Stat 8501 Lecture Notes
Baby Measure Theory
Charles J. Geyer
July 26, 2023

1 Old Probability Theory and New

All of probability theory can be divided into two parts. I call them master's level and PhD level probability theory. Other terms are classical probability theory and measure-theoretic probability theory. Historically, the dividing line is 1933 when *Grundbegriffe der Wahrscheinlichkeitsrechnung* (Foundations of the Theory of Probability) by Andrey Kolmogorov was published (although Kolmogorov was Russian, he wrote in German to reach a wider audience). What Kolmogorov did was to say that the new real analysis that had started with the PhD thesis of Henri Lebesgue (1902) and had been rapidly generalized to integrals of real-valued functions on arbitrary spaces by Radon, Fréchet, and others (called Lebesgue integration or abstract integration) should also be used in probability theory.

1.1 Discrete, Continuous, and None of the Above

In master's level probability theory we have the distinction between discrete and continuous distributions. The discrete ones have probability mass functions (PMF) and we calculate probabilities and expectations with sums; the continuous ones have probability density functions (PDF) and we calculate probabilities and expectations with integrals. We are aware that we can have distributions that are neither — consider a random vector whose first component is Poisson and second component is normal — but the theory does not cater to such distributions.

One place where we discuss such things in a master's level theory course is Bayesian inference with discrete data x and continuous parameter θ, as when we figure out, if x is binomial given θ and the prior for θ is beta, then the posterior is also beta. If we fuss about this, we notice that we have a mixed joint distribution with x discrete and θ continuous. But in deriving the posterior we are conditioning on x, essentially treating it as fixed, so we are really just working with a continuous distribution of one variable θ.

Another place where we discuss such things in master's level theory is in the discussion of distribution functions (DF). We are told that every DF

corresponds to a distribution (and vice versa), but we are not given any way to deal with the ones that are not discrete or continuous.

Here is an example

$$F(x) = \begin{cases} 0, & x < 0 \\ (1 + x)/2, & 0 \le x \le 1 \\ 1, & x \ge 1 \end{cases}$$

There is an atom at zero, so the distribution is partly discrete, but otherwise is continuous. We can describe this distribution as a 50-50 mixture of the distribution concentrated at zero and the uniform distribution on $(0, 1)$. There really isn't any mystery, but the formulas in master's level theory books don't cover this situation.

Yet another place where we discuss such things in master's level theory is in the discussion of multivariate normal distributions. We are told that any linear function of a normal random vector is another normal random vector. But this gives rise to degenerate normal random vectors. We are also told that any mean vector and any variance matrix correspond to some multivariate normal distribution, but singular variance matrices correspond to degenerate distributions that do not have PDF. Again, there really isn't any mystery. If μ is the mean vector and M is the variance matrix of a normal random vector X, Then M is singular if and only if $v^T M v = 0$ for some nonzero vector v, but

$$v^T M v = \operatorname{var}(v^T X),$$

and a random vector having variance zero is almost surely constant (actually this result requires measure theory, Theorem 1 below), hence $v^T X$ is almost surely constant, and the constant has to be $v^T \mu$. Writing v_i for the components of v and similarly for X and μ, the assumption that v is nonzero means it has at least one nonzero component, say v_k, so we have

$$X_k = \mu_k - \frac{1}{v_k} \sum_{i \ne k} v_i (X_i - \mu_i) \tag{1}$$

with probability one, and we can eliminate X_k from our calculations, using (1) to deal with it. After several such steps we have partitioned our general normal random vector into two parts, one of which has a nondegenerate normal distribution and a PDF and the other of which is a linear function of the first part. Clumsy, but not problematic.

So master's level theory can deal with many distributions that are neither discrete nor continuous (but have some aspects of both) by ad hoc devices, but it has no unified methodology for dealing with all such cases.

Also the notion that every DF corresponds to a probability distribution (which comes from measure-theoretic probability theory) allows much more bizarre distributions than master's level theory can handle.

