

ii. If $f : X \times Y \to \mathbb{R}$ is $\mathcal{X} \times \mathcal{Y}$-measurable then f^y is $\mathcal{X}$-measurable and f_x is $\mathcal{Y}$-measurable.

Proof. For i. it suffices to show that the sets

$$(\mathcal{X} \times \mathcal{Y})_x = \{E \subset X \times Y : E_x \in \mathcal{Y},\ x \in X\}$$

and

$$(\mathcal{X} \times \mathcal{Y})^y = \{E \subset X \times Y : E^y \in \mathcal{X},\ y \in Y\}$$

are both σ-algebras containing all measurable rectangles. Then it follows immediately that both must contain $\mathcal{X} \times \mathcal{Y}$ because this is the σ-algebra generated by the set of measurable rectangles. We prove that $(\mathcal{X} \times \mathcal{Y})_x$ is a σ-algebra and the proof for the other set is the same. First note that for $x \in X$, $(X \times Y)_x = Y \in \mathcal{Y}$ and also $\emptyset_x = \emptyset \in \mathcal{Y}$. Now, it is easy to check that, if $E \in (\mathcal{X} \times \mathcal{Y})_x$ then $((X \times Y) \setminus E)_x = (X \times Y)_x \setminus E_x \in \mathcal{Y}$ since $\mathcal{Y}$ is a σ-algebra. Now let $E_n \in (\mathcal{X} \times \mathcal{Y})_x$. It is also easy to check that

$$\left(\bigcup_n E_n\right)_x = \bigcup_n (E_n)_x \in \mathcal{Y}$$

because by assumption $(E_n)_x \in \mathcal{Y}$ by assumption. Then $(\mathcal{X} \times \mathcal{Y})_x$ is a σ-algebra and hence it must contain $\mathcal{X} \times \mathcal{Y}$.
It is easy to check that

$$(f^y)^{-1}[(-\infty, a)] = (f^{-1}[(-\infty, a)])^y$$

and

$$(f_x)^{-1}[(-\infty, a)] = (f^{-1}[(-\infty, a)])_x$$

Both of these sets are measurable by part i. since f is a measurable function, so f^y, f_x are measurable also. □

We are almost ready to prove the Fubini Theorem. We need one more technical lemma, which we give below.

Definition 4.6. A *monotone class* is a non-empty collection $\mathcal{C}$ of subsets of some set X such that

i. For every increasing sequence $\{E_n\} \subset \mathcal{C}$, we have $\bigcup_n E_n \in \mathcal{C}$.

ii. For every decreasing sequence $\{E_n\} \subset \mathcal{C}$, we have $\bigcap_n E_n \in \mathcal{C}$.

Proposition 4.7. (Monotone Class Lemma) *Let $\mathcal{A}$ be an algebra of sets. Then the monotone class $\mathcal{C}(\mathcal{A})$ generated by $\mathcal{A}$ is the same as the σ-algebra $\sigma(\mathcal{A})$ generated by $\mathcal{A}$.*

Proof. The proof is basically formal and is omitted. One starts by showing that the intersection of monotone classes is again a monotone class, and that every σ-algebra is a monotone class. □

Corollary 4.8. *Let $\mathcal{C}$ be a monotone class containing an algebra $\mathcal{A}$. Then $\mathcal{C}$ contains the σ-algebra generated by $\mathcal{A}$.*

We now have all the tools we will need to prove our lemma. This next result is essentially the Fubini Theorem for characteristic functions. Once we have this result proving the more general Fubini Theorem is straightforward.

Lemma 4.9. *Let $(X, \mathcal{X}, \mu)$ and $(Y, \mathcal{Y}, \nu)$ be σ-finite measure spaces. Then for any $E \in \mathcal{X} \times \mathcal{Y}$ the functions $f_E : X \to \mathbb{R}$ and $g_E : Y \to \mathbb{R}$ given by*

$$f_E(x) = \nu(E_x) \text{and } g_E(y) = \mu(E^y)$$

are measurable and

$$\int_X f_E \, \mathrm{d}\mu = \pi(E) = \int_Y g_E \, \mathrm{d}\nu$$

Proof. We begin by proving this theorem assuming our measure spaces are finite. Then we will extend the result to the class of σ-finite spaces. Suppose $\mathcal{P}$ is the algebra of sets generated by measurable rectangles. Set

$$\mathcal{S} = \{f_E \in \mathcal{M}(X, \mathcal{X})^+, g_E \in \mathcal{M}(Y, \mathcal{Y})^+, \text{and } \int_X f_E \, \mathrm{d}\mu = \pi(E) = \int_Y g_E \, \mathrm{d}\mu\}.$$

We want to show that $\mathcal{S}$ is a σ-algebra containing $\mathcal{P}$. If we succeed in this, because $\mathcal{X} \times \mathcal{Y}$ is generated by $\mathcal{P}$, we will have shown that $\mathcal{S}$ contains $\mathcal{X} \times \mathcal{Y}$. It is easy to see that this set contains every measurable rectangle $A \times B$. Indeed, note that for any $x \in X, y \in Y$ we have $f_E = \mu(E^y) = \mu(A)\mathbb{1}_B(y)$ and similarly $g_E = \nu(E_x) = \nu(B)\mathbb{1}_A(x)$. Hence

$$\int_X f_E \, \mathrm{d}\mu = \int_X \nu(B)\mathbb{1}_A \, \mathrm{d}\mu = \nu(B) \int_X \mathbb{1}_A \, \mathrm{d}\mu = \mu(A)\nu(B),$$

and tracing the argument backwards this is also equal to $\displaystyle\int_Y g_E \, \mathrm{d}\nu$.

Now for any set $P \in \mathcal{P}$ we can write $P = \bigcup_i A_i \times B_i$ as the union of disjoint rectangles. Consequently, if π denotes the measure on $\mathcal{X} \times \mathcal{Y}$ as before, we have

$$\begin{aligned}
\pi(P) &= \pi(\bigcup_i A_i \times B_i) \\
&= \sum_i \pi(A_i \times B_i) \\
&= \sum_i \int \nu(B_i)\mathbb{1}_{A_i} \, \mathrm{d}\mu \\
&= \int \sum_i \nu(B_i)\mathbb{1}_{A_i} \, \mathrm{d}\mu \\
&= \int \sum_i \mu((A_i \times B_i)^y) \, \mathrm{d}\mu,
\end{aligned}$$

and also

$$\sum_i \mu((A_i \times B_i)^y) = \mu(\bigcup_i (A_i \times B_i)^y) = \mu((\bigcup_i A_i \times B_i)^y)$$

since the sets above are disjoint and it is simple to check that the union of y-sections is the y-section of the union. We also used the Monotone Convergence Theorem in the second-to-last equality of the array above. Thus

$$\pi(P) = \int \mu((\bigcup_i A_i \times B_i)^y)\,\mathrm{d}\mu = \int f_{\bigcup_i A_i \times B_i}\,\mathrm{d}\mu = \int f_P\,\mathrm{d}\mu.$$

A similar argument establishes the equality

$$\pi(P) = \int_Y g_P\,\mathrm{d}\mu,$$

and we deduce $\mathcal{P} \subset \mathcal{S}$. Now, it remains to show that $\mathcal{S}$ is a monotone class; we will then apply the Monotone Class Lemma to finish the proof. To this end, let $\{E_n\}$ be an increasing sequence of sets in $\mathcal{S}$. For all n we have

$$\int_X f_{E_n}\,\mathrm{d}\mu = \pi(E_n) = \int_Y g_{E_n}\,\mathrm{d}\nu.$$

Now for any $x \in X$ and $y \in Y$, the sequences $\{(E_n)_x\}$ and $\{(E_n)^y\}$ are also increasing. Clearly $\lim_n f_{E_n} = f_E$ and $\lim_n g_{E_n} = g_E$, and furthermore these are monotonically increasing sequences. Also, $\lim_n \pi(E_n) = \pi(\bigcup_n E_n) = \pi(E)$.

