Contract no: BAP098M-RG-15-Med

Probability Theory

J.M. Steele

Department of Statistics,

Wharton School,

University of Pennsylvania

Probability theory is a branch of mathematics that has evolved from the investigation of social, behavioral, and physical phenomena that are influenced by randomness and uncertainty. For much of its early life, probability theory dealt almost exclusively with gambling games, and, even though there were important contributions made by such distinguished mathematicians as Pierre de Fermat, Blaise Pascal, and Pierre-Simon de Laplace, the field lacked respectability and failed to attract sustained attention.

One cause of the slow development of probability theory in its early days was the lack of a widely accepted foundation. Unlike the geometry of Euclid, or even the analytical investigations of Newton and Leibnitz, the theory of probability seemed to be eternally tied to the modelling process. In this respect, probability theory had greater kinship with the theories of heat or elasticity than with the pristine worlds of geometry or algebra.

Over time, foundations for probability were proposed by a number of deep thinking individuals including von Mieses, de Finetti, and Keynes, but the approach that has come to be the most widely accepted is the one that was advanced in 1933 in the brief book *Foundations of Probability Theory* by Andrey Nikolayevich Kolmogorov.

Kolmogorov's approach to the foundations of probability theory developed naturally from the theory of integration that was introduced by Henri Lebesgue and others during the first two of decades of the twentieth century. By leaning on the newly developed theory of integration, Kolmogorov demonstrated that probability theory could be viewed simply as another branch of mathematics. After Kolmogorov's work, probability theory had the same relationship to its applications that one finds for the theory of differential equations. As a consequence, the stage was set for a long and productive mathematical development.

Too be sure, there are some philosophical and behavioral issues that are not well addressed by Kolmogorov's bare-bones foundations, but, over the years, Kolmogorov's approach has been found to be adequate for most purposes. The Kolmogorov axioms are remarkably succinct, yet they have all the power that is needed to capture the physical, social, and behavioral intuition that a practical theory of probability must address.

1 Kolmogorov's Axiomatic Foundations

Central to Kolmogorov's foundation for probability theory was his introduction of a triple $(\Omega, \mathcal{F}, P)$ that is now called a probability space. The triple's first element, the *sample space* Ω, is only required to be a set, and, on the intuitive level, one can think of Ω as the set of all possible outcomes of an experiment. For example, in an experiment where one rolls a traditional six-faced die, then one can take $\Omega = \{1, 2, 3, 4, 5, 6\}$.

The second element of Kolmogorov's probability space is a collection $\mathcal{F}$ of subsets of Ω that satisfy three basic consistency properties that will be described shortly. On the intuitive level, one can think of the elements of $\mathcal{F}$ as "events" that may occur as a consequence of the experiment described by $(\Omega, \mathcal{F}, P)$. Thus, to continue with the example of rolling a die, the set $A = \{1, 3, 5\} \subset \Omega$ would correspond to the event that one rolls an odd number.

Two of the three consistency properties that Kolmogorov imposes on the $\mathcal{F}$ are quite trivial. First, $\mathcal{F}$ is required to contain Ω. Second, $\mathcal{F}$ must be closed under complementation; so, for example, if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$ where $A^c = \{\omega : \omega \in \Omega \text{ and } \omega \notin A\}$. The third condition is only a bit more elaborate; the collection $\mathcal{F}$ must be closed under countable unions. Thus, if $A_k \in \mathcal{F}$ for $k = 1, 2, ...$, then one requires that the union $A_1 \cup A_2 \cup \cdots$ of all of the events in the countable set $\{A_k : 1 \leq k < \infty\}$ must again be an element of $\mathcal{F}$.

The most interesting element of Kolmogorov's triple $(\Omega, \mathcal{F}, P)$ is the *probability*

measure P. Formally, P is just a function that assigns a real number to each of the elements of $\mathcal{F}$, and, naturally enough, one thinks of $P(A)$ as the probability of the event A. Thus, one can specify a probability model for the outcome of rolling a fair die, by *defining* P for $A \subset \Omega$ by $P(A) = \frac{1}{6}|A|$, where $|A|$ denotes the number of elements of the set A.

For sample spaces Ω that are not finite, more care is required in the specification of the probability measures P. To deal with general Ω, Kolmogorov restricted his attention to those P that satisfy three basic axioms:

Axiom 1. For all $A \in \mathcal{F}$, one has $P(A) \geq 0$.

Axiom 2. $P(\Omega) = 1$.