Here is one example. The *Cantor set* is defined as follows. From the open unit interval remove the middle third leaving an open set; call it C_1. Then remove the middle third of every interval in C_1 obtaining C_2, and keep on going, removing the middle third of every interval in C_{n-1} to obtain C_n. Then

$$C_\infty = \bigcap_{i=1}^{\infty} C_n$$

is the Cantor set. We can use the same construction to make a DF. Let F_n denote the DF of the continuous uniform distribution on C_n and then define

$$F_\infty(x) = \lim_{n \to \infty} F_n(x), \qquad \text{for all } x.$$

It is fairly easy to see that F_∞ is a continuous function. Therefore the random variable having this distribution is continuous in some sense, but not in the master's level theory sense. This distribution attributes probability zero to each of the intervals removed, and the lengths of these intervals add up to one. So all of the probability is concentrated on the Cantor set C_∞, which is what the measure-theoretic jargon calls a set of Lebesgue measure zero, Lebesgue measure being the measure-theoretic analog of ordinary length. This distribution does not have a PDF, and it is totally mysterious from the master's level theory point of view. The master's level recipe for finding the probability density function by differentiating the DF fails: F_∞ is not differentiable anywhere on C_∞ and has derivative zero off C_∞. The function F_∞ is called the "devil's staircase" because of its bizarre properties (continuous but not differentiable), so we can call the probability distribution with DF F_∞ the devil's staircase distribution.

So what is the point of that? Does this distribution have any applications? Why do we care? Answers: no point, no applications, and we don't care. The reason for inventing this stuff wasn't to produce weird examples. The reason was limit theorems.

1.2 Limit Theorems

In the nineteenth century mathematicians started studying limits of sequences of functions and right away found out that the limit of a sequence

of continuous functions needn't be continuous, the limit of a sequence of differentiable functions needn't be differentiable, the limit of a sequence of integrable functions needn't be integrable, and started studying conditions under which these do hold. In particular,

$$\lim_{n\to\infty} f_n(x) = f(x), \qquad \text{for all } x, \tag{2}$$

does not imply that f is an integrable function in the sense of ordinary calculus (so-called Riemann integrable) or even if it is that

$$\lim_{n\to\infty} \int f_n(x)\,dx \to \int f(x)\,dx. \tag{3}$$

At the beginning of the twentieth century mathematicians (led by Lebesgue) remedied this situation by redefining integration. Lebesgue integration or abstract integration gives the same result as Riemann integration when the latter exists, so nothing you know from calculus changes, but a lot more functions are integrable.

There are three limit theorems of abstract integration theory. *Fatou's lemma* says, if (2) holds with with all of the f_n integrable and nonnegative, then f is integrable and

$$\liminf_{n\to\infty} \int f_n(x)\,dx \ge \int f(x)\,dx. \tag{4}$$

The *monotone convergence theorem* says, if (2) holds with with all of the f_n integrable and the sequence monotone, that is, $f_1(x) \le f_2(x) \le \cdots$ for all x or the same except with the inequalities reversed, then f is integrable and (3) holds, possibly with $+\infty$ or $-\infty$ as the limit. The *dominated convergence theorem* says, if (2) holds and the sequence is dominated by an integrable function, that is, there exists an integrable function m such that $|f_n(x)| \le m(x)$ for all x and all n, then f is integrable and (3) holds. These limit theorems are the only method for calculating Lebesgue integrals or abstract integrals when they aren't also Riemann integrals.

Now the question about applications seems different. There are lots of applications of the limit concept. We want our use of the limit concept to be unimpeded by needless restrictions. Removing the restrictions requires the move from Riemann to Lebesgue integration.

Here is a truly bizarre application. Cantor discovered the rational numbers are countable $\mathbb{Q} = \{r_1, r_2, \ldots\}$ for some sequence r_1, r_2, ... Define

$$f_n(x) = \begin{cases} 1, & x = r_m \text{ for some } m \le n \\ 0, & \text{otherwise} \end{cases}$$

Then we have (2) with

$$f(x) = \begin{cases} 1, & x \in \mathbb{Q} \\ 0, & \text{otherwise} \end{cases}$$

Now f is discontinuous everywhere. It is certainly not Riemann integrable. But each f_n is Riemann integrable, because the value of a function at a finite set of points does not affect a Riemann integral, and the integral is zero. Hence by the monotone convergence theorem

$$\int_{-\infty}^{\infty} f(x)\, dx = 0$$

if the integral is Lebesgue integration.

But integrating bizarre functions isn't the point. The point is limit theorems.