Hence, applying the Monotone Convergence Theorem we have

$$\lim_n \int_X f_{E_n}\,\mathrm{d}\mu = \int_X f_E\,\mathrm{d}\mu = \lim_n \int_Y g_{E_n}\,\mathrm{d}\mu = \int_Y g_E\,\mathrm{d}\mu = \lim_n \pi(E_n) = \pi(E).$$

Similarly, assume $\{E_n\}$ is a decreasing sequence. Again it is clear that

$$\lim_n f_{E_n} = \lim_n \nu((E_n)_x) = \nu(\bigcap_n (E_n)_x) = \nu((\bigcap_n E_n)_x) = \nu(E) = f_E$$

and similarly $\lim_n g_{E_n} = g_E$. Again, $\lim_n \pi(E_n) = \lim_n \pi(\bigcap_n E_n) = \pi(E)$. Here our sequence of functions is not increasing, but note that $0 \leq f_{E_n} \leq K$ and $0 \leq g_{E_n} \leq K$ for $K = \max\{\mu(X), \nu(Y)\}$ which exists since our measure spaces are finite. Thus, we may apply the Dominated Convergence Theorem and after taking limits we deduce

$$\int_X f_E\,\mathrm{d}\mu = \pi(E) = \int_Y g_E\,\mathrm{d}\nu.$$

Hence $\mathcal{S}$ is a monotone class containing $\mathcal{P}$, and thus contains $\mathcal{X} \times \mathcal{Y}$.

Now let us assume our measure spaces are σ-finite but not necessarily finite. Clearly the product measure space is also. Let $\{A_n\}$ be a measurable covering of $X \times Y$ such that for all $n \in \mathbb{N}$, $A_n \subset A_{n+1}$ and $\pi(A_n) < \infty$. For each n we define a new finite measure space on the σ-algebra $\mathcal{P}(A_n) \cap \mathcal{X} \times \mathcal{Y}$. We denote

μ or ν restricted to this space by μ_n and ν_n respectively. Let E be a measurable set in $(X \times Y, \mathcal{X} \times \mathcal{Y})$ and define $E_n = E \cap A_n$. First note that

$$\begin{aligned}\lim_n f_{E_n}(x) &= \lim_n \nu((E \cap A_n)_x) = \lim_n \nu(E_x \cap (A_n)_x) \\ &= \nu(\bigcup_n E_x \cap (A_n)_x) = \nu(E_x \cap (\bigcup_n A_n) \\ &= \nu(E_x \cap X) = \nu(E_x) = f_E(x)\end{aligned}$$

Where we have used the fact that $\{A_n\}$ is an increasing sequence. Similarly, $\lim_n g_{E_n}(y) = g_E(y)$. Further, it is obvious that these sequences are increasing. Since $\mathcal{P}(A_n) \cap \mathcal{X} \times \mathcal{Y}$ is a finite measure space we have

$$\int_X f_{E_n} \, \mathrm{d}\mu = \int_X f_{E_n} \, \mathrm{d}\mu_n = \int \mathbb{1}_{E_n} \, \mathrm{d}\pi = \int_Y g_{E_n} \, \mathrm{d}\nu_n = \int_Y g_{E_n} \, \mathrm{d}\nu$$

where we have used the fact that since $E_n \subset A_n, \mu_n$ and μ coincide, as do ν_n and ν. Hence, the Monotone Convergence Theorem tells us that taking the limit of this expression yields

$$\int_X f_E \, \mathrm{d}\mu = \int_E \mathrm{d}\pi = \int_Y g_E \, \mathrm{d}\nu$$

□

Now we will prove the Fubini Theorem for non-negative functions.

Theorem 4.10. (Fubini-Tonelli) *Let $(X, \mathcal{X}, \mu)$ and $(Y, \mathcal{Y}, \nu)$ be σ-finite measure spaces. If $h \in \mathcal{M}(X \times Y, \mathcal{X} \times \mathcal{Y})^+$ then f_h and g_h given by*

$$f_h(x) = \int_Y h_x \, \mathrm{d}\nu; \qquad g_h(y) = \int h^y \, \mathrm{d}\mu$$

belong respectively to $\mathcal{M}(X, \mathcal{X})^+$ and $\mathcal{M}(Y, \mathcal{Y})^+$ and

$$\int_X f_h \, \mathrm{d}\mu = \int_{X \times Y} h \, \mathrm{d}\pi = \int_Y g_h \, \mathrm{d}\nu.$$

Proof. First, let us assume $h = \mathbb{1}_E$ for some measurable set E. It follows from the definition of section that $(\mathbb{1}_E)_x = \mathbb{1}_{E_x}$ and $(\mathbb{1}_E)^y = \mathbb{1}_{E_y}$. But then, we have $f_h(x) = \nu(E_x)$ and $g_h(y) = \mu(E_x)$. Hence Lemma 4.9 gives the desired result immediately. Linearity of the integral then extends this result to all positive measurable simple functions. Finally, assume $h \in \mathcal{M}(X \times Y, \mathcal{X} \times \mathcal{Y})^+$ and let $\{h_n\}$ be a sequence of monotonically increasing positive simple functions converging to h. It is easy to verify that $\{(h_n)_x\}$ and $\{(h_n)^y\}$ are also sequences of monotonically increasing positive simple functions. Furthermore, basic properties of the integral show that $\{f_{h_n}\}$ and $\{g_{h_n}\}$ are monotonically increasing sequences of simple function as well. The ubiquitous Monotone Convergence Theorem applied to $\{(h_n)_x\}$ and $\{(h_n)^y\}$ then implies $\lim_n f_{h_n} = f_h$ and $\lim_n g_{h_n} = g_h$. Finally, since for all n we have

$$\int_X f_{h_n} \, \mathrm{d}\mu = \int_{X \times Y} h_n \, \mathrm{d}\pi = \int_Y g_{h_n} \, \mathrm{d}\nu,$$

taking the limit as n goes to infinity and applying the Monotone Convergence Theorem completes the proof of the general statement. □

