John Nachbar
Washington University in St. Louis
January 10, 2018

Basic Lebesgue Measure Theory[1]

1 Introduction.

This is an introduction to Measure Theory. I focus on the motivation for and the definition of Lebesgue measure on $[0, 1]$; separate notes provide a brief introduction to Lebesgue integration. My treatment largely follows Royden (2010), but I also draw on Kolmogorov and Fomin (1970), Stein and Shakarchi (2005), and Tao (2011), among others.

Although Measure Theory has a deserved reputation for being subtle, its underlying motivation is straightforward. Lebesgue measure can by viewed as a natural generalization of length to sets that are more complicated than intervals or finite unions of intervals. Viewed as a probability, Lebesgue measure is the appropriate generalization of *equally likely*, as captured by the uniform distribution, to complicated events.

As an illustration of what is at issue, Section 3 examines the Law of Large Numbers for tosses of a fair coin. Recall that the Strong Law of Large Numbers says that, with probability one, the empirical frequency of heads converges to 1/2. In contrast, the Weak Law of Large Numbers says, roughly, that with high probability the empirical frequency is close to 1/2. The Strong Law implies the Weak Law but not conversely.

The Strong Law uses measure theoretic ideas. To understand why, note that in order to state that, "with probability one, the empirical frequency of heads converges to 1/2," one has to specify exactly what set has probability one. In the present context, the set of outcomes is the set of all possible realizations of coin tosses, which are infinite strings of the form HTTHHH.... One has to compute the probability of the subset of such strings for which the empirical frequency converges to 1/2. This subset turns out to be extremely complicated. In contrast, the Weak Law does not require measure theoretic ideas because the sets involved are finite unions of intervals.

2 An Overview of Terminology.

Consider first a *probability space*, $(X, \mathcal{A}, \mu)$. X is the (non-empty) set of underlying *states of the world* (e.g., whether it is raining). $\mathcal{A}$ is the set of possible *events*, which

are subsets of X: $\mathcal{A} \subseteq \mathbb{P}(X)$. μ is a *probability measure*: given an event $A \in \mathcal{A}$, the probability of A is $\mu(A)$. Since μ takes sets as its argument, it is called a *set function*, which should not be confused with a set-*valued* function (also known as a *correspondence*).

For $(X, \mathcal{A}, \mu)$ to qualify as a probability space, $\mathcal{A}$ and μ must satisfy certain conditions.

I assume throughout that $\mathcal{A}$, the set of events, is non-empty. I assume in addition that $\mathcal{A}$ is a *σ-algebra* (also known as a σ-field), meaning that $\mathcal{A}$ satisfies the following requirements.

1. If $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$.
2. If $\{A_t\} \subseteq \mathcal{A}$ is countable then $\bigcup_t A_t \in \mathcal{A}$.

By DeMorgan's Laws, countable intersections of sets in $\mathcal{A}$ are also in $\mathcal{A}$. Since $A \cup A^c = X$, it is automatically true that $X \in \mathcal{A}$, hence we also have $X^c = \emptyset \in \mathcal{A}$.

I assume that μ has the following properties.

1. *Probabilities are non-negative:* $\mu : \mathcal{A} \to \mathbb{R}_+$.
2. *Some state is realized*: $\mu(X) = 1$.
3. *Probability is countably additive:* If $\{A_t\} \subseteq \mathcal{A}$ is countable (finite or countably infinite) and pairwise disjoint ($A_s \cap A_t = \emptyset$ for any s, t) then,

$$\mu\left(\bigcup_t A_t\right) = \sum_t \mu(A_t).$$

   Countable additivity is also called *σ-additivity.*

These conditions imply that $\mu(\emptyset) = 0$ and that no event can have a probability larger than one.

Weaker criteria on $\mathcal{A}$ and μ are that $\mathcal{A}$ is an algebra (rather than a σ-algebra) and that μ is *finitely additive* (rather than countably additive). $\mathcal{A}$ is an algebra iff (a) if $A \in \mathcal{A}$ implies $A^c \in \mathcal{A}$ (this is the same property as above) and (b) if $\{A_t\} \subseteq \mathcal{A}$ is *finite* then $\bigcup_t A_t \in \mathcal{A}$. μ is finitely additive iff whenever $\{A_1, \ldots, A_T\}$ is a finite set of disjoint sets in $\mathcal{A}$ then,

$$\mu\left(\bigcup_{t=1}^{T} A_t\right) = \sum_{t=1}^{T} \mu(A_t).$$

The benchmark example of a probability space is $([0,1], \mathcal{B}, \lambda)$, which is the probability space associated with the uniform distribution on $[0,1]$. The σ-algebra $\mathcal{B}$, called the *Borel σ-algebra*, is the smallest σ-algebra that contains all intervals in $[0,1]$ (i.e., $\mathcal{B}$ is contained in any σ-algebra that contains all intervals). The measure λ is Lebesgue measure. For any interval $A \subseteq [0,1]$ with end points $a \leq b$,

$\lambda(A) = b - a$, which is the probability of A under the uniform distribution. But λ also assigns probability to events that are much more complicated than intervals or finite unions of intervals.

A natural question is: why is the set of events $\mathcal{B}$ taken to be a proper subset of the power set $\mathbb{P}([0,1])$? Why not just use all of $\mathbb{P}([0,1])$? The reason, discussed in detail starting in Section 4.1, is that λ cannot be extended to all of $\mathbb{P}([0,1])$ without losing finite, let alone countable, additivity.

Finally, a *measure space* $(X, \mathcal{A}, \mu)$ generalizes a probability space to allow $\mu(X) \neq 1$ ($\mu(X)$ is allowed to be "$+\infty$").

3 Some Motivation.

3.1 The Uniform Distribution.

Consider the uniform distribution on $[0,1]$. If $A \subseteq [0,1]$ is an interval with endpoints $a \leq b$, then the probability of A, written Prob$[A]$, equals the length of A, written $\ell(A)$: Prob$[A] = \ell(A) = b - a$. This definition of probability extends easily to events that are finite unions of intervals. I want to extend this notion of probability to more complicated events.

Given a set $A \subseteq [0,1]$, define the indicator function for A to be,

$$1_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases}$$

Define the probability of an event A to be

$$\text{Prob}[A] = \int_0^1 1_A(x)\, dx,$$

where $\int f(x)\, dx$ denotes the standard Riemann integral of introductory calculus.

Note that under this construction, Prob$[A]$ is defined iff the associated Riemann integral exists. The basic problem with this construction is that there are events A for which the Riemann integral does not exist. The next two subsections develop examples.

3.2 The Dirichlet Function.

Let A be the set of rational numbers in $[0,1]$. For this A, the indicator function 1_A is sometimes called the Dirichlet function.

As defined above, in terms of the Riemann integral, Prob$[A]$ is not defined. In particular, any (non-degenerate) interval contains both rational and irrational numbers. This implies that the upper Riemann integral for 1_A is 1 and the lower Riemann integral is 0: 1_A is not Riemann integrable.

The next subsection discusses an important example that exhibits similar behavior.

3.3 The Strong Law of Large Numbers.

Imagine an infinite string of coin flips, which I represent as an element of $X = \{0,1\}^\omega$, where 1 represents H (heads) and 0 represents T (tails). Thus, I represent the string $HTHH\ldots$ as $(1,0,1,1,\ldots)$. For each $x \in X$, define

$$S_T(x) = \frac{1}{T}\sum_{t=1}^{T} x_t.$$

$S_T(x)$ is thus the fraction of H in the first T flips. Let

$$A = \{x \in X : \lim_{T\to\infty} S_T(x) = 1/2\}.$$

A is the set of strings of coin flips for which the frequency of H converges to $1/2$. A basic intuition in probability theory is that if the coin is "fair" (roughly: the probability of H on any given flip is $1/2$, and the outcome of the flips on any finite set of dates has no effect on the probabilities at other dates) then,

$$\text{Prob}[A] = 1.$$

In the language of probability theory, this says that S_T converges *almost surely* to S, where $S(x) = 1/2$ for all $x \in X$.[2] This is a special case of the *Strong Law of Large Numbers (SLLN)*.

How should we formalize SLLN? In particular, how should we define a probability on X to formalize the idea that an element of X is a realization of an infinite number of flips of a fair coin? One approach, perhaps the most obvious one, is to note that any element of x can be identified with a binary expansion of a number in $[0,1]$: the string $x = (x_1, x_2, x_3, \ldots)$ becomes the binary expansion $0.x_1x_2x_3\ldots$, which represents the number,

$$\frac{x_1}{2} + \frac{x_2}{4} + \frac{x_3}{8} + \ldots.$$

As usual, there are two different expansions for some numbers (e.g., the binary expansion of $1/2$ is both $0.1000\ldots$ and $0.0111\ldots$). I discuss this issue briefly in Remark 1 below; ignore it, for the moment.

Now consider the uniform distribution on $[0,1]$. It is easy to confirm that the probability is $1/2$ that the first term in the binary expansion of x is 1 (i.e., H). And the probability is $1/4$ that the first two terms in the binary expansion of x are, in order, 1 and 0 (i.e., HT). And so on. In general, one can confirm that for any finite set of terms (not necessarily consecutive) in the binary expansion of x, the probabilities are the same as for the corresponding flips of a fair coin. The uniform distribution on $[0,1]$ thus provides a formalization of probability for flips of a fair

[2] Measure theory and probability theory use slightly different terminology for the same concepts. In particular, almost surely in probability theory translates to *almost everywhere* in measure theory.

coin. The important point is that the uniform distribution assigns probability not only to finite strings of coin flips, something we can easily do without going through the trouble of mapping coin flips to the $[0, 1]$ interval, but also to infinite strings of coin flips.