1.3 Discussion

Master's level probability theory has its virtues. It requires only calculus and not any higher level real analysis. It is the only theory that the vast majority of people that have any training in probability theory have been exposed to (including the vast majority of scientists). It works for many applications. But it also has vices.

It is annoying that every time one proves something and writes down a sum or an integral, one has to redo the whole thing with sums replacing integrals (or vice versa) or leave the redo as an "exercise for the reader" with just a comment that, as usual, we need different notation for the discrete case and the continuous case. Also sometimes it is not just a matter of different notation, for example, the conversion of DF to PMF or PDF.

It is also worrying that master's level theory does not cover all of probability theory. Maybe the devil's staircase distribution has no applications, but how can we be sure that there isn't some useful application that master's level theory cannot handle?

Also ad hockery has its limits. It is one thing to say that master's level theory can deal with (some) distributions that are neither discrete nor continuous by ad hoc methods, but it only partially deals with them. It is not at all clear how much of master's level theory can be applied once one gets out of the discrete and continuous cases the theory is designed to handle.

Measure-theoretic probability theory fixes all the vices. It deals with all probability distributions and does so with unified methods, having no

need for separate discussion of the discrete case and the continuous case (or separate discussion of any special cases).

Measure-theoretic probability theory is so clean and so elegant that most theoretical probability and statistics has been written in it for many decades. This means that if one wants to read the research literature, one has to have some idea what it is all about. Of course, there are still a lot of textbooks presenting the old (master's level) theory of probability and statistics. But when one comes to advanced topics, the only literature available may be measure theoretic. An example is Markov chain theory. There are master's level presentations of the theory of Markov chains, for example, Hoel, Port, and Stone (1986). But they only do the discrete case: Markov chains on finite or countable state spaces. The reason is partly historical. The theory of Markov chains on general state spaces was highly unsatisfactory until a technique was independently discovered by Nummelin (1978) and by Athreya and Ney (1978) that made the general state space theory work as cleanly as the countable state space theory (actually Nummelin and Athreya and Ney proposed somewhat different techniques that do the same job). The complete theory of general state space Markov chains is presented in the books Nummelin (1984) and Meyn and Tweedie (2009). All of the this literature uses measure-theoretic probability. No one has tried to "dumb it down" to master's level theory. This makes it nearly impossible to discuss the theory of general state space Markov chains without using measure-theoretic probability theory. Geyer (2011) does attempt this job, managing to discuss some of the basic theory of Markov chains and Markov chain Monte Carlo without measure theory, but has to give up at the end, finding it impossible to discuss the Metropolis-Hastings-Green algorithm (Green, 1995) without measure theory. A similar story could be told about any advanced topic that originated long after 1933. The literature is all measure-theoretic, so you need at least some understanding of measure-theoretic probability to read it.

2 Measure Theory, First Try

One way to unify the discrete and continuous cases is to use the P and E operators (and related operators like var, cov, and cor) exclusively rather than writing sums and integrals. Of course, every master's level theory book uses these, but few use them as much as possible.

One problem with this is that if we are going to take P and E seriously as operators, we have to be more careful about them. What do they operate

on?

P gives probabilities of events, so it is a function $A \mapsto P(A)$ that maps events to real numbers. So we need a notation $\mathcal{A}$ for the family of all events. Then $P : \mathcal{A} \to \mathbb{R}$ is a function just like any other function in mathematics. Such a function is called a *probability measure*.

Note that P is a *different function* for each different probability distribution (this is not made clear in all treatments of master's level theory — if you were under the impression that there is just one P that is kind of an abbreviation for probability, you are not alone). Since it is a different function for every different distribution, we need notation to distinguish different ones. We can "decorate" P, for example, we can use P_θ to distinguish the different distributions in a parametric family of distributions. But, as everywhere else in mathematics, we should avoid "frozen letters." Advanced probability theory often uses other letters, for example, let P, Q, and R be probability measures.

An issue that we leave hanging for now (see Sections 4.1 and 4.5 below) is what is $\mathcal{A}$? An event is a subset of the sample space, but is $\mathcal{A}$ all of the subsets of the sample space or just some of them? It turns out that, for very abstruse technical reasons, the answer is the latter. But we won't worry about that now.