Theorem 4.11. (Fubini) *Let $(X, \mathcal{X}, \mu)$ and $(Y, \mathcal{Y}, \nu)$ be σ-finite measure spaces. If $h \in \mathcal{L}(X \times Y, \mathcal{X} \times \mathcal{Y}, \pi)$ then the functions $f_h : X \to \mathbb{R}$ and $g_h : Y \to \mathbb{R}$ defined almost everywhere by*

$$f_h(x) = \int_Y h_x \, \mathrm{d}\nu \text{ and } g_h(y) = \int_X h^y \, \mathrm{d}\mu$$

belong respectively to $\mathcal{L}(X, \mathcal{X}, \mu)$ and $\mathcal{L}(Y, \mathcal{Y}, \nu)$, and

$$\int_X f_h \, \mathrm{d}\mu = \int_{X \times Y} h \, \mathrm{d}\pi = \int_Y g_h \, \mathrm{d}\nu.$$

Proof. The functions h^+ and h^- satisfy the hypothesis of Theorem 4.10. Thus, we have

$$\int_X f_{h^+} \, \mathrm{d}\mu = \int_{X \times Y} h^+ \, \mathrm{d}\pi = \int_Y g_{h^+} \, \mathrm{d}\nu$$

and

$$\int_X f_{h^-} \, \mathrm{d}\mu = \int_{X \times Y} h^- \, \mathrm{d}\pi = \int_Y g_{h^-} \, \mathrm{d}\nu.$$

Combining these gives

$$\int_X (f_{h^+} - f_{h^-}) \, \mathrm{d}\mu = \int_{X \times Y} (h^+ - h^-) \, \mathrm{d}\pi = \int_Y (g_{h^+} - g_{h^-}) \, \mathrm{d}\nu.$$

But note that

$$f_{h^+} - f_{h^-} = \int_Y (h^+)_x - (h^-)_x \, \mathrm{d}\nu = \int_Y (h_x)^+ - (h_x)^- \, \mathrm{d}\nu = \int_Y h_x \, \mathrm{d}\nu = f_h(x).$$

The analagous equality for g_h also holds; all we have used here is the fact that $(h^\pm)_x = (h_x)^\pm$, which follows at once from their definitions. Hence the above equality may be written as

$$\int_X f_h \, \mathrm{d}\mu = \int_{X \times Y} h \, \mathrm{d}\pi = \int_Y g_h \, \mathrm{d}\nu,$$

as desired. □

Remark 4.12. Actually, this proof involved a bit of hand-waving in that we glossed over an unilluminating subtlety. We do not a priori know that f_{h^+} and f_{h^-} are everywhere finite. However, we do know that they have finite integrals by the Tonelli Theorem. Consequently, there may be a set N of measure zero on which our above argument does not hold. That is, we may not be able to sensibly write $f_{h^+} - f_{h^-}$ on the entire space X. In practice one simply excises the set N from discussion and ignores it, which is why in the statement of the theorem we noted that f_h and g_h were defined *almost everywhere*.

The Fubini Theorem, while perhaps not vital for understanding the Central Limit Theorem, will be important to us later when dealing with indicator functions. It is interesting to note that proving Fubini for the Riemann integral is somewhat trivial, but the theorem for the Lebesgue integral requires much more machinery. However, our approach based around developing the notion of sections seems to provide more insight than the estimation of upper and lower sums that is used in the Riemann case, mostly because our assumptions are much less restrictive than in the Riemann case. We required only that h be integrable, whereas for the Riemann case one much check that h is integrable and that all the iterated integrals exist. The reason we do not require these assumptions is because they are roughly analogous to the results of Proposition 4.5, which followed from the fact that h is measurable.

5 Distributions and Indicator Functions

Now we wish to develop the basic concepts of probability theory. We already know what a probability space is - it is a measure space with total mass 1. We often denote such a space by Ω, and an element of such a space $\omega \in \Omega$, called an *outcome*. Measurable subsets of Ω are called *events*, and measurable functions on a probability space are called *random variables*. We introduce a probabilistic way of thinking about these functions, and then introduce distribution functions, first as derived from random variables and then as totally separate quantities. Finally we introduce indicator functions as a means of working with distribution functions. The reason for this approach will become clear in the next section, where we will need the theory of indicator functions to prove the Central Limit Theorem.

Definition 5.1. If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, then a *random variable* is a measurable function $X : \Omega \to \mathbb{R}$. If $\int |X| \, \mathrm{d}\mathbb{P} < \infty$ then we define the *expectation*, *expected value*, or *mean* of X to be

$$\mathbb{E}(X) = \int X \, \mathrm{d}\mathbb{P}.$$

We also briefly define one more concept which we will require later.

Definition 5.2. The *variance* of a random variable is given by

$$\mathrm{Var}(X) := \sigma(X)^2 = \mathbb{E}(X - \mathbb{E}(X)).$$

It is easy to check that $\sigma(X)^2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2$. Often, the square root of the variance, $\sigma = \sigma(X)$, is referred to as the *standard deviation* of X.

The simplest example of a random variable is the *indicator function* $\mathbb{1}_A$ that we have already used frequently We call this the characteristic function of the set A and denote it χ_A when we are thinking of our space as a measure space.

We may encounter a case where we have two spaces, $(\Omega, \mathcal{F}, \mathbb{P})$ and $(\Omega_1, \mathcal{F}_1)$, and a measurable function $h : \Omega \to \Omega_1$. Then we can define a measure ν_h on $(\Omega_1, \mathcal{F}_1)$ by $\nu_h(A_1) = \mathbb{P}[h^{-1}(A_1)]$. Furthermore, if f_1 is a random variable on $(\Omega_1, \mathcal{F}_1, \nu_h)$ then $f(x) = f_1[h(x)]$ is a random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ and they have the same expected value.

Proposition 5.3. *If either side exists, then*

$$\mathbb{E}(f) = \int f \, \mathrm{d}\mathbb{P} = \int f_1 \, \mathrm{d}\nu_h$$

Proof. We prove this first for the indicator function $\mathbb{1}_A : \Omega_1 \to \mathbb{R}$. There

$$\int (\mathbb{1}_A \circ h) \, \mathrm{d}\mathbb{P} = \mathbb{P}[h^{-1}(A)] = \nu_h(A) = \int \mathbb{1}_A \, \mathrm{d}\nu_h.$$