Remark 1. Different strings of coin flips can map to the same number in [0,1]. For example, THHH... and HTTT... both map to 1/2. This complication is inessential in two respects. First, as already noted, the construction works perfectly for finite strings of coin flips. Second, the set of numbers in [0,1] for which the problem occurs is $\{0, 1, 1/2, 1/4, 3/4, 1/8 \dots\}$, which is countable, and countable sets have zero probability (Theorem 11 in Section 5.2). □

Therefore, reinterpret the coin flip experiment with $X = [0, 1]$, so that $A \subseteq [0, 1]$. We again want to show that $\text{Prob}[A] = 1$. The problem is that, with this probability model for infinite strings of coin flips, $\text{Prob}[A]$, the probability of the frequency of H converging to 1/2, is undefined using Riemann integration: for any non-degenerate interval, there are points in A (points for which the corresponding empirical frequency of H converges to 1/2) and points not in A. Hence the upper Riemann integral is 1 and the lower Riemann integral is 0. The situation is very similar to what happened in Section 3.2. Stating, and proving, even this simple version of SLLN requires new ideas about how to assign probability to sets.

3.4 Comparison with the Weak Law of Large Numbers.

One can avoid measure theoretic ideas entirely and still get a form of the Law of Large Numbers by using a weaker convergence criterion.

For each $\varepsilon > 0$, let A_T^ε be the set of points x for which $S_T(x)$ is within ε of 1/2. In notation,

$$A_T^\varepsilon = \{x \in [0, 1] : S_T(x) \in N_\varepsilon(1/2)\}.$$

For any T and any $\varepsilon > 0$, $I_{A_T^\varepsilon}$ is a finite union of intervals. Therefore, one can define

$$\text{Prob}[A_T^\varepsilon]$$

without appealing to measure theoretic ideas. One can show, again without appealing to measure theoretic ideas, that for any $\varepsilon > 0$,

$$\lim_{T\to\infty} \text{Prob}[A_T^\varepsilon] = 1.$$

One says that function S_T converges in *probability* to S, where, again, $S(x) = 1/2$ for all $x \in [0, 1]$. This is a special case of the *Weak Law of Large Numbers* (WLLN). Roughly, the almost sure convergence of SLLN says that the probability that S_T converges to S *is* 1, while the convergence in probability of WLLN says that the probability that S_T is *close* to S *converges* to 1. Put differently, SLLN is the probability of a limit while WLLN is the limit of a probability.

To understand why convergence in probability is relatively weak, consider another, artificial, example. Define sets B_t by,

$$\begin{aligned} B_1 &= [0, 1/2], \\ B_2 &= [1/2, 1], \\ B_3 &= [0, 1/4], \\ B_4 &= [1/4, 1/2], \\ B_5 &= [1/2, 3/4], \\ &\dots \end{aligned}$$

For each T, define

$$f_T(x) = \begin{cases} 0 & \text{if } x \in B_T, \\ 1 & \text{if } x \notin B_T. \end{cases}$$

Let f be the constant function $f(x) = 1$. Then f_T converges to f in probability but *not* almost surely. In fact, letting

$$B = \{x \in [0, 1] : f_T(x) \to f(x)\},$$

one has

$$B = \emptyset,$$

hence

$$\text{Prob}[B] = 0.$$

Thus, f_T converges to f in probability even though there isn't even one point x for which $f_T(x)$ converges to $f(x)$. Instead, one has that, for all x, $f_T(x)$ is close to (in fact, equal to) $f(x)$ for "most" T.

4 Lebesgue Outer Measure on $[0, 1]$.

I now begin the construction of Lebesque measure, which will solve the problems illustrated in Section 3. Throughout, I focus on the metric space $[0, 1]$ (with the standard Euclidean metric). The construction can be extended to $\mathbb{R}^N$ with only a few complications, mostly stemming from the fact that $\mathbb{R}^N$, unlike $[0, 1]$, is unbounded.

If A is an interval, meaning that $A = [a, b]$, (a, b), $[a, b)$ or $(a, b]$, with $a, b \in [0, 1]$, then the length of A is denoted $\ell(A) = b - a$. A degenerate interval, which is just a point, has length 0. The goal is to define length for more complicated sets. For reasons that I discuss in Section 4.1, this turns out to be highly non-trivial.

For the moment, take a naive approach. It is natural to approximate the "length" of a set A by covering A with a countable set of open intervals and then adding up the lengths of those intervals. This gives an upper bound on the possible length of A. The "outer measure" of A is the smallest such upper bound, the infimum over all countable sets of open intervals covering A.

Formally, for any non-empty set $A \subseteq [0,1]$, consider a countable set $\mathcal{I}$ of open intervals in $[0,1]$ that cover A:

$$A \subseteq \bigcup_{I \in \mathcal{I}} I.$$

Note that the interval $[0,1]$ is open since the metric space is $X = [0,1]$. Similarly the intervals $[0,b)$ and $(a,1]$ are open for any $0 \leq a < b \leq 1$. (Alternative language is that an interval such as $[0,b)$ is *relatively* open.)

Let $\mathcal{I}$ be any countable set of open intervals. Since $\ell(I) \geq 0$ for every I, the infinite sum $\sum_{I \in \mathcal{I}} \ell(I)$ is well defined (but it could be infinite), and in particular it is independent of the order in which the summation is taken over $\mathcal{I}$.

Definition 1. *For any non-empty set $A \subseteq [0,1]$, the* Lebesgue outer measure of *A, written $\lambda^*(A)$, is defined to be the infimum of $\sum_{I \in \mathcal{I}} \ell(I)$, where the infimum is taken over all countable sets $\mathcal{I}$ of open intervals that cover A. Set $\lambda^*(\emptyset) = 0$.*

Here are a few preliminary observations.

- For any $A \subseteq [0,1]$, $\lambda^*(A) \geq 0$.
- For any $A \subseteq [0,1]$, $\lambda^*(A) \leq 1$, since $A \subseteq [0,1]$, and $[0,1]$ is an open interval.
- For any $A, B \subseteq [0,1]$, if $A \subseteq B$ then $\lambda^*(A) \leq \lambda^*(B)$, since any $\mathcal{I}$ that covers B also covers A (but not necessarily vice versa). This property is sometimes called *monotonicity*.

Theorem 1. *λ^* is* countably sub-additive*: for any countable set $\{A_t\}$ of subsets of $[0,1]$, letting $A = \bigcup_t A_t$,*

$$\lambda^*(A) \leq \sum_t \lambda^*(A_t).$$

Proof. Fix $\varepsilon > 0$. By definition of λ^*, for each t there is a countable cover of A_t by open intervals, call this cover $\mathcal{I}_t$, such that

$$\sum_{I \in \mathcal{I}_t} \ell(I) < \lambda^*(A_t) + 2^{-t}\varepsilon.$$

Then $\mathcal{I} = \bigcup \mathcal{I}_t$ is a countable cover of A by open intervals. Then, by definition of λ^*,

$$\lambda^*(A) \leq \sum_{I \in \mathcal{I}} \ell(I) \leq \sum_t \sum_{I \in \mathcal{I}_t} \ell(I) < \sum_t \lambda^*(A_t) + \varepsilon.$$

Since ε was arbitrary, this implies the result.

As an aside, note that it is possible that $\sum_t \lambda^*(A_t) = \infty$. In this case, the result holds trivially. ■

Remark 2. It is not hard to think up examples in which a set function might be *super*-additive rather than sub-additive. For example, suppose $X = \{a, b\}$ and $\mu(a) = \mu(b) = 1/4$ while $\mu(\{a, b\}) = 1$. An economic interpretation might be that μ gives total output when inputs are either a, b, or $\{a, b\}$, and there are increasing returns from using a and b together rather than separately. □

A consequence of countable sub-additivity is the following fact.

Theorem 2. *If $A \subseteq [0, 1]$ is countable then $\lambda^*(A) = 0$.*

Proof. For any $x \in [0, 1]$, $\lambda^*(\{x\}) = 0$ since the singleton set $\{x\}$ can be covered by an open interval of arbitrarily small length. For a countable set A, $\lambda^*(A) = 0$ then follows by countable sub-additivity. ■

Example 1. $\lambda^*(\mathbb{Q} \cap [0, 1]) = 0$. □

If A is an interval then it seems obvious that $\lambda^*(A) = \ell(A)$. In particular, if A is an open interval, then $\mathcal{I} = \{A\}$ is a (trivial) cover of A by a single open interval and hence $\lambda^*(A) \leq \ell(A)$. It seems equally obvious that no other $\mathcal{I}$ can yield a smaller approximation to $\lambda^*(A)$, and hence $\lambda^*(A) = \ell(A)$. This is correct but actually demonstrating it requires work. For intuition as to why this direction is not immediate, note that if our space is $\mathbb{Q} \cap [0, 1]$ rather than $[0, 1]$ then, as in Example 1, the outer measure of the interval $\mathbb{Q} \cap [0, 1]$ is 0, not 1. Thus, the fact that we are working in $\mathbb{R}$ rather than $\mathbb{Q}$ must play a role in the proof.