E gives expectations of random variables, so it is a function $X \mapsto E(X)$ that maps random variables to real numbers. There is a problem that not every random variable has an expectation (for example, the expectation of a Cauchy random variable does not exist). So we need a notation $L^1(P)$ for the family of all random variables that have expectation. Then $E : L^1(P) \to \mathbb{R}$ is a function just like any other function in mathematics. Such a function is called an *expectation operator*.

Again note that E is a *different function* for each different probability distribution (and again this is not made clear in all treatments of master's level theory). Since it is a different function for every different distribution, we need notation to distinguish different ones. We can "decorate" E, for example, we can use E_θ to distinguish the different distributions in a parametric family of distributions. Avoiding "frozen letters" is harder here. People really insist on E for expectation and not some other letter. So decoration is the only option people tolerate.

There is, of course, a tight relation between a probability measure P and the corresponding expectation operator E. This is apparent in the notation $L^1(P)$ for the domain of E. Sometimes we want to make this explicit. The

measure theoretic notations for this are

$$E(X) = \int X \, dP = \int X(\omega) \, dP(\omega) = \int X(\omega) \, P(d\omega), \tag{5}$$

which are four different notations for exactly the same concept. The integral signs here do not mean integration in the sense of calculus (so-called Riemann integration). For now we just take it to be another notation for expectation. If you know what $E(X)$ means, then you know what the other notations in (5) mean. The actual operation may be ordinary integration or summation or some combination of the two if the distribution is neither discrete nor continuous or something else entirely (more on that later).

When we use P and E as actual mathematical functions and when we can make the connections between P and E in (5), we are doing measure theory. We don't need to know all the technicalities to use the notations and (hopefully) read this notation in measure-theoretic literature.

3 Measure Theory, Second Try

So what are probability measures and expectation operators? Since Kolmogorov (1933) it has been considered the thing to do to present probability theory as an axiomatic theory based on the following axioms.

Recall from the preceding section, that a *probability measure* is a function $P : \mathcal{A} \to \mathbb{R}$, where $\mathcal{A}$ is the family of events. Such a function is a probability measure if it satisfies three axioms.

(P1)
$$P(A) \geq 0, \qquad A \in \mathcal{A}.$$

(P2)
$$P(\Omega) = 1,$$
where Ω is the sample space (the largest element of $\mathcal{A}$).

(P3) If I is a countable set and $\{A_i : i \in I\}$ a disjoint subfamily of $\mathcal{A}$, meaning $A_i \cap A_j = \varnothing$ whenever $i \neq j$, then
$$P\left(\bigcup_{i \in I} A_i\right) = \sum_{i \in I} P(A_i).$$

The complicated axiom (P3) is called *countable additivity*. Countable additivity includes finite additivity (finite sets are countable).

In order for (P3) to make sense, it must be a requirement on $\mathcal{A}$ that if the A_i are events, then so is their union $\bigcup_{i \in I} A_i$ (more on this in Section 4.1 below). A textbook of measure-theoretic probability theory, such as Billingsley (1995) or Fristedt and Gray (1996), develops all of the theory of probability and expectation from these three axioms. But that takes a huge amount of work that we want to avoid.

So we just present a parallel set of axioms for expectation theory. This is not commonly done, although one book (Whittle, 2005) does develop probability theory along these lines.

Recall from the preceding section, that an *expectation operator* is a function $E : L^1(P) \to \mathbb{R}$, where $L^1(P)$ is the family of random variables having expectation. Such a function is the expectation operator associated with a probability measure P if it satisfies five axioms.

(E1) If X and Y are random variables having expectation, then $X + Y$ also has expectation and

$$E(X + Y) = E(X) + E(Y).$$

(E2) If X is a random variable having expectation, and a is a real number, then $Y = aX$, meaning $Y(\omega) = aX(\omega)$ for all $\omega \in \Omega$, where Ω is the sample space, also has expectation, and

$$E(Y) = aE(X).$$

(E3) If X and Y are nonnegative random variables such that $X \le Y$, meaning $X(\omega) \le Y(\omega)$ for all $\omega \in \Omega$, and Y has expectation, then X also has expectation, and

$$E(X) \le E(Y).$$

(E4) If X_1, X_2, ... is a monotone sequence of random variables having expectation, meaning $X_1(\omega) \le X_2(\omega) \le \cdots$ for all $\omega \in \Omega$ or the same with the inequalities reversed, and X is another random variable satisfying

$$X(\omega) = \lim_{n \to \infty} X_n(\omega), \qquad \omega \in \Omega,$$

then X has expectation and

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

provided the limit exists; otherwise X does not have expectation.