Now we consider the case of a simple function $\varphi : \Omega_1 \to \mathbb{R}$ given by

$$\varphi(x) = \sum_{i=1}^{n} \alpha_i \mathbb{1}_{A_i}(x)$$

where $\alpha_i \in \mathbb{R}$ and $A_i \subset \Omega_1$. In this case,

$$\begin{aligned}
\int (\varphi \circ h) \, \mathrm{d}\mathbb{P} &= \int \sum_{i=1}^{n} \alpha_i (\mathbb{1}_{A_i} \circ h) \, \mathrm{d}\mathbb{P} \\
&= \sum_{i=1}^{n} \alpha_i \int (\mathbb{1}_{A_i} \circ h) \, \mathrm{d}\mathbb{P} \\
&= \sum_{i=1}^{n} \alpha_i \int \mathbb{1}_{A_i} \, \mathrm{d}\nu_h \\
&= \int \sum_{i=1}^{n} \alpha_i \mathbb{1}_{A_i} \, \mathrm{d}\nu_h \\
&= \int \varphi \, \mathrm{d}\nu_h,
\end{aligned}$$

by linearity of the integral. Finally, let $f_1 : \Omega_1 \to \mathbb{R}$ be an arbitrary measurable function. Then, as usual, we can choose a sequence of monotonically increasing simple functions $\{f_{1n}\}$ which converge to f_1. The sequence $f_n = f_{1n} \circ h$ also increases monotonically, and converges to f. Hence, by the Monotone Convergence Theorem we have

$$\mathbb{E}(f) = \int f \, \mathrm{d}\mathbb{P} = \lim_n \int f_n \, \mathrm{d}\mathbb{P} = \lim_n \int f_{1n} \, \mathrm{d}\nu_h = \lim_n \int f_1 \, \mathrm{d}\nu_h$$

which finishes the proof. □

We are often interested in a random variable X and the probability that it takes values in various sets with respect to a probability measure μ.

Definition 5.4. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X a random variable. We define the *probability distribution* or *law* of X to be the function

$$\mathbb{P}_X(B) = \mu(X^{-1}(B))$$

for $B \subset X$. We define the *distribution function of* X to be the function

$$F_X(x) = \mathbb{P}_X((-\infty, x])$$

We have the obvious consequence that for $a < b$ we have $F(b) - F(a) = \mu\{(a, b]\}$.

Corollary 5.5. *If $g : \mathbb{R} \to \mathbb{R}$ then we have*

$$\mathbb{E}[g \circ X] = \int_\Omega (g \circ X)\, \mathrm{d}\mathbb{P} = \int_{\mathbb{R}} g\, \mathrm{d}\mathbb{P}_X.$$

Proof. This is Proposition 5.3 applied to the machinery of Definition 5.4. □

Proposition 5.6. *Let X be a random variable and F be its distribution function. Then F has the following properties*

i. *F is non-decreasing*

ii. *F is continuous from the right*

iii. $\lim_{x\to\infty} F(x) = 1$

iv. $\lim_{x\to-\infty} F(x) = 0$

Proof. i. follows immedietely since $x \leq y$ implies $(-\infty, x) \subset (-\infty, y)$ which shows that $X^{-1}[(-\infty, x)] \subset X^{-1}[(-\infty, y)]$ so in particular

$$F(x) = \mathbb{P}(X^{-1}[(-\infty, x)]) \leq \mathbb{P}(X^{-1}[(-\infty, y)]) = F(y)$$

ii. follows since if $\{x_n\}$ is a sequence converging to x from the right

$$\lim_{x_n \to x+} F(x_n) = \lim_n \mathbb{P}(X^{-1}[(-\infty, x_n)]$$

but $\{X^{-1}[(-\infty, x_n)]\}$ is a decreasing sequence, hence by Proposition 2.9

$$\begin{aligned}
\lim_n \mathbb{P}(X^{-1}[(-\infty, x_n)]) &= \mathbb{P}\left(\bigcap_n X^{-1}[(-\infty, x_n)]\right) \\
&= \mathbb{P}\left(X^{-1}\left[\bigcap_n (-\infty, x_n)\right]\right) \\
&= \mathbb{P}\left(X^{-1}(-\infty, x]\right) = F(x)
\end{aligned}$$

So F is right continuous.
iii. and iv. are similar and are left to the reader. □

Remark 5.7. This proof used that fact that if x_n converges to x from the right then $\bigcap_n (-\infty, x_n) = (-\infty, x]$. Note that if x_n converges to x from the left but $x_n \neq x$ for any n, then $\bigcap_n (-\infty, x_n) = (-\infty, x)$. This subtle point is the reason a distribution function may fail to be continuous.

Conversely, it is often the case that for a function satisfying some of the above properties that it is the distribution of some random variable. This is not always the case, but it motivates the following more general definition.

Definition 5.8. A function $F : \mathbb{R} \to \mathbb{R}$ is said to be a *distribution function* if

i. $0 \leq F \leq 1$.

ii. F is non-decreasing.

iii. F is continuous from the right.

We define the *variation* of a distribution function by

$$\mathrm{Var}(F) = \lim_{x \to \infty} F(x) - \lim_{x \to -\infty} F(x).$$

If F is the distribution function of a random variable we have $\mathrm{Var}(F) = 1$. We use the same notation for variation and variance, but no confusion should result as the setting will always make it clear which concept is intended.

Definition 5.9. If F is a distribution function, we define the *characteristic function* ϕ of F by

$$\phi(t) = \phi_F(t) = \int_{\mathbb{R}} e^{itx} \, \mathrm{d}F,$$

where this integral is taken with respect to the measure induced by the distribution function, as described in the prelude to Proposition 5.3.

Remark 5.10. Proposition 5.3 also implies that

$$\phi(t) = \int_{\mathbb{R}} e^{itx} \, \mathrm{d}F = \int_{\Omega} e^{itF} \, \mathrm{d}\mathbb{P}.$$

We have not provided any motivation for such an object; however, it will turn out to be very useful because, as we shall prove in a moment, the characteristic function uniquely defines the distribution function. The characteristic function is an interesting object of study not only because it is equivalent to the distribution function in this sense, but also because it is frequently easier to work with.

Proposition 5.11. *Let X be a random variable with characteristic function ϕ_X. Then*

i. For $\alpha, \beta \in \mathbb{R}$, we have $\phi_{(\alpha X + \beta)}(t) = e^{i\beta t}\phi(\alpha t)$.

ii. If $\mathbb{E}(X^n) < \infty$ *then*

$$\frac{\mathrm{d}^n}{\mathrm{d}t^n}\phi_X(t)|_{t=0} = i^n\mathbb{E}(X^n).$$

iii. If f *is an arbitrary distribution function, then for any* t, $|\phi(t)| \leq \mathrm{Var}(f)$*; furthermore* $\phi(0) = \mathrm{Var}(f)$*. In particular, if* $f = F_X$ *is the distribution function of a random variable, then* $|\phi(t)| \leq 1$ *and* $\phi(0) = 1$*.*

Proof. To prove i., we have

$$\begin{aligned}
\phi_{(\alpha X+\beta)}(t) &= \int_{\mathbb{R}} e^{it(\alpha X+\beta)}\,\mathrm{d}\mathbb{P} \\
&= \int_{\mathbb{R}} e^{i(\alpha t)X}e^{i\beta t}\,\mathrm{d}\mathbb{P} \\
&= e^{i\beta t}\int_{\mathbb{R}} e^{i\alpha tx}\,\mathrm{d}F(x) \\
&= e^{i\beta t}\phi(\alpha t).
\end{aligned}$$