Theorem 3. *For any interval $A \subseteq [0, 1]$, $\lambda^*(A) = \ell(A)$.*

Proof. Let A be an interval with endpoints $a \leq b$. If $a = b$, so that $A = \{a\}$, then, by Theorem 2, $\lambda^*(A) = 0 = b - a = \ell(A)$. Henceforth, assume $b > a$. I prove first that $\lambda^*(A) \leq b - a$ and then that $\lambda^*(A) \geq b - a$.

1. $\lambda^*(A) \leq b - a$.

   For any $\varepsilon > 0$, $A \subseteq (a - \varepsilon/2, b + \varepsilon/2) \cap [0, 1]$, hence, taking $\mathcal{I} = \{(a - \varepsilon, b + \varepsilon) \cap [0, 1]\}$, $\lambda^*(A) \leq b - a + 2\varepsilon$. Since $\varepsilon > 0$ was arbitrary, $\lambda^*(A) \leq b - a$, as was to be shown.

2. $\lambda^*(A) \geq b - a$.

   (a) Suppose that $A = [a, b]$. From the definition of λ^*, for any $\varepsilon > 0$, there is a set $\mathcal{I}$, a countable cover of A via open intervals, such that

   $$\sum_{I \in \mathcal{I}} \ell(I) < \lambda^*(A) + \varepsilon.$$

   Intuitively, it must be the case that $b - a = \ell(A) \leq \sum_{I \in \mathcal{I}} \ell(I)$. I verify this below. Assuming this is true, then

   $$b - a < \lambda^*(A) + \varepsilon.$$

Since this holds for any $\varepsilon > 0$, the result follows.

It remains to show that $b - a \leq \sum_{I \in \mathcal{I}} \ell(I)$. By Heine-Borel, A is compact. This is where the proof exploits the fact that we are working in $\mathbb{R}$ rather than $\mathbb{Q}$. Since A is compact, there is a finite subset $\hat{\mathcal{I}} \subseteq \mathcal{I}$ that also covers A.

Since $\hat{\mathcal{I}}$ covers A, there is an $I \in \hat{\mathcal{I}}$ such that $a \in I$. Denote this I as I_1, with endpoints $a_1 < b_1$ (the inequality must be strict since I_1 is open).

If $b \in I_1$, then $b \leq b_1$ (with equality iff $b = 1$). Since $a_1 \leq a$ (since $a \in I_1$, with equality iff $a = 0$),

$$\sum_{I \in \hat{\mathcal{I}}} \ell(I) \geq \ell(I_1) = b_1 - a_1 \geq b - a.$$

On the other hand, if $b \notin I_1$, then $b_1 \leq b$. Then $b_1 \in A$ (since also $a \leq b_1$, since $a \in I_1$; in fact, one can show that $a < b_1$). Again since $\hat{\mathcal{I}}$ covers A, and since $b_1 \notin I_1$, there is an $I \in \hat{\mathcal{I}}$, $I \neq I_1$, such that $b_1 \in I$. Denote this I as I_2, with endpoints $a_2 < b_2$.

If $b \in I_2$ then $b \leq b_2$. Since $a_2 \leq b_1$ (since $b_1 \in I_2$; in fact, one can show that $a_2 < b_1$) and since $a_1 \leq a$ (since $a \in I_1$),

$$\begin{aligned} \sum_{I \in \hat{\mathcal{I}}} \ell(I) &\geq \ell(I_2) + \ell(I_1) \\ &= b_2 - a_2 + b_1 - a_1 \\ &\geq b_2 - a_1 \\ &\geq b - a. \end{aligned}$$

On the other hand, if $b \notin I_2$, then $b_2 \leq b$. Then $b_2 \in A$ (since also $a \leq b_2$, since $a \leq b_1$ and $b_1 \leq b_2$, since $b_1 \in I_2$; in fact, one can show that $a < b_2$). And so on.

Since, $\hat{\mathcal{I}}$ is finite, this process must terminate with a T such that $a_1 \leq a$, $b \leq b_T$ and also $a_2 \leq b_1$, $\ldots$, $a_T \leq b_{T-1}$. Then,

$$\begin{aligned} \sum_{I \in \hat{\mathcal{I}}} \ell(I) &\geq \ell(I_T) + \cdots + \ell(I_1) \\ &= b_T - a_T + b_{T-1} - \cdots - a_2 + b_1 - a_1 \\ &\geq b_T - a_1 \\ &\geq b - a. \end{aligned}$$

Thus, in all cases, $\sum_{I \in \hat{\mathcal{I}}} \ell(I) \geq b - a$. Since $\sum_{I \in \mathcal{I}} \ell(I) \geq \sum_{I \in \hat{\mathcal{I}}} \ell(I)$, the claim follows.

(b) Suppose that $A \subseteq [0,1]$ is any interval, not necessarily closed. Thus, A is of the form $[a,b]$, $(a,b]$, $[a,b)$, or (a,b). For $\varepsilon > 0$ small enough (any $\varepsilon < (b-a)/2$), let $A_\varepsilon = [a+\varepsilon, b-\varepsilon]$. Then $A_\varepsilon \subseteq A$, hence $\lambda^*(A_\varepsilon) \leq \lambda^*(A)$. By step 2a, $\lambda^*(A_\varepsilon) = b - a - 2\varepsilon$, hence

$$b - a - 2\varepsilon \leq \lambda^*(A).$$

Since ε was arbitrary, $\lambda^*(A) \geq b - a$, as was to be shown.

■

Theorem 2 and Theorem 3 together imply that $[0,1]$, and hence $\mathbb{R}$, is uncountable.

Theorem 4. *$[0,1]$ is uncountable.*

Proof. By Theorem 3, $\lambda^*([0,1]) = 1 > 0$. By Theorem 2, this implies that $[0,1]$ is not countable. ■

Finally, the following result says that any subset of [0,1] can be approximated, in an outer measure sense, by an open set.

Theorem 5. *For any set $A \subseteq [0,1]$ and any $\varepsilon > 0$, there is an open set O such that $A \subseteq O$ and $\lambda^*(O) - \lambda^*(A) < \varepsilon$.*

Proof. Fix $\varepsilon > 0$. By the definition of λ^*, there is a cover of A by open intervals, call this cover $\mathcal{I}$, with the property that

$$\sum_{I \in \mathcal{I}} \ell(I) < \lambda^*(A) + \varepsilon.$$

Let $O = \bigcup_{I \in \mathcal{I}} I$. Then O is open, $A \subseteq O$, and (by sub-additivity) $\lambda^*(O) \leq \sum_{I \in \mathcal{I}} \ell(I)$. Thus, $\lambda^*(O) < \lambda^*(A) + \varepsilon$, hence

$$\lambda^*(O) - \lambda^*(A) < \varepsilon.$$

■

Remark 3. It is not hard to show that if we had instead defined outer measure using closed intervals (call this λ^C) or just any intervals (call this λ^I), then for any set A, $\lambda^*(A) = \lambda^C(A) = \lambda^I(A)$. The open interval characterization of outer measure facilitates the proof of Theorem 4, which exploits the compactness of $[0,1]$. □.

Remark 4. If we had approximated the length of a set A by using covers by *finite* sets of intervals then we would have gotten *Jordan outer measure* instead of Lebesgue outer measure. Historically, Jordan measure (also called Jordan content) is a precursor to Lebesgue measure. It is not hard to see that the Jordan outer measure of

a set $A \subseteq [0,1]$ equals the Riemann upper integral of the indicator function for that set. As a consequence, the measure theory developed from Jordan outer measure suffers the drawbacks already discussed for Riemann integration, hence our focus instead on Lebesgue measure. I discuss Jordan measure a bit more in Remark 10 in Section 5.4. □

4.1 Vitali Sets.

The main result of this subsection, Theorem 8, says that λ^* is not countably additive and hence $([0,1], \mathbb{P}([0,1]), \lambda^*)$ is not a measure space. The proof is due to Vitali.

Restrict attention to $[0,1)$ and for any $x \in [0,1)$ and $r \in [0,1)$ define

$$x + r \pmod 1 = \begin{cases} x + r & \text{if } x + r < 1 \\ x + r - 1 & \text{if } x + r \geq 1. \end{cases}$$

This is called *mod 1 addition*. If you think of the points in $[0,1)$ as being arranged along the edge of a disk, like numbers on the face of an analog clock, then $x + r \pmod 1$ corresponds to moving a pointer, like the minute hand of an analog clock, first from 0 to x and then by an additional r. For example, $1/4 + 3/4 \pmod 1 = 0$, for the same reason that, on a clock, if the minute hand is on 0 (it is, say, 1 PM exactly), then 15 minutes plus another 45 minutes brings the minute hand of a clock back to 0. Mod 1 addition is commutative and associative.

Given a set $A \subseteq [0,1)$, define

$$A + r \pmod 1 = \{x \in [0,1) : \exists a \in A \text{ s.t. } x = a + r \pmod 1\};$$

that is, $A + r \pmod 1$ is the set formed by adding r, mod 1, to every element of A. In the clock analogy, if one thinks of A as points distributed along the rim of the clock face, then $A + r \pmod 1$ is a rotation of A by the amount r.

The following result says that λ^* is *translation invariant*, mod 1.