(E5) If A is an event, then I_A, the indicator function of the set A defined by

$$I_A(x) = \begin{cases} 0, & x \notin A \\ 1, & x \in A \end{cases}$$

is a random variable having expectation, and

$$E(I_A) = P(A).$$

and, if A is the whole sample space, then $P(A) = 1$.

Properties (E1) and (E2) can be used separately, but together they are called *linearity of expectation*. Property (E3) is called *monotonicity of expectation*. Property (E4) is called *monotone convergence*. Property (E5) is the relationship between probability and expectation.

Properties (E1) and (E2) imply that $L^1(P)$ is a vector space (closed under addition and scalar multiplication) so we can also think of random variables that have expectation as elements of the vector space $L^1(P)$. This is not particularly helpful, since this vector space is infinite-dimensional unless the sample space is finite.

Properties (E2), (E3), and (E5) imply that every bounded random variable has expectation. Many unbounded random variables have expectation, but there are usually some unbounded random variables that do not have expectation whenever the sample space is infinite.

Properties (E2), (E3), and (E5) also imply $0 \le P(A) \le 1$ for all events A.

4 Measure Theory, Third Try

4.1 Sigma-Algebras

Let Ω be an arbitrary set. A *sigma-algebra* for Ω is a family $\mathcal{A}$ of subsets of Ω that contains Ω and is closed under complements and countable unions and intersections. De Morgan's laws say $(\cup_i A_i)^c = \cap_i A_i^c$ and $(\cap_i A_i)^c = \cup_i A_i^c$ and imply that closed under complements and countable unions implies closed under countable intersections, and the same holds with unions and intersections swapped, so the definition given above is redundant, but we don't wish to privilege unions over intersections or vice versa. Another term for sigma-algebra is sigma-field. These terms are usually written σ-algebra and σ-field by those who like their writing ugly.

The smallest sigma-algebra is $\{\varnothing, \Omega\}$. It must contain Ω by definition, and it must contain $\varnothing$ because it is Ω^c. Unions and intersections of Ω and $\varnothing$ give us the same sets back, no new sets.

The largest sigma-algebra is the set of all subsets of Ω, called the *power set* of Ω.

4.2 Measurable Spaces

A set Ω equipped with a sigma-algebra $\mathcal{A}$ is called a *measurable space* and usually denoted as a pair $(\Omega, \mathcal{A})$. In this context, the elements of $\mathcal{A}$ are called *measurable sets*.

4.3 Measures

For various reasons, we want to generalize the concept of probability measure. So now we drop axiom (P2) and perhaps also (P1) keeping only the notion of countable additivity (P3).

4.3.1 Positive Measures

When we drop only (P2) but retain (P1), we get the concept of a *positive measure*. In this context, it is useful to allow values of ∞ for the measure, with the convention that $x + \infty = \infty$.

A *positive measure* on a measurable space $(\Omega, \mathcal{A})$ is a function $\mu : \mathcal{A} \to \mathbb{R} \cup \{\infty\}$ that satisfies

$$\mu(A) \geq 0, \qquad A \in \mathcal{A},$$

and, if I is a countable set and $\{A_i : i \in I\}$ a disjoint subfamily of $\mathcal{A}$, meaning $A_i \cap A_j = \varnothing$ whenever $i \neq j$, then

$$\mu\left(\bigcup_{i \in I} A_i\right) = \sum_{i \in I} \mu(A_i), \tag{6}$$

this property being called *countable additivity*.

4.3.2 Signed Measures

When we drop both (P1) and (P2), we get the concept of a *signed measure*. In this context, it is not useful to allow infinite values because there would be no reasonable definition for $\infty - \infty$ if both positive and negative infinite values were allowed.

A *signed measure* on a measurable space $(\Omega, \mathcal{A})$ is a function $\mu : \mathcal{A} \to \mathbb{R}$ that is countably additive, that is, satisfies (6) and the restrictions on I and A_i given just above it.