Next,

$$\frac{\mathrm{d}^n}{\mathrm{d}t^n}\phi(t) = \frac{\mathrm{d}^n}{\mathrm{d}t^n}\int_{\mathbb{R}} e^{itx}\,\mathrm{d}F(x),$$

and it is a consequence of the Dominated Convergence Theorem that we may interchange the derivative and the integral. The reader may see [3] for the precise statement of the necessary proposition. Thus we have

$$\int_{\mathbb{R}} \frac{\mathrm{d}^n}{\mathrm{d}t^n}e^{itx}\,\mathrm{d}F(x) = \int_{\mathbb{R}} (ix)^n e^{itx}\,\mathrm{d}F(x),$$

and hence

$$\frac{\mathrm{d}^n}{\mathrm{d}t^n}\phi(t)|_{t=0} = \int_{\mathbb{R}} (ix)^n\,\mathrm{d}F(x) = i^n\mathbb{E}(X^n).$$

Finally,

$$|\phi(t)| = \left|\int_{\mathbb{R}} e^{itx}\,\mathrm{d}F(x)\right| \leq \int_{\mathbb{R}} \mathrm{d}F(x) = \lim_{x\to\infty} F(x) - \lim_{x\to-\infty} F(x) = \mathrm{Var}(F).$$

Since $e^0 = 1$, this is an equality for $\phi(0)$. □

6 Convergence of Distribution Functions

This section's function is primarily technical. Our aim is to determine when convergence in characteristic function implies convergence in distribution. It is not obvious yet why this will be of interest to us. However, we will find that the characteristic function of a sum of random variables is much easier to work with than the corresponding distribution function.

Definition 6.1. A sequence of distribution functions $\{F_n\}$ is said to converge *weakly* if $F_n(x) \to F(x)$ for each $x \in C(F) = \{x \in X : F \text{ is continuous at } x\}$. It is said to converge *up to an additive constant* to a distribution function F if for all subsequences $\{F_{n_k}\}$ and $\{F_{n_j}\}$ such that F_{n_k} converges weakly to F_1 and F_{n_j}converges weakly to F_2 we have $F_1 - F_2 = c$ for some constant c.

In most cases we will be dealing with continuous distribution functions and $C(F) = X$. We include this requirement only for completeness. As we noted in the previous section it is possible for a distribution function to be discontinuous. Distribution functions are necessarily right continuous, but in strange situations, such as when certain singletons have positive measure, continuity may fail. Setting aside these technical considerations, we begin our study of the convergence of distribution functions with the following proposition.

Proposition 6.2. *Let $\{F_n\}$ be a sequence of distribution functions. Then there exists a subsequence F_{n_k} converging weakly to some distribution function F.*

Proof. We make use of a diagonal argument. Let $\{q_i\}$ be an enumeration of the rational numbers. Consider the sequence $\{F_n(q_1)\}$. $0 \leq F_n \leq 1$ by assumption, so this is a bounded sequence of real numbers which by the Bolzano-Weierstrass Theorem has a convergent subsequence. Denote such a subsequence by $\{F_{n_{i1}}\}$. Now take a subsequence of this sequence such that $\{F_{n_{i_k 2}}(q_2)\}$ converges and label this sequence $\{F_{n_{i2}}\}$. Iterate this process to produce a diagonal sequence $\{F_{n_{ii}}\}$. For sufficiently large i this is a subsequence of some $\{F_{n_{kj}}\}$ and hence $\{F_{n_kj}(q)\}$ is convergent for each q rational.

At all rational points we define $F(x) = \lim_i F_{n_{ii}}(x)$. F is increasing since it is the limit of increasing functions. Now, we define

$$F(x) = \inf\{F(q) : q \in \mathbb{Q}, q \geq x\}.$$

It is apparent from this definition that F, now defined on $\mathbb{R}$, is increasing and right continuous. Since F is bounded over the rationals it must be bounded over the reals also because it is monotonically increasing and right continuous. Thus F is a distribution function. □

Sometimes this result is referred to as the *Helly Selection Principle*.

Now we turn our attention to the characteristic function once again. The following notion will be useful to us later.

Definition 6.3. Let F be a distribution function with characteristic function ϕ. The *integral characteristic function* $\overline{\phi}$ for F is the function

$$\overline{\phi}(t) = \int_0^t \phi(v)\,\mathrm{d}v.$$

Remark 6.4. With the setup of Definition 6.3, we have

$$\overline{\phi} := \int_{\mathbb{R}} \phi(v)\,\mathrm{d}v = \int_0^t \int_{\mathbb{R}} e^{ivx}\,\mathrm{d}F(x)\,\mathrm{d}v = \int_{\mathbb{R}} \int_0^t e^{ivx}\,\mathrm{d}v\,\mathrm{d}F(x)$$

by Fubini. Integrating, we have

$$\overline{\phi}(t) = \int_{\mathbb{R}} \frac{e^{itx} - 1}{ix} \, \mathrm{d}F(x)$$

We will need one further result to analyze convergence.

Remark 6.5. It is a standard result from analysis that if $\{a_n\}$ is a sequence of real numbers, then $a_n \to a$ if and only if for every subsequence a_{n_k} there is a further (sub)subsequence $a_{n_{k_j}}$ which converges to a. This strange characterization of convergence will be useful for proving the next proposition.

Proposition 6.6. *i. Let $\{F_n\}$ and F be distribution functions with integral characteristic functions $\{\overline{\phi_n}\}$ and $\overline{\phi}$, respectively. Then if $F_n \to F$ up to an additive constant, we also have $\overline{\phi_n} \to \overline{\phi}$ pointwise over $\mathbb{R}$.*

ii. If $\overline{\phi_n} \to \overline{g}$ pointwise over $\mathbb{R}$ then there is a distribution function F with integral characteristic function $\overline{\phi}$ such that $F_n \to F$ up to an additive constant and $\overline{g} = \overline{\phi}$.

Proof. Suppose that $F_n \to F$ up to an additive constant. Let F_{n_1} and F_{n_2} be subsequences such that $F_{n_1} \to F_1$ and $F_{n_2} \to F_2$. Following Remark 6.4, since

$$\left| \frac{e^{itx} - 1}{ix} \right| \leq \left| \frac{2}{x} \right|,$$

which is bounded and continuous away from 0, we see that as $x \to \pm\infty$

$$\frac{e^{itx} - 1}{ix} \to 0$$

Given this condition, it is a consequence of a version of the Helley-Bray Lemma, which may be found in chaper 8 of [3], that

$$\int_{\mathbb{R}} \frac{e^{itx} - 1}{ix} \, \mathrm{d}F_{n_1}(x) \to \int_{\mathbb{R}} \frac{e^{itx} - 1}{ix} \, \mathrm{d}F_1(x)$$

and the analogous result is true for F_2. Thus $\overline{\phi_{n_1}} \to \phi_1$ and $\overline{\phi_{n_2}} \to \phi_2$. We have by assumption $F_1 - F_2 = c$ for some $c \in \mathbb{R}$ which implies that the measures induced by these functions are the same and the same as that induced by F, hence

$$\int_{\mathbb{R}} \frac{e^{itx} - 1}{ix} \, \mathrm{d}F_1(x) = \int_{\mathbb{R}} \frac{e^{itx} - 1}{ix} \, \mathrm{d}F_2(x)$$

This shows that given any convergent subsequence of F_n, the corresponding subsequence of $\overline{\phi_n}$ converges, and the limits of each subsequence coincide. Now, picking an arbitrary subsequence F_{n_k}, there is a further subsequence $F_{n_{kl}}$ which does in fact converge by Proposition 6.2. Further, the limits of all such subsequences differ by only an additive constant. The above argument shows that given any subsequence $\overline{\phi_{n_k}}$ there is a further subsequence $\overline{\phi_{n_{kl}}}$ which converges,

and the limits of all such subsequences are equal. Then, in particular we have that $\overline{\phi_n(t)} \to \overline{\phi(t)}$, for all t by Remark 6.5; hence $\overline{\phi_n} \to \overline{\phi}$.