Axiom 3. For any countable collection $\{A_k \in \mathcal{F} : 1 \leq k < \infty\}$ with the property that $A_j \cap A_k = \emptyset$ for all $j \neq k$, one has

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

Axioms 1 and 2 just reflect the intuitive view that $P(A)$ measures the frequency with which A occurs in an imagined sequence of repetitions of the experiment $(\Omega, \mathcal{F}, P)$. For most mathematicians, Axiom 3 also just reflects the most basic intuition about the way probabilities should behave. Nevertheless, Axiom 3 is more subtle because it deals with an infinite collection of sets, and, in some ways, such collections are outside of our direct experience. This has lead some philosophers to examine the possibility of avoiding Kolmogorov's third axiom, and, over the years,

various attempts have been made to replace Kolmogorov's third axiom with the simpler assumption of *finite additivity*.

2 Random Variables

In most applications of probability theory, the triple $(\Omega, \mathcal{F}, P)$ that gives life to probability as a rigorous mathematical subject is almost invisible. In practice, builders of probability models take advantage of various shortcuts that have the effect of keeping the probability triple at a respectful distance. The most important of these shortcuts is the notion of a *random variable*.

On the intuitive level, a random variable is just a number that depends on a chance-driven experiment. More formally, a random variable X is a function from Ω to the real numbers with the property that $\{\omega : X(\omega) \leq t\} \in \mathcal{F}$ for all t. What drives this definition is that one inevitably wants to talk about the probability that X is less than t, and, for such talk to make sense under Kolmogorov's framework, the set $\{\omega : X(\omega) \leq t\}$ must be an element of $\mathcal{F}$. Random variables make the modeler's job easier by providing a language that relates directly to the entities that are of interest in a probability model.

3 The Distribution Function and Related Quantities

There are several ways to specify the basic probabilistic properties of a random variable, but the most fundamental description is given by the *distribution function*

of X, which is defined by $F(t) = P(X \leq t)$. Knowledge of the distribution function tells one everything that there is to know about the probability theory of the random variable. Sometimes it even tells too much.

The knowledge one has of a random variable is often limited, and in such cases it may be useful to focus on just part of the information that would be provided by the distribution function. One common way to specify such partial information is to use the *median* or the *quantiles* of the random variable X. The median is defined to be a number m for which one has $P(X < m) \leq 1/2 \leq P(X \leq m)$. Roughly speaking, the median m splits the set of outcomes of X into two halves so that the top half and the bottom half each have probability that is close to one-half.

The quantile x_p does a similar job. It splits the the possible values of X into disjoint sets, so that one has $P(X < x_p) \leq p \leq P(X \leq x_p)$. When $p = 1/2$, the quantile x_p reduces to the median, and, when $p = 1/4$ or $p = 3/4$, then x_p is called the *lower quartile*, or the *upper quartile*, respectively.

4 Mass Functions and Densities

If a random variable X only takes values from a finite or countable set S, then X is called a *discrete* random variable. In such cases, one can also give a complete description of the probabilistic properties of X by specification of the *probability mass function*, which is defined by setting $f(x) = P(\{x\})$ for all $x \in S$. Here, one should note that the probability mass function f, permits one to recapture the

distribution function by the relation

$$F(t) = \sum_{x \le t, x \in S} f(x) \quad \text{for all } t.$$

There are many important random variables whose values are not confined to any countable set, but, for some of these random variables, there is still a description of their distribution functions that is roughly analogous that provided for discrete random variables. A random variable X is said to be *absolutely continuous* provided that there exist a function f such that the distribution function of X has a representation of the form

$$F(t) = \int_0^t f(x)\, dx \quad \text{for all } t.$$

The function f in this representation is called the *density* of X, and many random variables of practical importance have densities. The most famous of these are the *standard normal* (or, *standard Gaussian*) random variables that have the density

$$f(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2} \quad \text{for all } -\infty < x < \infty,$$

which has a key role in the Central Limit Theorem that will be discussed shortly.

Much of the theory of probability can be developed with just the notion of discrete or absolutely continuous random variables, but there are natural random variables that do not fit into either class. For this reason, the distribution function remains the fundamental tool for describing the probabilities associated with a random variable; it alone is universal.

5 Mathematical Expectation: A Fundamental Concept

If X is a discrete random variable, then its *mathematical expectation* (or just expectation, for short) is defined by

$$E(X) = \sum_{x \in S} x f(x),$$

where f is the probability mass function of X. To illustrate the this definition, one can take X to be the outcome of rolling a fair die, so that $f(x) = 1/6$ for all $x \in S = \{1, 2, ..., 6\}$ and

$$E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = \frac{7}{2}.$$

In a way that is parallel — yet not perfectly so — the expectation of an absolutely continuous random variable X is defined by the integral

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx.$$

Here one needs to work a bit harder to illustrate the definition. Consider, for example, an experiment where one chooses a number at random out of the unit interval $[0, 1]$. From the intuitive meaning of the experiment, one has $P(A) = a$ for $A = [0, a]$, and, from this relationship, one can calculate that the density function for X is given by $f(x) = 1$ for $x \in [0, 1]$ and by $f(x) = 0$ otherwise. For the density f, one therefore finds that

$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx = \int_{0}^{1} x \cdot 1\, dx = \frac{1}{2}.$$

From this formula, one sees that the expected value of a number chosen at random from the unit interval is equal to one-half, and for many people this assertion is perfectly reasonable. Still, one should note that the probability that X equals one-half is in fact equal to zero, so the probabilistic use of the word "expected value" differs modestly from the day-to-day usage.