4.3.3 Examples

Example 4.1.
Counting measure is a positive measure that counts the number of points in a set: $\mu(A)$ is the number of points in A. If Ω is infinite then $\mu(A) = \infty$ when A is any infinite subset of Ω.

Example 4.2.
Lebesgue measure on $\mathbb{R}$ corresponds to the dx of ordinary calculus:

$$\mu(A) = \int_A dx \tag{7}$$

whenever A is a set over which the Riemann integral is defined. For other sets, we have to use countable additivity to extend the measure from Riemann measurable sets to Lebesgue measurable sets.

If we take Ω to be the whole real line, then $\mu(\Omega) = \infty$. So again, we need to allow ∞ as a value.

The same idea works for $\mathbb{R}^n$. Just take the integrals in (7) to be multiple integrals.

Example 4.3.
If P and Q are probability measures, then $P - Q$ is a signed measure, so we need signed measures to compare probability measures.

If μ and ν are signed measures and a and b are real numbers, then $a\mu + b\nu$ is a signed measure, so the family of all signed measures on a measurable space is a vector space.

The latter explains why signed measures are of interest in real analysis. The former explains why they are of interest in probability theory.

4.4 Measure Spaces

A measurable space $(\Omega, \mathcal{A})$ equipped with a measure μ, either a positive measure or a signed measure, is called a *measure space* and usually denoted as a triple $(\Omega, \mathcal{A}, \mu)$.

Probability measures are special cases of both positive measures and signed measures. If μ is a probability measure, then $(\Omega, \mathcal{A}, \mu)$ is called a *probability space*.

Again, the elements of $\mathcal{A}$ are called *measurable sets*. If more specificity is required, μ-measurable sets.

4.5 Existence

Does Lebesgue measure exist? Of course it does; otherwise a very large area of mathematics would be nonsense. But the existence question turns out to be very tricky.

According to the widely accepted (but not universally accepted) axiomatic set theory foundations of mathematics, Zermelo-Fraenkel set theory with the axiom of choice (ZFC), Lebesgue measure does not exist if the sigma-algebra is taken to be the set of all subsets of the real line (the power set of $\mathbb{R}$).

According to the Banach-Tarski theorem (also called the Banach-Tarski *paradox* because the whole point of the theorem is to show that ZFC has bizarre consequences), a solid sphere in 3 dimensions can be divided into five pieces that can rotated, translated, and reassembled to make two solid spheres of the same size. This violates countable additivity. It even violates finite additivity. Rotating and translating sets does not change their Riemann measure, so it should not change their Lebesgue measure. But here it does. Thus the "paradox."

According to the Carathéodory extension theorem, any measure on an algebra can be extended to one on some sigma-algebra containing it, where an algebra for Ω is is only required to contain Ω and be closed under complements and finite unions and intersections. The set of all finite unions of intervals, including degenerate intervals (single points) is an algebra. And we can evaluate (7) for each set in this algebra (it is just the sum of the lengths of the intervals that make up the set). Now the Carathéodory extension theorem says that Lebesgue measure exists, but the sigma-algebra, called the sigma-algebra of *Lebesgue measurable* sets, is not the whole power set of $\mathbb{R}$. The Banach-Tarski theorem tells us it cannot be (the five pieces of the sphere that the theorem claims exist cannot be Lebesgue measurable).

Could I show you an example of a set that is not Lebesgue measurable? No. They are literally indescribable. Any set that could be obtained by starting with intervals and applying a countable sequence of operations is Lebesgue measurable. The axiom of choice implies nonmeasurable sets exist, but the assertion is completely nonconstructive. There is no way to actually describe one.

5 Integration Theory

5.1 Measurable Functions

A function f from one measurable space $(S, \mathcal{A})$ to another $(T, \mathcal{B})$ is *measurable* if

$$f^{-1}(B) \in \mathcal{A}, \qquad B \in \mathcal{B},$$

where

$$f^{-1}(B) = \{\, x \in A : f(x) \in B \,\}.$$

You might think it is hard to verify that a function is measurable, and, in general, it is. But the task becomes easier when the target space is the real numbers. Then it is a theorem that is only moderately difficult to prove that it is enough to check

$$f^{-1}\big((-\infty, t)\big) \in \mathcal{A}, \qquad t \in \mathbb{R}.$$

5.2 Abstract Integrals

If $(\Omega, \mathcal{A}, \mu)$ is a measure space and f is a real-valued measurable function on Ω, we want to define the integral of f, which is usually written $\int f \, d\mu$ or $\int f(x) \, d\mu(x)$ or $\int f(x) \, \mu(dx)$. We usually use the latter (it goes better with Markov chain theory).