Conversely, suppose that $\overline{\phi_n} \to g$. Let F_{n_1} and F_{n_2} be subsequences such that $F_{n_1} \to F_1$ and $F_{n_2} \to F_2$, where F_1 and F_2 are distribution functions with characteristic functions $\overline{\phi_1}$ and $\overline{\phi_2}$. Then by the above $\overline{\phi_{n_1}} \to \overline{\phi_1}$ and $\overline{\phi_{n_2}} \to \overline{\phi_2}$ and we must have $\overline{\phi_1} = \overline{\phi_2} = g$ by assumption. But if this is the case we must have $F_1 - F_2 = c$, so that $F_n \to F$ up to an additive constant. □

Definition 6.7. We say that F_n converges to F *completely up to an additive constant*, $F_n \xrightarrow{c} F$, if $F_n \to F$ up to an additive constant and $\mathrm{Var}(F_n) \to \mathrm{Var}(F)$.

Theorem 6.8. *Let $\{F_n\}$ and F be bounded distribution functions with characteristic functions ϕ_n and ϕ respectively. Then*

i. If $F_n \to F$ up to an additive constant then $\phi_n \to \phi$ over $\mathbb{R}$.

ii. If $\{\phi_n\}$ converges to a function g that is continuous at zero, then $F_n \xrightarrow{c} F$ up to an additive constant for some distribution function F with characteristic function $\phi = g$.

Proof. First assume that $F_n \to F$ up to an additive constant. Then take any subsequence $\{F_m\} \subset \{F_n\}$. By Proposition 6.2 we have a (sub)subsequence $\{F_r\}$ such that $F_r \to F'$. Then $F - F' = c$ for some $c \in \mathbb{R}$. Note that e^{itx} is a bounded continuous function and hence again we apply the Helley-Bray Lemma ([3]) which yields

$$\phi_r(t) = \int_{\mathbb{R}} e^{itx} \,\mathrm{d}F_r(x) \to \int_{\mathbb{R}} e^{itx} \,\mathrm{d}F(x) = \phi(t)$$

So $\phi_n \to \phi$ by Remark 6.5.

Conversely, suppose that $\phi_n \to g$, with g continuous at zero. Since $|e^{itx}| = 1$ and F a distribution function implies that the measure induced by F is a finite measure $|\phi_n|$ is bounded by some constant. Then we can apply the dominated convergence theorem which yields:

$$\overline{\phi_n(t)} = \int_0^t \phi_n(v) \,\mathrm{d}v \to \int_0^t g(v) \,\mathrm{d}v = \overline{g}$$

By proposition we then have that $F_n \to F$ up to an additive constant for some distribution function F such that $\overline{\phi_f} = \overline{g}$. In particular

$$\frac{1}{t} \int_{\mathbb{R}} \phi(v) \,\mathrm{d}v = \frac{1}{t} \int_{\mathbb{R}} g(v) \,\mathrm{d}v$$

Taking the limit as $t \to 0$ an easy continuity argument shows that $\phi(0) = g(0)$. Since g is continuous at 0 we have $g(v) = g(0) + h(v)$ where h is a function such that

$$\lim_{v \to 0} h(v) = 0$$

Then

$$\frac{1}{t}\int_0^t g(v)\,\mathrm{d}v = \frac{1}{t}\int_0^t g(0) + h(v)\,\mathrm{d}v = g(0) + \frac{1}{t}\int_0^t h(v)\,\mathrm{d}v$$

And we have

$$\left|\frac{1}{t}\int_0^t h(v)\,\mathrm{d}v\right| \leq \max_{v\in(0,t)} |h(v)|$$

which tends to zero as t does by continuity. Hence we have $\phi(0) = g(0)$. But note

$$\mathrm{Var}(F_n) = \phi_n(0) \to g(0) = f(0) = \mathrm{Var}(F)$$

So $F_n \xrightarrow{c} F$ up to an additive constant. □

Corollary 6.9. (Lévy Continuity Theorem) *Suppose $\{F_n\}$ is a sequence of distribution functions of random variables with characteristic functions $\{\phi_n\}$.*

i. If $F_n \to F$, where F is the distribution function of some random variable with characteristic function ϕ, then $\phi_n \to \phi$.

ii. If $\phi_n \to g$, then $F_n \to F$ for some distribution function of a random variable F continuous at zero with characteristic function $\phi = g$.

Proof. This is exactly the previous theorem, with the additional restriction that all the distribution functions considered arise from random variables. The fact that the F in the second part of the theorem is actually the distribution function of some random variable follows from some unilluminating and technical results presented in [3]. We do not present those details here. □

The Lévy Continuity Theorem is precisely the result we need to prove the Central Limit Theorem. It is basically a technical tool and has little interest to us otherwise.

7 The Central Limit Theorem

Now we turn our attention to the Central Limit Theorem itself. We have described all the necessary machinery and now it only remains to build our intuition about the normal distribution—or the "bell curve". This distribution is interesting to us because it arises frequently. Although we expect it is familiar to many readers we define it here for convenience.

Definition 7.1. The *normal distribution* with mean μ and standard deviation σ is the distribution function given by

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\,\mathrm{d}t$$

We often denote this distribution function by $N(\mu, \sigma^2)$.

Note that we refer to integrand in the definition as the *probability density function*. Speaking imprecisely,

$$e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

is the probability that the random variable we are considering takes a value in the interval $(x, x + \mathrm{d}x)$.

Since the normal distribution arises so frequently it seems there should be some intuitive way to understand its appearance, and there is, at least qualitatively. If we are analyzing an ability distributed "by chance" we expect that most individuals possess close to the average amount of this ability. We expect the probability that an individual has a larger or smaller share of ability to taper off rapidly. That is, the probability density for a continuous random variable should have a shape similar to the bell curve but it is not obvious a Gaussian is the proper variation of this shape. This only becomes apparent *after* we establish the Central Limit Theorem.

Remark 7.2. It is possible to show that the characteristic function of the normal distribution with mean 0 and variance 1 is $e^{-\frac{t^2}{2}}$. We do not prove this here because it is well known; the proof may be found in many texts (for instance, [4]) and it adds little to our discussion.