Theorem 6. *For any $A \subseteq [0,1)$ and any $r \in [0,1)$, $\lambda^*(A) = \lambda^*(A + r \pmod 1)$.*

Proof. Let $\mathcal{O}$ be any open cover of A by open intervals in $[0,1]$. For any $O \in \mathcal{O}$, $O + r \pmod 1$ is either an open interval or a union of two open intervals. Moreover, one can check that the length of O is the same as the length of $O + r \pmod 1$. Since the sets $O + r \pmod 1$ cover $A + r \pmod 1$, this shows that $\lambda^*(A + r \pmod 1) \leq \lambda^*(A)$. The proof of the reverse inequality is essentially the same. ■

Let $\hat{\mathbb{Q}} = \mathbb{Q} \cap [0,1)$. For each $x \in [0,1)$ let $[x]$ denote the set defined by: $b \in [x]$ iff $b \in [0,1)$ and there is an $r \in \hat{\mathbb{Q}}$ such that $b = x + r \pmod 1$. $[x]$ is an equivalence class, meaning that $b \in [x]$ iff $x \in [b]$. To see this, suppose that $b \in [x]$. I must show that $x \in [b]$. Since $b \in [x]$, there is an $r \in \hat{\mathbb{Q}}$ such that $b = x + r \pmod 1$. I have the following cases.

- If $b = x$ (i.e., $r = 0$), then, trivially, $x \in [b]$.
- If $b > x$ then $b = x + r$ with $r \in (0,1) \cap \mathbb{Q}$, hence $x = b - r$, hence $x = b + (1 - r) - 1$. Since $r \in (0,1) \cap \mathbb{Q}$, $1 - r \in (0,1) \cap \mathbb{Q}$. And since $b = x + r$, it follows that $b + (1 - r) = x + r + (1 - r) = x + 1 \geq 1$. Putting this all together, $x = b + (1 - r) \pmod 1$, hence $x \in [b]$.
- If $x > b$, then, since $b = x + r \pmod 1$ with $r \in (0,1) \cap \mathbb{Q}$, $b = x + r - 1$. Hence $x = b + (1 - r)$ with $1 - r \in (0,1) \cap \mathbb{Q}$, hence $x \in [b]$.

Because each $[x]$ is an equivalence classes, any two equivalence classes are either equal or disjoint.

$[0] = \hat{\mathbb{Q}}$. The other equivalence classes are (mod 1) translations of $\hat{\mathbb{Q}}$ by a non-rational number. In particular, each $[x]$ is countable. Since $[0,1)$ is not countable, the set of all $[x]$ is uncountable. That is, the equivalence classes partition $[0,1)$ into uncountably many disjoint sets, each of which is countable.

By the Axiom of Choice, there exists a set V, called a Vitali set, that contains exactly one element from each equivalence class $[x]$. (I am not claiming that V is unique; it suffices that there exists at least one such set.) V is uncountable. It is (for me) impossible to visualize; its properties imply that it is not an interval; it is some sort of cloud of points.

Theorem 7. *Let V be a Vitali set. For each $r \in \hat{\mathbb{Q}}$, define*

$$V_r = V + r \text{ (mod 1)}.$$

The following hold.

1. $$[0,1) = \bigcup_{r \in \hat{\mathbb{Q}}} V_r.$$
2. *For any $r, r' \in \hat{\mathbb{Q}}$, $r \neq r'$, V_r and $V_{r'}$ are disjoint.*

Proof.

1. Note first that since each $V_r \subseteq [0,1)$, $\bigcup_{r \in \hat{\mathbb{Q}}} V_r \subseteq [0,1)$. It remains to show that $[0,1) \subseteq \bigcup_{r \in \hat{\mathbb{Q}}} V_r$. Consider any $x \in [0,1)$. By the definition of V, there is a $v \in V$ such that $v \in [x]$, hence $x \in [v]$. Thus, $x = v + r \pmod 1$ for some $r \in \hat{\mathbb{Q}}$, which implies that $x \in V_r$, hence $x \in \bigcup_{r \in \hat{\mathbb{Q}}} V_r$, as was to be shown.

2. Contraposition: suppose $x \in V_r \cap V_{r'}$ for some $r, r' \in \hat{\mathbb{Q}}$. Without loss of generality, assume $r' \geq r$. Since $x \in V_r$, there is a $v \in V$ such that $x = v + r \pmod 1$. Since $x \in V_{r'}$, there is a $v' \in V$ such that $x = v' + r' \pmod 1$. Thus $x \in [v]$ and $x \in [v']$. As discussed above, the equivalence classes are either disjoint or equal. Since x lies in the intersection, $[v] = [v']$. The construction of V then implies that $v = v'$, which implies that $r = r'$. By contraposition, if $r \neq r'$ then V_r and $V_{r'}$ are disjoint.

■

Theorem 8. *λ^* is not finitely additive, hence not countably additive.*

Proof. Let V be a Vitali set as in Theorem 7 and its proof. Since the set $\{V_r\}$ is countable, it can be enumerated as, say, $\{V_1, V_2, \dots\}$. (So each V_t, where t is a positive natural number, corresponds to some V_r, where r is a rational number in $[0, 1)$.)

By countable sub-additivity (Theorem 1), and since $\lambda^*(V_t) = \lambda^*(V)$ for every t (by Theorem 6),

$$1 = \lambda^*([0,1)) \leq \sum_t \lambda^*(V_t) = \sum_t \lambda^*(V) = \lim_{T\to\infty} T\ \lambda^*(V).$$

This implies that $\lambda^*(V) > 0$. But then finite additivity fails. For any natural number T,

$$\lambda^*\left(\bigcup_{t=1}^{T} V_t\right) \leq 1$$

since $\bigcup_{t=1}^{T} V_t \subseteq [0, 1)$. But if $T\lambda^*(V) > 1$,

$$\sum_{t=1}^{T} \lambda^*(V_t) = T\lambda^*(V) > 1.$$

Hence

$$\lambda^*\left(\bigcup_{t=1}^{T} V_t\right) < \sum_{t=1}^{T} \lambda^*(V_t).$$

■

The situation is actually worse than Theorem 8 suggests. One might be tempted to argue that the above shows a defect in the definition of λ^*, and that perhaps there is another way to define "equally likely" for all subsets of $[0, 1)$. Unfortunately, the proof of Theorem 8 shows that for any $\mu : \mathbb{P}([0, 1]) \to \mathbb{R}_+$, if

1. $\mu([0, 1)) > 0$, and
2. μ is translation invariant mod 1 on $[0, 1)$,

then μ violates countable additivity, and that if μ is also countably sub-additive then it violates even finite additivity.

Alternatively, one might argue that the argument above shows a flaw in the Axiom of Choice, which was invoked in the construction of the Vitali set. Intuitively, without Choice, the collection of points that I have called V might still exist but

might not be a *set*, in which case it would not be in $\mathbb{P}([0,1])$, which is the domain of λ^*. Solovay (1970) shows that, indeed, if Choice is dropped then λ^* is countably additive on all of $\mathbb{P}([0,1])$. But, as I discuss in Section 5.5, you cannot get rid of these sorts of problems without simultaneously giving up other, desirable, mathematical properties. Loosely, the existence of sets like the Vitali sets is the price that one must pay for the convenience of the continuum.

Yet another response, even less appealing (to me) than dropping Choice, is to drop translation invariance. But if the Continuum Hypothesis holds, then even dropping translation invariance will not help.[3] Recall that the uniform distribution assigns probability zero to each singleton set $\{x\}$, $x \in [0,1]$.

Theorem 9. *If the Continuum Hypothesis holds, then for any $\mu : \mathbb{P}([0,1]) \to \mathbb{R}_+$, if*

1. *$\mu([0,1]) > 0$, and*
2. *$\mu(\{x\}) = 0$ for all $x \in [0,1]$,*

then μ violates countable additivity.

Proof. See Billingsley (1995). ■

In summary, we have the following alternatives: we can drop additivity, we can drop the Axiom of Choice (implicitly restricting the domain of λ^*), or we can explicitly restrict the domain of λ^*. Standard Measure Theory does the last.

5 Lebesgue Measure on $[0,1]$.

5.1 Measurablility.

By Theorem 8, λ^* is not additive on $\mathbb{P}([0,1])$. I handle this problem by restricting the domain to exclude sets like the Vitali sets defined in Section 4.1. The following approach is due to Carathéodory. In Section 5.5, I verify explicitly that this approach does indeed eliminate the Vitali sets.

Definition 2. *A set $A \subseteq [0,1]$ is* measurable *iff for any set $E \subseteq [0,1]$,*

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

Let $\mathcal{M}$ denote the set of measurable subsets of $[0,1]$.

[3] Recall from Set Theory that the Continuum Hypothesis states that there is no set whose cardinality is intermediate between that of $\mathbb{N}$ and that of $\mathbb{R}$. The Continuum Hypothesis is provably independent of the standard (Zermelo-Fraenkel) axioms of Set Theory. Thus, it must be explicitly assumed whenever it is used.

Note that, by sub-additivity (Theorem 1), for any sets $A, E \subseteq [0,1]$, since $E = (E \cap A) \cup (E \cap A^c)$,

$$\lambda^*(E) \leq \lambda^*(E \cap A) + \lambda^*(E \cap A^c).$$

Measurability requires that this inequality hold with equality. Put differently, if A is not measurable then there is a set E such that,

$$\lambda^*(E) < \lambda^*(E \cap A) + \lambda^*(E \cap A^c),$$

even though $E \cap A$ and $E \cap A^c$ are disjoint. As I discuss further in Section 5.5, if A is not measurable then, in fact, this strict inequality holds for $E = [0,1]$, in which case $1 < \lambda^*(A) + \lambda^*(A^c)$. An interpretation is that if A is not measurable then it is so complicated that it is not be possible to closely approximate A and A^c by covering them with open intervals: there is a non-vanishing amount of overlap between the covering of A and the covering of A^c.