We want this new kind of integral to have all the familiar properties of integrals from calculus (Riemann integrals) and to obey the limit theorems, that is, for positive measures we want (E1), (E2), (E3), (E4), and (E5) to hold, and for signed measures we want all of these except (E3) to hold.

Let I_A denote the indicator function of the set A. Then I_A is a measurable function if and only if A is a measurable set (an element of $\mathcal{A}$). And we define

$$\int I_A(x) \, \mu(dx) = \mu(A).$$

(This is the analog of (E5) for general measures.) Then we want this new kind of integral to have the linearity properties (E1) and (E2), so if f is a measurable function having a finite set of values, so we can write

$$f(x) = \sum_{i=1}^{n} b_i I_{A_i}(x)$$

for some positive integer n, some real numbers b_1, $\ldots$, b_n, and some measurable sets A_1, $\ldots$, A_n, we define

$$\int f(x)\,\mu(dx) = \sum_{i=1}^{n} b_i \mu(A_i).$$

Then we extend this by monotone convergence to all nonnegative-valued measurable functions that have expectation. Given a nonnegative-valued measurable function f and a positive integer n, define a function f_n by

$$f_n(x) = \begin{cases} (k-1)/n, & (k-1)/n \le f(x) < k/n \text{ and } k = 1, \ldots, n^2 \\ n, & f(x) \ge n \end{cases}$$

Then we know how to integrate each f_n and we have $f_n \uparrow f$, so monotone convergence says $\int f_n\,d\mu \to \int f\,d\mu$, provided the limit exists.

If $\int f_n\,d\mu \to \infty$, then we say, strictly speaking, that f is not integrable and the integral does not exist, but we also say, loosely speaking, that the integral has the value ∞.

We then deal with arbitrary real-valued functions by decomposing them into positive and negative parts

$$f^+(x) = \begin{cases} f(x), & f(x) \ge 0 \\ 0, & \text{otherwise} \end{cases}$$

$$f^-(x) = \begin{cases} -f(x), & f(x) \le 0 \\ 0, & \text{otherwise} \end{cases}$$

Then $f = f^+ - f^-$ and we already know how to integrate f^+ and f^-, and in order to maintain the linearity properties we must define

$$\int f\,d\mu = \int f^+\,d\mu - \int f^-\,d\mu, \tag{8}$$

and we do define it this way when f^+ and f^- are both integrable. Otherwise, we say, strictly speaking, that the integral does not exist and f is not integrable. When either f^+ or f^- is integrable, we can still, loosely speaking, use (8) with the conventions $r - \infty = -\infty$ and $\infty - r = +\infty$ for $r \in \mathbb{R}$. When neither f^+ or f^- is integrable, we are stuck, there is no sensible definition of $\infty - \infty$ in this context, and we have to say the integral of f does not exist, even loosely speaking.

Now the hard work starts. We have to show that the integral so defined obeys the linearity properties and also obeys the convergence theorems.

As this takes weeks in a PhD level real analysis course or in a PhD level probability theory course, we won't try. (As the jargon says, it is "beyond the scope of this course.")

Another thing we have to show is that the integral with respect to counting measure is summation. If μ is counting measure on a set Ω, then

$$\int f \, d\mu = \sum_{\omega \in \Omega} f(\omega).$$

Another thing we have to show is that the integral with respect to Lebesgue measure is Riemann integration, when the latter exists: if μ is Lebesgue measure on $\mathbb{R}$ and f is a Lebesgue measurable real-valued function of one real variable, then

$$\int f \, d\mu = \int f(x) \, dx,$$

and, more generally, if f is Lebesgue measure on $\mathbb{R}^d$ and f is a Lebesgue measurable function of d real variables, then

$$\int f \, d\mu = \iint \ldots \int f(x_1, \ldots, x_d) \, dx_1 \ldots dx_d.$$

But these are also "beyond the scope of this course."