As our remarks above make clear, the motivation for the Central Limit Theorem does not evolve from studying the normal distribution itself. Rather, it comes from considering sums of random variables. Often such sums are of great interest when we are considering complex systems. For instance, if we use the random variables indexed by integers X_k to describe the outcome of a simple test such as a coin flip, the sum

$$S_n = \sum_{k=1}^{n} X_k$$

describes the results of the first n tests. Unfortunately, it is difficult a priori to say much about the sum of two random variables. Of course if X and Y are random variables, so is $X + Y$ by Proposition 2.5, and we can even find the distribution of this random variable with the method of convolutions (see [3]). But things become hopelessly complicated when we are considering more than two variables. Consequently, we are interested in simplifying our problem.

We will take the example of a repeated experiment such as the coin flips referred to above as our model and attempt to abstract characteristics from it. First we notice that the value of each random variable is independent of the others in the sense that if we flip our coin once, this has no effect on the value of our next flip. If we suppose that A and B are the events that X_1 and X_2 are heads respectively then we want to compute the probability $\mathbb{P}(A \cap B)$. Since these random variables are independent in the sense we have described above, we expect that we can somehow compute this probability in terms of $\mathbb{P}(A)$ and $\mathbb{P}(B)$.

Indeed, in the case of the coinflip it is easy to see that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. We formalize this natural sort of independence in the following definition.

Definition 7.3. We say that two events $A_1, A_2 \in \mathcal{F}$ are *independent* if $\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(A_1)\mathbb{P}(A_2)$. If instead we have events $\{A_i : i \in I\}$ then we say they are independent provided for any finite set $\{A_{j_1}, \ldots, A_{j_n}\} = J \subset I$ we have $\mathbb{P}(A_{j_1} \cap A_{j_2} \cap \ldots \cap A_{j_n}) = \mathbb{P}(A_{j_1})\mathbb{P}(A_{j_2}) \ldots \mathbb{P}(A_{j_n})$. This leads us to a natural characterization of independence of classes of events $\mathcal{C}_1, \ldots, \mathcal{C}_n \subset \mathcal{A}$ where $\mathcal{A}$ is some σ-algebra. We say these classes are *independent* if for any choice of $A_j \in \mathcal{C}_j$ the A_j's are independent.

Remark 7.4. It seems more natural to define independence as a pairwise property. That is, a first guess might have been that $\{A_i : i \in I\}$ are independent events if $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$ for $i \neq j$. However, it can be shown that this characterization does not possess all of the properties we want.

Independence of σ-algebras provides us a natural approach to defining independence of random variables. This approach relies on the fact that if X is a random variable and $\mathcal{F}$ is a σ-algebra then $X^{-1}(\mathcal{F})$. We leave proof of this fact to the reader.

Definition 7.5. Let $\mathcal{B}$ be the Borel σ-algebra over a space Ω. Note that $X_j^{-1}(\mathcal{B})$ is a σ-algebra. We say that random variables X_j are independent if the σ-algebras $X_j^{-1}(\mathcal{B})$ are independent.

Independence is the most important characteristic of the random variables we have gleaned from our model. However, there is another important characteristic. Note that if we are considering multiple copies of the same experiment our random variables must have the same distribution function.

Definition 7.6. We say that random variables X_1 and X_2 are *identically distributed* if they have the same distribution functions.

Although it appears that we lose a lot by making these two assumptions they are so natural that we will hardly notice the difference. But even these simplifications do not simplify the distributions of sums of random variables appreciably. However, they make the characteristic functions of these sums extremely simple. This observation is what should motivate the study of the material in Section 6. The material in this section is in some sense antecedent to the technical results we have already presented. However, we included those results first because we did not want to interrupt the chain of intuition we present here.
Now we are almost ready to prove the Central Limit Theorem. Our strategy will rely on showing that the characteristic function of S_n approaches the characteristic function of $N(0, 1)$ as n approaches infinity. To make dealing with characteristic functions easier we first want to show that they are as nice as possible; in particular, they are analytic.

Proposition 7.7. *Suppose X is a random variable such that $\mathbb{E}(X^n) < \infty$ for some $n \in \mathbb{N}$. Let ϕ be the characteristic function of X and let $m^{(k)} = \mathbb{E}(X^k)$. Let f be the characteristic function of X, then we have*

$$\phi(t) = \sum_{k=0}^{n} \frac{m^{(k)}}{k!}(it)^k + o(t^n)$$

Proof. By Proposition 5.11 we have

$$\frac{\mathrm{d}^n}{\mathrm{d}t^n}\phi(t)|_{t=0} = i^n\mathbb{E}(X^n) = i^n m^{(n)}$$

Hence this proposition will follow from Taylor's Theorem provided we can show ϕ is holomorphic. Consider

$$\lim_{t_0\to 0}\frac{f(t+t_0)-f(t)}{t_0} = \lim_{t_0\to 0}\int e^{itx}\frac{(e^{it_0x}-1)}{t_0}\,\mathrm{d}F(x)$$

We know that

$$\lim_{t_0\to 0}\frac{e^{it_0x}-1}{t_0} = ix$$

since e^{itx} is holomorphic. Hence, in particular if we take t_0 sufficiently small we can assume that

$$\left|\frac{f(t+t_0)-f(t)}{t_0} - ix\right| \leq C$$

for some $C \in \mathbb{R}$. By the triangle inequality,

$$\left|\frac{f(t+t_0)-f(t)}{t_0}\right| \leq |x| + C$$

which is integrable, hence by the Dominated Convergence Theorem we take the limit inside the integral above. But then, in particular the limit exists so f is holomorphic. Thus f is analytic, so there is an open neighborhood of zero about which

$$\phi(t) = \sum_{k=0}^{\infty}\frac{\mathrm{d}^n}{\mathrm{d}t^n}\phi|_{t=0}\frac{t^k}{k!} = \sum_{k=0}^{n}\frac{m^{(k)}}{k!}(it)^k + o(t^n)$$

□

This is the final piece of machinery necessary. Now we state our main theorem.

Theorem 7.8. (Classical Central Limit Theorem) *Let $X_1, \ldots, X_n$ be independent, identically distributed random variables such that $\mathbb{E}(X_1)$ and $\sigma^2(X_1) = \sigma^2$ are finite. Also allow*

$$S_n = \sum_{j=1}^{n} X_j \text{ and } \overline{X_n} = \frac{1}{n}\sum_{j=1}^{n} X_j$$

Then we have

$$\frac{S_n - \mathbb{E}(S_n)}{\sigma(S_n)} = \frac{\sqrt{n}[\overline{X_n} - m^{(1)}]}{\sigma} \xrightarrow[n\to\infty]{d} Z \sim N(0,1)$$

Remark 7.9. Before we prove the theorem we briefly demonstrate the first equivalence using the fact that the random variables are independent and identically distributed.