Theorem 18 below implies that $\mathcal{M}$ is the largest σ-algebra on which λ^* is countably additive and that $\mathcal{M}$ consists of sets that are either in $\mathcal{B}$ (the smallest σ-algebra containing the intervals) or are "almost" in $\mathcal{B}$ (differ from sets in $\mathcal{B}$ by sets that have an outer measure of zero). For the moment, however, I focus on showing that $\mathcal{M}$ is a non-empty σ-algebra and that λ^* is countably additive on $\mathcal{M}$.

To see that $\mathcal{M}$ is not empty note that $[0,1] \in \mathcal{M}$ since for any $E \subseteq [0,1]$, $E \cap [0,1] = E$ while $E \cap [0,1]^c = \emptyset$. The next result records that, in addition, any set with outer measure zero is measurable.

Theorem 10. *For any $A \subseteq [0,1]$, if $\lambda^*(A) = 0$ then $A \in \mathcal{M}$.*

Proof. As noted above, it suffices to show that,

$$\lambda^*(E) \geq \lambda^*(E \cap A) + \lambda^*(E \cap A^c).$$

$E \cap A \subseteq A$ implies $\lambda^*(E \cap A) \leq \lambda^*(A) = 0$, which implies $\lambda^*(E \cap A) = 0$ (since outer measure is always non-negative). $E \cap A^c \subseteq E$ implies $\lambda^*(E \cap A^c) \leq \lambda^*(E)$. Therefore,

$$\lambda^*(E \cap A) + \lambda^*(E \cap A^c) \leq \lambda^*(E),$$

as was to be shown. ■

Therefore, by Theorem 2, any countable set is measurable. In particular, the rationals in $[0,1]$ are measurable. For ease of reference, I record this as a separate result.

Theorem 11. *If $A \subseteq [0,1]$ is countable then it has Lebesgue outer measure zero and is therefore measurable.*

Proof. As just noted, this follows from Theorem 2 and Theorem 10. ■

Another useful fact is the following.

Theorem 12. *If $A, B \subseteq [0,1]$, $B \subseteq A$ and B is measurable then,*

$$\lambda^*(A) - \lambda^*(B) = \lambda^*(A \setminus B).$$

Proof. $\lambda^*(A) = \lambda^*(A \cap B) + \lambda^*(A \cap B^c)$, since B is measurable. The result then follows, since $A \cap B = B$ (since $B \subseteq A$) and $A \cap B^c = A \setminus B$. ■

I now record the main result of this subsection.

Theorem 13. *$\mathcal{M}$ is a σ-algebra and λ^* is countably additive on $\mathcal{M}$.*

Proof.

1. *For any $A \in \mathcal{M}$, $A^c \in \mathcal{M}$.*

   This is immediate since the definition of measurability is symmetric in A and A^c.

2. *For any $A_1, A_2 \in \mathcal{M}$, $A_1 \cup A_2 \in \mathcal{M}$.*

   Take any $E \subseteq [0,1]$. By sub-additivity (Theorem 1), it suffices to show that,

   $$\lambda^*(E) \geq \lambda^*(E \cap (A_1 \cup A_2)) + \lambda^*(E \cap (A_1 \cup A_2)^c).$$

   Since A_1 is measurable,

   $$\lambda^*(E) = \lambda^*(E \cap A_1) + \lambda^*(E \cap A_1^c).$$

   Since A_2 is measurable,

   $$\lambda^*(E \cap A_1) = \lambda^*(E \cap A_1 \cap A_2) + \lambda^*(E \cap A_1 \cap A_2^c)$$

   and

   $$\lambda^*(E \cap A_1^c) = \lambda^*(E \cap A_1^c \cap A_2) + \lambda^*(E \cap A_1^c \cap A_2^c).$$

   Combining,

   $$\lambda^*(E) = \lambda^*(E \cap A_1 \cap A_2) + \lambda^*(E \cap A_1 \cap A_2^c) + \lambda^*(E \cap A_1^c \cap A_2) + \lambda^*(E \cap A_1^c \cap A_2^c).$$

   Since

   $$E \cap (A_1 \cup A_2) = (E \cap A_1 \cap A_2) \cup (E \cap A_1 \cap A_2^c) \cup (E \cap A_1^c \cap A_2),$$

   sub-additivity implies

   $$\lambda^*(E \cap (A_1 \cup A_2)) \leq \lambda^*(E \cap A_1 \cap A_2) + \lambda^*(E \cap A_1 \cap A_2^c) + \lambda^*(E \cap A_1^c \cap A_2).$$

   On the other hand, since

   $$E \cap (A_1 \cup A_2)^c = E \cap A_1^c \cap A_2^c,$$

   it follows that

   $$\lambda^*(E \cap (A_1 \cup A_2)^c) = \lambda^*(E \cap A_1^c \cap A_2^c).$$

   Combining all this yields the desired inequality.

3. *$\mathcal{M}$ is an algebra and λ^* is finitely additive on $\mathcal{M}$.*

   By step 2 and induction, if $\{A_t\}$ is a finite set of elements of $\mathcal{M}$ then $\bigcup_t A_t$ is also in $\mathcal{M}$. Thus, given step 1, $\mathcal{M}$ is an algebra.

   As for finite additivity, if A_1, A_2 are disjoint then $(A_1 \cup A_2) \cap A_1 = A_1$ and $(A_1 \cup A_2) \cap A_1^c = A_2$. Therefore, if A_2 is measurable then, taking $E = A_1 \cup A_2$, $\lambda^*(A_1 \cup A_2) = \lambda^*((A_1 \cup A_2) \cap A_1) + \lambda^*((A_1 \cup A_2) \cap A_1^c) = \lambda^*(A_1) + \lambda^*(A_2)$. Finite additivity then follows by induction.

4. *λ^* is countably additive on $\mathcal{M}$.* Since we already know that λ^* is finitely additive, let $\{A_t\}$ be a countably infinite set of disjoint measurable sets. Let $A = \bigcup_t A_t$. By sub-additivity, $\lambda^*(A) \leq \sum_t \lambda^*(A_t)$. On the other hand, since $\bigcup_{t=1}^T A_t \subseteq A$,

$$\lambda^* \left(\bigcup_{t=1}^T A_t \right) \leq \lambda^*(A),$$

   hence, by finite additivity,

$$\sum_{t=1}^T \lambda^*(A_t) \leq \lambda^*(A).$$

   Thus, the sequence of partial sums, which is weakly increasing, is bounded above, and hence converges, and $\lambda^*(A) \leq \sum_t \lambda^*(A_t) \leq \lambda^*(A)$, which implies the result.

5. *$\mathcal{M}$ is a σ-algebra.* Let $\{A_t\}$ countable set of measurable sets. Since we already know that $\mathcal{M}$ is an algebra, it suffices to consider the case in which $\{A_t\}$ is countably infinite. Let $A = \bigcup_t A_t$. Consider any set $E \subseteq [0, 1]$. As usual, because of countable sub-additivity, it suffices to show that

$$\lambda^*(E) \geq \lambda^*(E \cap A) + \lambda^*(E \cap A^c).$$

   For the moment, suppose that the A_t are disjoint. For each T, let $B_T = \bigcup_{t=1}^T A_t$. Since $\mathcal{M}$ is an algebra, B_T is measurable. Thus

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

   I claim that

$$\lambda^*(E \cap B_T) = \sum_{t=1}^T \lambda^*(E \cap A_t).$$

   (This would follow from finite additivity *if* the $E \cap A_t$ were measurable.) The argument is by induction. It is trivially true for $T = 1$. Suppose that it is true for T. Since A_{T+1} is measurable,

$$\lambda^*(E \cap B_{T+1}) = \lambda^*(E \cap B_{T+1} \cap A_{T+1}) + \lambda^*(E \cap B_{T+1} \cap A_{T+1}^c).$$

But $B_{T+1} \cap A_{T+1} = A_{T+1}$ and, since the A_t are disjoint, $B_{T+1} \cap A_{T+1}^c = B_T$. Substituting in the induction hypothesis that $\lambda^*(E \cap B_T) = \sum_{t=1}^T \lambda^*(E \cap A_t)$ proves the claim.

Also, since $B_T \subseteq A$, it follows that $E \cap A^c \subseteq E \cap B_T^c$, hence $\lambda^*(E \cap A^c) \leq \lambda^*(E \cap B_T^c)$. Combining all this,

$$\lambda^*(E) \geq \sum_{t=1}^{T} \lambda^*(E \cap A_t) + \lambda^*(E \cap A^c).$$

This implies that the sequence of partial sums $\sum_{t=1}^T \lambda^*(E \cap A_t)$, which is weakly increasing, is bounded above, and hence converges. Therefore,

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

By sub-additivity, $\lambda^*(E \cap A) \leq \sum_{t=1}^{\infty} \lambda^*(E \cap A_t)$, and so

$$\lambda^*(E) \geq \lambda^*(E \cap A) + \lambda^*(E \cap A^c),$$

as was to be shown.