We merely make one more comment before finishing with this subject. It is clear from (8) that f is integrable, strictly speaking, if and only if $|f|$ is integrable, in other words, f is integrable if and only if it is absolutely integrable. Thus abstract integration theory has no concept like conditional convergence in ordinary calculus (integrable but not absolutely integrable or summable but not absolutely summable for infinite series).

6 A Theorem

Here is one example of the use of these axioms, just to show some of the flavor of measure-theoretic argument.

Theorem 1. *If X is a nonnegative random variable, then $E(X) = 0$ if and only if $X = 0$ with probability one.*

Proof. Suppose $E(X) = 0$. Then for any $\varepsilon > 0$ the set

$$A_\varepsilon = \{ \, \omega \in \Omega : X(\omega) \geq \varepsilon \, \}$$

is measurable and $\varepsilon I_{A_\varepsilon} \le X$. Hence by axioms (E2), (E3), and (E5)

$$\varepsilon P(A_\varepsilon) = E(\varepsilon I_{A_\varepsilon}) \le E(X) = 0,$$

from which we conclude that $P(A_\varepsilon) = 0$. The sequence of random variables $I_{A_{1/n}}$, $n = 1, 2, \ldots$ is monotone and increases to I_{A_0}, where

$$A_0 = \bigcup_{n=1}^{\infty} A_{1/n}$$

hence $P(A_0) = 0$ by axiom (E4). And $\omega \in A_0$ if and only if $X(\omega) > 0$. This proves one direction.

Now assume $X = 0$ with probability one, which means, if we define

$$A_0 = \{\, \omega \in \Omega : X(\omega) > 0 \,\},$$

then $P(A_0) = 0$. For any positive integer n, define X_n by

$$X_n(\omega) = \begin{cases} X(\omega), & X(\omega) \le n \\ n, & \text{otherwise} \end{cases}$$

Then by axioms (E2), (E3), and (E5) we have

$$E(X_n) \le nP(I_{A_0}) + 0 \cdot P(I_{A_0^c}) = 0$$

and X_n is clearly a monotone sequence increasing to X, so $E(X) = 0$ by axiom (E4). □

We actually need the monotone convergence theorem (which we are taking as an axiom) to get this result. There is a subject called *finitely additive probability theory* which replaces axiom P3 with the weaker axiom of finite additivity, which is P3 with finite sets replacing countable sets. In finitely additive probability theory, Theorem 1 does not hold. We can reason as far as $X \ge 0$ and $E(X) = 0$ implies $P(X \ge \varepsilon) = 0$ for every $\varepsilon > 0$. But we can go no farther without the monotone convergence theorem (which is a consequence of countable additivity but not of mere finite additivity).

References

Athreya, K. B., and Ney, P. (1978). A new approach to the limit theory of recurrent Markov chains. *Transactions of the American Mathematical Society*, **245**, 493–501.

Billingsley, P. (1995). *Probability and Measure*, 3rd ed. New York: Wiley.

Breiman, L. (1968). *Probability*. Redding, MA: Addison-Wesley. Republished 1992, Philadelphia: Society for Industrial and Applied Mathematics.

Geyer, C. J. (2011). Introduction to Markov chain Monte Carlo. In *Handbook of Markov Chain Monte Carlo*, edited by Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. Boca Raton, FL: Chapman & Hall/CRC.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82, 711–732.

Fristedt, B. E. and Gray, L. F. (1996). *A Modern Approach to Probability Theory*. Boston: Birkhäuser.

Hoel, P. G., Port, S. C., and Stone, C. J. (1972). *Introduction to Stochastic Processes*. Boston: Houghton Mifflin. Republished, Waveland Press, Prospect Heights, Illinois, 1986.

Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung* Springer. English translation (1950): *Foundations of the Theory of Probability*. Chelsea.

Meyn, S. P. and Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability*, second edition. Cambridge: Cambridge University Press.

Nummelin, E. (1978). A splitting technique for Harris recurrent Markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **43**, 309–318.

Nummelin, E. (1984). *General Irreducible Markov Chains and Non-Negative Operators*. Cambridge: Cambridge University Press.

Whittle, P. (2005). *Probability via Expectation*, 4th ed. New York: Springer.