The probability distribution function and the expectation operation provide almost all of the language that is needed to describe the probability theory of an individual random variables. To be sure, there are several further notions of importance, but these may all be expressed in terms of the distribution or the expectation. For example, the most basic measure of dispersion for the random variable X is its *variance*, which is defined in terms of the expectation by the formula

$$\mathrm{Var(X)} = \mathrm{E(X} - \mu)^2 \quad \text{where } \mu = \mathrm{E(X)}.$$

Finally, the *standard deviation* of X, which is defined to be the square root of $\mathrm{Var(X)}$, provides a useful measure of scale for problems involving X.

6 Introducing Independence

The world of probability theory does not stop with the description of a single random variable. In fact, it becomes rich and useful only when one considers collections of random variables — especially collections of *independent* random variables.

Two events A and B in $\mathcal{F}$ are said to be independent provided that they satisfy the

identity,

$$P(A \cap B) = P(A)P(B).$$

What one hopes to capture with this definition is the notion that the occurrence of A has no influence on the occurrence of B — and vice versa. Nevertheless, the quantities that appear in the defining formula of independence are purely mathematical constructs, and any proof of independence ultimately boils down to the proof of the defining identity in a concrete model $(\Omega, \mathcal{F}, P)$.

Consider, for example, the experiment of rolling two fair dice, one red and one blue. The sample space Ω for this experiment may be taken to be the 36 pairs (j, k) with $1 \leq j \leq 6$ and $1 \leq k \leq 6$, where one thinks of j and k as the number rolled on the red and blue die, respectively. Here, for any $A \subset \Omega$ one can set $P(A) = |A|/|\Omega|$ in order to obtain a model for a pair of fair dice. Under this probability model, there many pairs of events that one can prove to be independent. In particular, one can prove that the event of rolling an even number on the blue die is independent of rolling an odd number on the red die. More instructively, one can also prove that the event of rolling an even number on the blue die is independent of the parity of the sum of the two dice.

7 Extending Independence

The concept of independence can be extended to random variables by defining X and Y to be independent provided that the events $\{X \leq s\}$ and $\{Y \leq t\}$ are independent for all real s and t. One easy consequence of this definition is that for

any pair of monotone functions ϕ and ψ the random variables $\phi(X)$ and $\psi(Y)$ are independent whenever X and Y independent. This mapping property reconciles nicely with the intuitive assertion: if X and Y have no influence on each other, then neither should $\phi(X)$ and $\psi(Y)$ have any influence on each other.

The importance of independence for probability theory will be underscored by the theorems of the next section, but first the notion of independence must be extended to cover more than just pairs of random variables. For a finite collection of n random variables $X_1, X_2, ..., X_n$, the condition for *independence* is simply that

$$P(X_1 \leq t_1, X_2 \leq t_2, ..., X_n \leq t_n) = P(X_1 \leq t_1)P(X_2 \leq t_2) \cdots P(X_n \leq t_n)$$

for all real values $t_1, t_2, ..., t_n$. Finally, an infinite collection of random variables $\{X_s : s \in S\}$ is said to be independent provided that every finite subset of the collection is independent.

8 The Law of Large Numbers

Any mathematical theory that hopes to reflect real-world random phenomena must provide some rigorous interpretation of the intuitive "law of averages." Kolmogorov's theory provides many such results, the most important of which is given in the following theorem.

Theorem 1 (Law of Large Numbers). Suppose that $\{X_i : 1 \leq i < \infty\}$ is a sequence of independent random variables with a common distribution function $F(\cdot)$, so $P(X_i \leq t) = F(t)$ for all $1 \leq i < \infty$ and all real t. If the expectation

$\mu = E(X_1)$ is well-defined and finite, then the random variables

$$\frac{1}{n}\{X_1 + X_2 + \ldots + X_n\}$$

converge to their common mean μ with probability one. More precisely, if one lets

$$A = \left\{\omega : \lim_{n\to\infty} \frac{1}{n}\{X_1(\omega) + X_2(\omega) + \ldots + X_n(\omega)\} = \mu\right\},$$

then the event A satisfies $P(A) = 1$.