Finally, if the A_t are not disjoint then define $\hat{A}_1 = A_1$, $\hat{A}_2 = A_2 \setminus A_1$, $\hat{A}_3 = A_3 \setminus (A_1 \cup A_2)$, and so on. The $\hat{A}_t$ are disjoint (some may be empty), they are measurable (since $\hat{A}_t = A_t \cap (A_1 \cup \cdots \cup A_{t-1})^c$ and $\mathcal{M}$ is an algebra, by step 3), and $A = \bigcup \hat{A}_t$. Apply the above argument using $\hat{A}_t$ instead of A_t.

■

Remark 5. Let X be any non-empty set. An *outer measure* on X is any set function $\mu^* : \mathbb{P}(X) \to \mathbb{R}_+$ (the target space may also include "$+\infty$") such that (a) $\mu(\emptyset) = 0$, (b) μ is monotone: if $A, B \subseteq X$ and $A \subseteq B$ then $\mu^*(A) \leq \mu^*(B)$ and (c) μ^* is countably sub-additive. λ^* is a special case of this more general concept of outer measure. The proof of Theorem 13 shows that if $\mathcal{A}$ is the subset of $\mathbb{P}(X)$ satisfying Carathéodory measurability (appropriately modified), then $\mathcal{A}$ is a σ-algebra and μ^* is a measure when restricted to $\mathcal{A}$. □

5.2 The Lebesgue measure space.

Let λ denote λ^* when its domain is restricted to $\mathcal{M}$. λ is called *Lebesgue measure*.

Theorem 14. $([0,1], \mathcal{M}, \lambda)$ *is a measure space.*

Proof. Immediate from Theorem 13. ■

For later reference, I record the following.

Theorem 15. *Let $A_1 \supseteq A_2 \supseteq \ldots$ be a sequence of measurable subsets of $[0,1]$. Let*

$$A = \bigcap_t A_t.$$

($A = \emptyset$ is possible.) Then

$$\lambda(A) = \lim_t \lambda(A_t).$$

An analogous statement holds for the union of a sequence $A_1 \subseteq A_2 \subseteq \ldots$ of measurable subsets of $[0,1]$.

Proof. $A_1 \setminus A = A_1 \cap A^c$ and hence is measurable, since $\mathcal{M}$ is an algebra. Similarly, $A_t \setminus A_{t+1}$ is measurable for any t.

One can verify that, because the sequence of A_t is nested,

$$A_1 = A \cup \bigcup_t (A_t \setminus A_{t+1}).$$

Moreover, again because the sequence is nested, A, $(A_1 \setminus A_2)$, $(A_2 \setminus A_3)$, $(A_3 \setminus A_4)$, ... are mutually disjoint. Hence, by countable additivity,

$$\lambda(A_1) = \lambda(A) + \sum_t \lambda(A_t \setminus A_{t+1}).$$

By Theorem 12, for any t,

$$\lambda(A_t \setminus A_{t+1}) = \lambda(A_t) - \lambda(A_{t+1}).$$

Combining,

$$\lambda(A_1) = \lambda(A) + \sum_t [\lambda(A_t) - \lambda(A_{t+1})].$$

Using the definition of $\sum_t [\lambda(A_t) - \lambda(A_{t+1})]$ as the limit of the associated sequence of partial sums, and noting that $\sum_{t=1}^{T} [\lambda(A_t) - \lambda(A_{t+1})] = \lambda(A_1) - \lambda(A_{T+1})$, it follows that,

$$\lambda(A_1) = \lambda(A) + \lambda(A_1) - \lim_t \lambda(A_t).$$

Rearranging gives the result. (As an aside, $\lim_t \lambda(A_t)$ exists since $A_1 \supseteq A_2 \supseteq \ldots$ implies that $\lambda^*(A_1) \geq \lambda^*(A_2) \geq \ldots$, hence $\lambda^*(A_t)$ is a decreasing sequence that is bounded below by zero.)

Finally, the proof of the result for unions is similar. ■

5.3 The Borel measure space.

Theorem 16. *Every interval in* $[0,1]$ *is in* $\mathcal{M}$.

Proof.

1. Consider any interval $A = (a, 1]$. Consider any set $E \subseteq [0,1]$. As usual, it suffices to show that

$$\lambda^*(E) \geq \lambda^*(E \cap A) + \lambda^*(E \cap A^c).$$

By definition of λ^*, for any $\varepsilon > 0$, there is a set $\mathcal{I}$ of open intervals that covers E such that

$$\sum_{I \in \mathcal{I}} \ell(I) \leq \lambda^*(E) + \varepsilon.$$

For each $I \in \mathcal{I}$, let $I_1 = I \cap A$ and $I_2 = I \cap A^c$. Note that since A is of the form $(a, 1]$, both I_1 and I_2 are intervals (not necessarily *open* intervals). Let $\mathcal{I}_1$ be the set of I_1 and $\mathcal{I}_2$ be the set of I_2. By sub-additivity,

$$\lambda^*(E \cap A) \leq \sum_{I_1 \in \mathcal{I}_1} \lambda^*(I_1) = \sum_{I_1 \in \mathcal{I}_1} \ell(I_1),$$
$$\lambda^*(E \cap A^c) \leq \sum_{I_2 \in \mathcal{I}_2} \lambda^*(I_2) = \sum_{I_2 \in \mathcal{I}_2} \ell(I_2).$$

For each $I \in \mathcal{I}$, since $I_1 \cup I_2 = I$, and $I_1 \cap I_2 = \emptyset$, $\ell(I_1) + \ell(I_2) = \ell(I)$. Hence

$$\sum_{I_1 \in \mathcal{I}_1} \ell(I_1) + \sum_{I_2 \in \mathcal{I}_2} \ell(I_2) = \sum_{I \in \mathcal{I}} \ell(I).$$

Therefore,

$$\lambda^*(E \cap A) + \lambda^*(E \cap A^c) \leq \sum_{I_1 \in \mathcal{I}_1} \ell(I_1) + \sum_{I_2 \in \mathcal{I}_2} \ell(I_2) = \sum_{I \in \mathcal{I}} \ell(I) \leq \lambda^*(E) + \varepsilon.$$

Since ε was arbitrary, the result follows.

2. Similarly, any interval $[a, 1]$, $[0, b)$ or $[0, b]$ is measurable.

3. Any interval in $[0,1]$ can be constructed from unions, intersections, and complements of the above. For example, if $a < b$ then

$$[a, b) = [0, a)^c \cap [0, b).$$

■

The *Borel σ-algebra*, denoted $\mathcal{B}$, is the smallest σ-algebra that contains all intervals in [0,1] or, equivalently, all open sets in $[0,1]$. ("Equivalently" since one can show that any open set in $[0,1]$ is a countable union of open intervals.) "Smallest" means that $\mathcal{B}$ is contained in every other σ-algebra that contains all the intervals in [0,1]. It is not hard to see that $\mathcal{B}$ equals the intersection of all σ-algebras containing the intervals in $[0,1]$. This intersection is not empty, since $\mathcal{M}$ (and, for that matter, also $\mathbb{P}([0,1])$) is a σ-algebra that contains all intervals in $[0,1]$. And it is easy to check that any intersection of σ-algebras is itself a σ-algebra.

Theorem 17. $\mathcal{B} \subseteq \mathcal{M}$.

Proof. Immediate from Theorem 16, since $\mathcal{B}$ is the intersection of all σ-algebras containing the intervals, and one such σ-algebra is $\mathcal{M}$. ■

The *Borel measure space* (also confusingly sometimes called the Lebesgue measure space) is $([0,1], \mathcal{B}, \lambda)$. $\mathcal{B}$ turns out to be large enough for nearly all applied work in economics. Hence, in economics, it is actually more common to encounter $([0,1], \mathcal{B}, \lambda)$ than $([0,1], \mathcal{M}, \lambda)$. I discuss what else is in $\mathcal{M}$ in Section 5.4.

Recall the set A defined in Section 3.3 in connection with the Strong Law of Large Numbers. The next example shows that $A \in \mathcal{B}$.

Example 2. Let $A = \{x \in [0,1] : \lim S_T(x) = 1/2\}$, where S_T is the frequency of 1s in the first T terms in the binary expansion of x. I claim that $A \in \mathcal{B}$.

Let A_ε be the set of x for which $S_T(x)$ is within ε of $1/2$ for T sufficiently large (where T can depend on x). That is,

$$A_\varepsilon = \{x \in [0,1] : \exists\, T_x \text{ such that } \forall\, T > T_x, S_T(x) \in N_\varepsilon(1/2)\}.$$

Equivalently,

$$A_\varepsilon = \bigcup_{\hat{T}=1}^{\infty} \bigcap_{T>\hat{T}} \{x \in [0,1] : S_T(x) \in N_\varepsilon(1/2)\}.$$

For fixed T, $\{x \in [0,1] : S_T(x) \in N_\varepsilon(1/2)\}$ is a finite union of intervals, and hence is in $\mathcal{B}$. Then,

$$\bigcap_{T>\hat{T}} \{x \in [0,1] : S_T(x) \in N_\varepsilon(1/2)\}$$

is a countable intersection of sets in $\mathcal{B}$ and hence is in $\mathcal{B}$. Then A_ε is a countable union of sets in $\mathcal{B}$ and hence is in $\mathcal{B}$. Finally,

$$A = \bigcap_{k \in \{1,2,3,\dots\}} A_{1/k},$$

which is a countable intersection of sets in $\mathcal{B}$, and hence is in $\mathcal{B}$. □

5.4 Further characterizing $\mathcal{M}$.

Taking stock, we know that $\mathcal{B} \subseteq \mathcal{M}$ but we don't yet have much, if any insight, into what is in $\mathcal{M} \setminus \mathcal{B}$. In addition, while we know that $\mathcal{M}$ "works" in the sense that it is a σ-algebra on which λ^* is countably additive, we don't know whether $\mathcal{M}$ is the largest σ-algebra that we could use. The goal of this subsection is to address these issues.