$$\begin{aligned}
\frac{S_n - \mathbb{E}(S_n)}{\sigma(S_n)} &= \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n \mathbb{E}(X_i)}{\sqrt{\mathbb{E}(S_n^2) - \mathbb{E}(S_n)^2}} \\
&= \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n \mathbb{E}(X_i)}{\sqrt{\mathbb{E}((\sum_{i=1}^n X_i)^2) - \mathbb{E}(\sum_{i=1}^n X_i)^2}} \\
&= \frac{\sum_{i=1}^n X_i - \mathbb{E}(X_i)}{\sqrt{\mathbb{E}(\sum_{i=1}^n \sum_{j=1}^n X_i X_j) - (\sum_{i=1}^n \mathbb{E}(X_i))^2}} \\
&= \frac{\sum_{i=1}^n X_i - \mathbb{E}(X_i)}{\sqrt{\sum_{i\neq j} \mathbb{E}(X_i)\mathbb{E}(X_j) + \sum_{i=1}^n \mathbb{E}(X_i^2) - \sum_{i,j} \mathbb{E}(X_i)\mathbb{E}(X_j)}} \\
&= \frac{\sum_{i=1}^n X_i - \mathbb{E}(X_i)}{\sqrt{\sum_{i=1}^n \mathbb{E}(X_i^2) - \mathbb{E}(X_i)^2}} \\
&= \frac{\sum_{i=1}^n X_i - \mathbb{E}(X_i)}{\sqrt{n\sigma^2}} \\
&= \frac{\sqrt{n}(\overline{X_n} - m^{(1)})}{\sigma}
\end{aligned}$$

Proof. First we "center" the random variables and normalize their variance by setting

$$Y_i = \frac{X_i - m^{(1)}}{\sigma}$$

It is easy to check that $\mathbb{E}(Y_i) = 0$ and $\sigma^2(Y_i) = 1$ Now, suppose f_1 is the characteristic function of Y_1 (note that it is also the characteristic function of Y_i since the Y_i are identically distributed and the characteristic function is defined in terms of the distribution function). Applying Proposition 7.1

$$f_1(t) = 1 + \frac{\mathbb{E}(Y_1)}{1!}(it) + \frac{\mathbb{E}(Y_1^2)}{2!}(it^2) + o(t^2) = 1 - \frac{t^2}{2} + t^2 o(1)$$

Then we have

$$f_{\frac{S_n - \mathbb{E}(s_n)}{\sigma(S_n)}}(t) = f_{\sum_{i=1}^n \frac{X_i - \mathbb{E}(X_i)}{\sigma\sqrt{n}}}(t) = f_{\sum_{i=1}^n \frac{Y_i}{\sqrt{n}}}(t) = f_1^n\left(\frac{t}{\sqrt{n}}\right)$$

where the last equality follows from Proposition 5.11.

$$f_1^n\left(\frac{t}{\sqrt{n}}\right) = \left[1 - \frac{t^2}{2n} + \frac{t^2}{n}o(1)\right]^n$$

Thus, using a well known limit we clearly have

$$\lim_n f_1^n\left(\frac{t}{\sqrt{n}}\right) = \lim_n \left[1 - \frac{t^2}{2n} + \frac{t^2}{n}o(1)\right]^n = \lim_n \left[1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right]^n = e^{-\frac{t^2}{2}}$$

But note that this is the characteristic function for the normal distribution. Hence, our above random variable converges in distribution to $N(0,1)$ by the Lévy Continuity Theorem. □

Remark 7.10. Note that we tacitly used the assumption our random variables are independent. This assumption allows us to claim that

$$f_{X_1+X_2} = f_{X_1} f_{X_2}$$

This natural condition might otherwise have failed.

There are a number of interesting situations the Central Limit Theorem can be exploited to provide information about. One basic but extremely important example is the random walk.
Suppose at time $t = 0$ we place a particle at the origin of the real line. Suppose further that at each subsequent integer time the particle moves randomly a distance of one in the positive or negative direction. If we allow X_k to be a random variable with distribution $\mathbb{P}(X_k = 1) = \mathbb{P}(X_k = -1) = \frac{1}{2}$ then X_k encodes the possible movements of this particle at time k. More generally we could assume that $\mathbb{P}(X_k = 1) = p$ and $\mathbb{P}(X_k = -1) = 1 - p$; any variable of this form is called a *Bernoulli Random Variable*. However, this would complicate our application of the Central Limit Theorem so we only consider the case where $p = \frac{1}{2}$. Under this assumption if we denote the position of the particle at time n by the random variable S_n we have

$$S_n = \sum_{k=1}^{n} X_k$$

Notice that

$$\mathbb{E}(X_k) = (1)(\frac{1}{2}) + (-1)(\frac{1}{2}) = 0$$

hence each variable has mean zero, and $\mathbb{E}(X_k^2) = (1)(1)$ hence

$$\sigma^2(X_k) = 1 \text{ so } \sigma(X_k) = 1$$

and X_k has zero variance also. Apparently, the X_k's for $t \in \mathbb{N}$ are also identically distributed. Then the Central Limit Theorem tells us that

$$\frac{\sqrt{n}\frac{1}{n}S_n}{1} = \frac{S_n}{\sqrt{n}} \xrightarrow[n\to\infty]{d} Z \sim N(0,1)$$

Ultimately, the Central Limit Theorem is important because it explains why

the normal distribution appears so frequently and is such a useful statistical tool. It is one of the most important results of probability theory, and the fact we can prove it using measure theory provides important posterior justification for developing this machinery.
In fact, numerous other variants of the Central Limit Theorem exist but we have opted to prove only the classical version. Probably the most significant other variant is the Lindeberg Central Limit Theorem which is discussed extensively in [3]. This theorem is more general, and implies the classical variant; however, the proof requires a number of additional technical lemmas and a number of results from complex analysis. It is not nearly so enlightening or accessible as the one we have presented here. If the reader is interested in this more general theorem, we invite him or her to explore the proof in [3] and refer him or her to [2] for a reference on Complex Analysis. Indeed, the curious reader may find numerous analogues of the Central Limit Theorem because it is so important. Given a new setting, an analogue of this theorem is one of the first things a Probabilist is interested in establishing.

8 Acknowledgements

I owe a great debt to the works of C. S. Kubrusly and G. G. Roussas. The first three sections of this paper are based heavily on [1] and much of the remaining is based on [3]. The important theorem showing that sums and products of measurable functions is due to [5]. In addition, although [5] appears rarely in the text, it was an invaluable resource for the author's developing understanding of measure theory and probability. Finally, I would like to thank my mentor, Zachary Madden, for providing much useful advice and help editing this paper. It should go without saying that any errors retained in the final version are my own.

References

[1] C. S. Kubrusly, *Measure Theory: A First Course*, Elsevier Academic Press, Burlington, MA, 2007.

[2] S. Lang, *Complex Analysis*, Springer-Verlag, New York, NY, 1999.

[3] G. G. Roussas, *An Introduction to Measure-Theoretic Probability*, Elsevier Academic Press, Burlinton, MA, 2005.

[4] J. S. Rosenthal, *A First Look at Rigorous Probability Theory*, World Scientific Publishing Company, Singapore, 2006.

[5] W. Rudin, *Real and Complex Analysis*, McGraw-Hill Book Company, New York, 1966.