9 The Central Limit Theorem

The second great theorem of probability theory is the famous Central Limit Theorem. Although it is not tied as tightly to the meaning of probability as the Law of Large Numbers, the Central Limit Theorem is key to many of the practical applications of probability theory. In particular, it often provides a theoretical interpretation for the *bell curve* that emerges in countless empirical investigations.

Theorem 2 (Central Limit Theorem). Suppose that $\{X_i : 1 \leq i < \infty\}$ is a sequence of independent random variables with a common distribution function F. If these random variables have a finite variance $\text{Var}(X_i) = \sigma^2$, then

$$\lim_{n\to\infty} P\Big(\frac{1}{\sigma\sqrt{n}}\{X_1 + X_2 + \ldots + X_n - n\mu\} \leq x\Big) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^{2/2}} du.$$

10 Stochastic Processes

The most fundamental results of probability theory address the behavior of sums of independent random variables, but many applications of probability theory lead

to sequences of random variables $\{X_n : 1 \leq n < \infty\}$ that may not be independent. Such sequences are called *stochastic processes*, and they are of many different types.

The simplest nontrivial stochastic process is the *Markov chain* which is used to model random phenomena where X_{n+1} depends on X_n, but, given X_n, the value of X_{n+1} does not depend on the rest of the past $X_{n-1}, X_{n-2}, ..., X_1$. To construct such a process, one most often begins with an $n \times n$ matrix $T = \{p_{ij}\}$ with entries that satisfy $0 \leq p_{ij} \leq 1$ and with row sums that satisfy $p_{i1} + p_{i2} + \ldots + p_{in} = 1$. One then generates the Markov chain by making sequential selections from the set $\mathcal{S} = \{1, 2, \ldots, n\}$ in accordance with the rows of the *transition matrix* T. Specifically, if $X_n = i$, then X_{n+1} is obtained by choosing an element of $\mathcal{S}$ in accordance with the probabilities (p_{ij}) given by the ith row of T.

A second group of stochastic processes that is of considerable practical importance is the set of *martingales*. Roughly speaking, these are stochastic process that have the property that the expectation of X_{n+1} given the values of $X_n, X_{n-1}, ..., X_1$ is just equal to X_n. One reason that martingales are important is that they provide a model for the fortune of an individual who gambles on a fair game. Such gambling games are relatively unimportant in themselves, but many economic and financial questions can be reframed as such games. As a consequence, the theory of martingales has become an essential tool in the pricing of stock options and other derivative securities.

The theory of stochastic processes is not confined to just those sequences $\{X_n : 1 \leq n < \infty\}$ with the discrete index set $\{1 \leq n < \infty\}$, and, in fact, almost any set $\mathcal{S}$ can serve as the index set. When one takes $\mathcal{S}$ to be the set of nonnegative real numbers, the index is often interpreted as time, and in this case one speaks of continuous time stochastic processes. The most important of these are the *Poisson process* and *Brownian motion*. Brownian motion is arguably the most important stochastic process.

11 Directions for Further Reading

For a generation, Feller (1968) has served as an inspiring introduction to probability theory. The text assumes only a modest background in calculus and linear algebra, yet it goes quite far. The text of Dudley (1989) is addressed to more mathematically sophisticated readers, but it contains much material that is accessible to readers at all levels. In particular, Dudley (1989) contains many informative historical notes with careful references to the original sources.

For an introduction to the theory of stochastic processes, the text by Çinlar (1975) is recommended, and, for an easy introduction to Brownian motion, martingales, and their applications in finance one can consult Steele (2000).

For background on the early development of probability theory, the books of David (1962) and Stigler (1986) are useful, and the article by Doob (1994) helps make the link to the present. Finally, for direct contact with the master, anyone does well to

read Kolmogorov (1933).

References

Çinlar, E. (1975). *Introduction to Stochastic Processes*, Prentice-Hall, Englewood Cliffs, NJ.

David, F.N. (1962). *Games, Gods, and Gambling: The Origins and History of Probability from the Earliest Times to the Newtonian Era*. Griffin, London.

Doob, Joseph L. (1994). The development of rigor in mathematical probability (1900-1950), in *Development of Mathematics 1900-1950*, J.-P. Pier, ed. Birkhauser-Verlag, Basel.

Dudley, R.M. (1989). *Real Analysis and Probability*. Wadsworth-Brooks/Cole, Pacific Grove.

Feller, W. (1968). *An Introduction to Probability and Its Applications*. Vol. I, 3rd Ed., Wiley, New York.

Kolmogorov, A.N. (1933). *Grundbegriffe der Wahrscheinlichtkeitrechnung*, Springer-Verlag, Berlin. (English translation: N. Morrison (1956), *Foundations of the Theory of Probability*, Chelsea, New York.)

Steele, J.M. (2000). *Stochastic Calculus and Financial Applications*, Springer, New

York.

Stigler, S.M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900.* Harvard University Press, Cambridge, MA.