To make the exposition a little easier, for this subsection I refer to measurable as *C-measurable* (C for Carathéodory). An alternative to C-measurability is what I will call L-measurability (L for Lebesgue, this being Lebesgue's original characterization of measurability): a set $A \subseteq [0,1]$ is L-measurable iff

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

It is almost immediate that C-measurability implies L-measurability: take $E = [0,1]$ in the definition of C-measurability.

Remark 6. Define the *inner* measure of a set $A \subseteq [0,1]$ to be $\lambda_*(A) = 1 - \lambda^*(A^c)$. Then $\lambda_*(A) \leq \lambda^*(A)$, since, by sub-additivity, $1 = \lambda^*([0,1]) \leq \lambda^*(A) + \lambda^*(A^c)$. L-measurability thus requires that the inner and outer measures be equal: A is L-measurable iff $\lambda_*(A) = \lambda^*(A)$. □

Informally, the next result, Theorem 18, says that L-measurability is equivalent to C-measurability, despite the fact that L-measurability seems weaker. Theorem 18 also says that measurability is equivalent to (various versions of) the property that the set can be approximated arbitrarily closely by open and closed sets.

In the statement of Theorem 18, $\mathfrak{G}_\delta$ is the set of all subsets of $[0,1]$ that can be written as a countable intersection of open sets in $[0,1]$. $\mathfrak{F}_\sigma$ is the set of all subsets of $[0,1]$ that can be written as a countable union of closed sets in $[0,1]$. Because $\mathcal{B}$ is a σ-algebra that contains all open and closed sets, $\mathfrak{G}_\delta, \mathfrak{F}_\sigma \subseteq \mathcal{B}$.

Theorem 18. *For any set $A \subseteq [0,1]$, the following are equivalent.*

1. *A is C-measurable.*
2. *A is L-measurable.*
3. *For any $\varepsilon > 0$ there is an open set O such that $A \subseteq O$ and $\lambda^*(O \setminus A) < \varepsilon$.*
4. *For any $\varepsilon > 0$ there is a closed set K such that $K \subseteq A$ and $\lambda^*(A \setminus K) < \varepsilon$.*
5. *There is a $G \in \mathfrak{G}_\delta$ such that $A \subseteq G$ and $\lambda^*(G \setminus A) = 0$.*
6. *There is an $F \in \mathfrak{F}_\sigma$ such that $F \subseteq A$ and $\lambda^*(A \setminus F) = 0$.*

Proof.

1. *(1) ⇒ (2).* As already noted above, this follows by taking $E = [0,1]$ when applying the definition of C-measurability.

2. *(2) ⇒ (3).* Fix $\varepsilon > 0$. By Theorem 5, there is an open set O with $A \subseteq O$ such that $\lambda^*(O) - \lambda^*(A) < \varepsilon$. The result follows if $\lambda^*(O \setminus A) = \lambda^*(O) - \lambda^*(A)$. I cannot simply cite Theorem 12 here because that assumed C-measurability of the set A, and I have not yet shown that the L-measurability of A implies that it is C-measurable.

   Since O is C-measurable (Theorem 17), it follows that $\lambda^*(A^c) = \lambda^*(A^c \cap O) + \lambda^*(A^c \cap O^c)$. Since $A \subseteq O$ (hence $O^c \subseteq A^c$), this implies

   $$\lambda^*(A^c) = \lambda^*(O \setminus A) + \lambda^*(O^c).$$

   Since A is L-measurable, $\lambda^*(A^c) = 1 - \lambda^*(A)$. Since O is C-measurable, it is L-measurable, hence $\lambda^*(O^c) = 1 - \lambda^*(O)$. It follows that,

   $$1 - \lambda^*(A) = \lambda^*(O \setminus A) + (1 - \lambda^*(O)),$$

   which implies $\lambda^*(O \setminus A) = \lambda^*(O) - \lambda^*(A)$ and hence the result.

3. *(2) ⇒ (4).* Fix $\varepsilon > 0$. By the previous step, since A^c is L-measurable if A is, there is an open set U with $A^c \subseteq U$ such that $\lambda^*(U \setminus A^c) < \varepsilon$. Define $K = U^c$. K is closed since U is open and $K \subseteq A$ since $A^c \subseteq U$. Then $A \setminus K = A \cap K^c = A \cap U = U \cap (A^c)^c = U \setminus A^c$, hence $\lambda^*(A \setminus K) < \varepsilon$.

4. *(3) ⇒ (5).* By (3), for any t, there is an open set O_t such that $A \subseteq O_t$ and $\lambda^*(O_t \setminus A) < 1/t$. Define $G = \bigcap_t O_t$. Then $G \in \mathfrak{G}_\delta$ and $A \subseteq G \subseteq O_t$. Since $G \setminus A \subseteq O_t \setminus A$, it follows that $\lambda^*(G \setminus A) \leq \lambda^*(O_t \setminus A) < 1/t$. Since this holds for all $1/t$, it follows that $\lambda^*(G \setminus A) = 0$.

5. *(4) ⇒ (6).* The proof is almost identical to that in the previous step, with $F = \bigcup_t (K_t)$.

6. *(5) ⇒ (1).* Since $\lambda^*(G \setminus A) = 0$, $G \setminus A$ is C-measurable (Theorem 10). Since G is C-measurable (Theorem 17), it follows that $A^c = G^c \cup (G \setminus A)$ is C-measurable, hence A is C-measurable.

7. *(6) ⇒ (1).* Similarly, since $A = F \cup (A \setminus F)$, and both F and $A \setminus F$ are C-measurable, A is C-measurable.

■

Turning to the two questions posed at the start of this subsection, consider first a set $A \in \mathcal{M} \setminus \mathcal{B}$. By property (6) of Theorem 18, A is the disjoint union of a set in $\mathcal{B}$ (namely F) and a measure zero set (namely $A \setminus F$).

The set of measure zero sets has strictly higher cardinality than $\mathcal{B}$ (see also Example 3 in Section 5.6). Therefore, $\mathcal{M}$ is *much* larger than $\mathcal{B}$. But this comparison is somewhat misleading because the measure zero sets are, after all, measure zero: the elements of $\mathcal{M} \setminus \mathcal{B}$ are "almost" elements of $\mathcal{B}$.

As for the second question, Theorem 18 also implies that $\mathcal{M}$ *is* the largest σ-algebra on which λ^* is countably additive. In particular, L-measurability is a necessary condition for λ^* to be *additive* on an algebra (let alone countably additive on a σ-algebra).

Remark 7. Theorem 18 raises the question why I did not define $\mathcal{M}$ using L- rather than C-measurability. This is a "pick your poison" issue and some developments do use L-measurability. L-measurability is, for me, much easier to understand and it makes the proof of Theorem 16 trivial. But C-measurability yields, arguably, an easier proof of Theorem 13, and C-measurability is easily adapted to general measure spaces. □

Remark 8. From Theorem 18, it is also almost immediate that measurability is equivalent to either of the following.

1. For any $\varepsilon > 0$ there is an open set O and a closed set K with $K \subseteq A \subseteq O$ such that $\lambda^*(O \setminus K) < \varepsilon$.

2. There is a $\mathfrak{G}_\delta$ set G and an $\mathfrak{F}_\sigma$ set F such that $F \subseteq A \subseteq G$ such that $\lambda^*(G \setminus F) = 0$.

□

Remark 9. Another characterization of measurability is that a set $A \subseteq [0,1]$ is measurable iff for any $\varepsilon > 0$ there exists a set $B \subseteq [0,1]$ such that B is a finite union of intervals and $\lambda^*(A \,\Delta\, B) < \varepsilon$. ($A \,\Delta\, B = (A \setminus B) \cup (B \setminus A)$.) For a proof, see, for example, Kolmogorov and Fomin (1970). That is, a set is measurable iff it is "almost" a finite union of intervals.

This characterization of measurability is the first of "Littlewood's Three Principles" of Measure Theory. The other two principles are that any measurable function is "almost" continuous (Lusin's Theorem) and any convergent sequence of measurable functions is "almost" uniformly convergent (Egorov's Theorem). I discuss measurable functions in companion notes. □

Remark 10. Suppose that, as in Remark 4 in Section 4, we had approximated the length of a set A by covering A with *finite* sets of intervals. This yields a set function called *Jordan outer measure*. As already noted in Remark 4, the Jordan outer measure of a set A equals the Riemann upper integral of the indicator function for A. One can show that the Jordan outer measure of $\mathbb{Q} \cap [0,1]$ is 1 even though the Jordan outer measure of any point, or any finite set of points, is 0. Thus, Jordan outer measure is not countably sub-additive, and hence is not a true outer measure in the sense of Remark 5 in Section 5.2. Moreover, the set of Jordan measurable sets (with Jordan measurability defined analogously to Lebesgue measurability) is not a σ-algebra, hence Jordan measure is not a true measure. A related point is that Jordan measurability excludes a number of important sets, including the Borel set that appears in the simple Law of Large Numbers discussed in Section 3.3; see

also Example 2 in Section 5.3. Tao (2011) discusses Jordan measure at some length, by way of motivating the introduction of the more complicated, but ultimately also more useful, Lebesgue measure. □

5.5 Non-measurable sets revisited.

Theorem 6, Theorem 8, and Theorem 13 imply that the Vitali set V of Section 4.1 is *not* measurable. More directly, as in the proof of Theorem 8, for any fixed enumeration $\{V_1, V_2, \dots\}$ of the V_r, there is a T such that

$$\lambda^*\left(\bigcup_{t=1}^{T} V_t\right) < \sum_{t=1}^{T} \lambda^*(V_t).$$

Consider the smallest such T; it must be that $T > 1$ since trivially $\lambda^*(V_1) = \lambda^*(V_1)$. Take

$$E = \bigcup_{t=1}^{T} V_t.$$

Thus $\lambda^*(E) < \sum_{t=1}^{T} \lambda^*(V_t)$. By construction of E,

$$E \cap V_T = V_T$$

and (since the V_t are pairwise disjoint)

$$E \cap V_T^c = \bigcup_{t=1}^{T-1} V_t.$$

Since (by the definition of T),

$$\lambda^*\left(\bigcup_{t=1}^{T-1} V_t\right) = \sum_{t=1}^{T-1} \lambda^*(V_t),$$

it follows that,

$$\begin{aligned}
\lambda^*(E) &< \sum_{t=1}^{T} \lambda^*(V_t) \\
&= \lambda^*(V_T) + \sum_{t=1}^{T-1} \lambda^*(V_t) \\
&= \lambda^*(V_T) + \lambda^*\left(\bigcup_{t=1}^{T-1} V_t\right) \\
&= \lambda^*(E \cap V_T) + \lambda^*(E \cap V_T^c).
\end{aligned}$$

This shows that V_T is not measurable. By translation invariance, no V_r is measurable, including V $(r = 0)$.

By Theorem 18, $\lambda^*(V) + \lambda^*(V^c) > 1$. This helps clarify part of what measurability is capturing. Informally, if a set is not measurable then neither it nor its complement can be well approximated by a countable set of open intervals: there will be too much overlap between the approximation to the set and the approximation to the complement.

One of the most famous examples of non-measurable sets arises in the Banach-Tarski paradox: any solid ball in $\mathbb{R}^3$, call it A, can be cut into a finite number of pieces (five is the minimum) that can then be rearranged (shifted, rotated) to form two solid balls, each identical to A. These pieces are difficult to visualize, but in one construction they look a bit like certain versions of a toy called a puffer ball. Here is a link to a YouTube video. The Axiom of Choice guarantees that these pieces are sets, and as sets, they are non-measurable (in the natural extension of measurability to $\mathbb{R}^3$).

As noted earlier, one possible response to measurability issues is to drop the Axiom of Choice, without which collections of points such as V, and the Banach-Tarski pieces, are not sets. But measurability issues are entangled with other, desirable, results in mathematics. For example, in infinite dimensional settings, the natural analog of the Separating Hyperplane Theorem is a consequence of the Hahn-Banach Extension Theorem. One can show that Hahn-Banach together with the standard Set Theory axioms (but dropping the Axiom of Choice) implies that some sets are not measurable; see Foreman and Wehrung (1991). There seems to be no recourse but to accept non-measurable sets, however weird they may seem.

5.6 More on Measure Zero.

Theorem 10 states that any set of Lebesgue outer measure zero is measurable and hence has Lebesgue measure zero. Sets of measure zero play a special role in Measure Theory. A property (say, continuity or differentiability) is said to hold *almost everywhere* (a.e.) iff the set of points on which the property fails has measure zero. Much of Measure Theory is devoted to studying properties that hold a.e., even though the properties may fail at some x. (Again, probability theory uses "almost surely"; measure theory uses "almost everywhere".)

Any countable set has measure zero (Theorem 11). There are also uncountable measure zero sets.

Example 3. Let $E_0 = [0,1] \setminus (1/3, 2/3)$, let $E_1 = E_2 \setminus ((1/9, 2/9) \cup (7/9, 8/9))$ and so on. In words, at each stage, I remove the (open) middle third from each remaining subinterval. Let $C = \bigcap E_t$. C is called the *Cantor set*. $\lambda(C) = 0$, since, for all t, $C \subseteq E_t$, hence $\lambda(C) \leq \lambda(E_t) = (2/3)^t$, and the latter goes to zero as t goes to infinity.

C has the same cardinality as $\mathbb{R}$. One way to see this is to write the elements

of [0,1] as ternary expansions (i.e., in the form $x = x_1/3 + x_2/9 + x_3/27 + \dots$) and then note that $x \in C$ iff it has a ternary expansion with no 1s in it. Thus $x \in C$ iff it has a ternary expansion that is a string of 0s and 2s. Writing a 1 in place of a 2 puts the set of such strings into 1-1 correspondence with the set of strings of 0s and 1s. The latter set of strings can be identified with the set of binary expansion of the elements of $[0, 1]$. Combining all this, the elements of C can be put into 1-1 correspondence with the elements of $[0, 1]$.

One implication of the Cantor set is that there are measurable sets that are not in the Borel σ-algebra. Informally, the reason is the following. Because C has the same cardinality as $\mathbb{R}$, $\mathbb{P}(C)$ has cardinality strictly greater than that of $\mathbb{R}$. On the other hand, one can show that $\mathcal{B}$ has the same cardinality as $\mathbb{R}$. Thus, there must exists sets in $\mathbb{P}(C)$ that are not in $\mathcal{B}$. But any subset of C has Lebesgue outer measure zero, and hence is measurable. □

In higher dimensions, an important class of measure zero sets (where Lebesgue measure can be generalized using open rectangles rather than open intervals), are lower dimensional surfaces. For example, in $\mathbb{R}^2$ a line has measure zero, in $\mathbb{R}^3$ a plane has measure zero.

An alternative mathematical definition of "small" is topological: a set is "small" if it is closed and nowhere dense, meaning that the complement is open and dense. The relationship between measure zero, on the one hand, and closed and nowhere dense, on the other, is subtle.

Example 4. The set $A = \mathbb{Q} \cap [0, 1]$, being countable, has measure zero. By definition of Lebesgue outer measure, this implies that for any ε, it is possible to cover A by open intervals the sum of whose lengths is no more than $\varepsilon > 0$. For $\varepsilon = 1/k$, let O_k be the union of one such cover. Then O_k is open (obviously) and it is dense in $[0, 1]$ since it contains $\mathbb{Q} \cap [0, 1]$, which is dense in $[0, 1]$. But $\lambda(O_k) \leq 1/k$. Conversely O_k^c is closed and nowhere dense but $\lambda(O_k^c) \geq 1 - 1/k$.

A set is said to be *residual* iff it is a countable intersection of sets that are open and dense. Residual is usually interpreted as "large, but not as large as "open and dense." From the above, $A = \cap_k O_k$ and hence A is a set that is residual, hence "large," but is measure zero, hence "small." □

5.7 Lebesgue integration.

I provide a brief introduction to Lebesgue integration in separate notes. Here, I make only a few remarks.

Lebesgue integration finesses the non-existence problems with the Riemann integral discussed in Section 3.2 and Section 3.3. In particular, for any measurable set $A \subseteq [0, 1]$,

$$\int 1_A = \lambda(A),$$

where $\int 1_A$ is notation for the Lebesgue integral of 1_A.

For $A = \mathbb{Q} \cap [0,1]$ as in Section 3.2, Theorem 11 implies $\lambda(A) = 0$ and hence

$$\int 1_A(x) = 0.$$

For $A = \{x \in [0,1] : S_T(x) \to 1/2\}$ as in Section 3.3, where S_T is the frequency of 1s in the first T terms in the binary expansion of x, one can prove that $\lambda(A) = 1$ (again, this is a version of the Strong Law of Large Numbers) and hence that

$$\int 1_A(x) = 1.$$

More generally, one can show that the Lebesgue integral equals the Riemann integral whenever the Riemann integral exists. And one can prove that the Riemann integral exists iff the function being integrated is continuous almost everywhere (i.e., except possibly on a set of Lebesgue measure zero).

References

Billingsley, Patrick. 1995. *Probability and Measure.* Third ed. New York: John Wiley and Sons.

Foreman, M. and F. Wehrung. 1991. "The Hahn-Banach theorem implies the existence of a non-Lebesgue measurable set." *Fundamenta Mathematicae* 138:13–19.

Kolmogorov, A.Ñ. and S.Ṽ. Fomin. 1970. *Introductory Real Analysis.* Mineola, NY: Dover.

Royden, Halsey. 2010. *Real Analysis.* Fourth ed. Pearson.

Solovay, Robert. 1970. "A Model of Set Theory In Which Every Set of Reals ls Lebesgue Measurable." *Annals of Mathematics* 92:1–56.

Stein, Elias and Rami Shakarchi. 2005. *Real Analysis: Measure Theory, Integration, and Hilbert Spaces.* Princeton University Press.

Tao, Terrence. 2011. *An Introduction to Measure Theory.* American Mathematical Society.