

ECONOMICS 245A  
INTRODUCTION TO MEASURE THEORY

The goal of this lecture is to take the axioms of probability, which are introduced as the basis for statistical theory, and relate them to measure theory.

*Probability*

Probability is a subject that can be studied independently of statistics, it forms the foundation for statistics. For example, what is the probability that a head comes up twice in a row if we toss an unbiased coin? The answer, .25, is calculated without need of statistical inference.

Kolmogorov (1933) related probability to the concept of measure in integration theory. In so doing he exploited analogies between set theory and the concept of a random variable and developed the axiomatic theory of probability.

*Axiomatic Theory of Probability*

Definitions of a few commonly used terms follow. These terms inevitably remain vague until they are illustrated.

*Random experiment.* An experiment that can be repeated under identical conditions for which all possible outcomes of the experiment are known beforehand, although on any trial the realized outcome is not known beforehand.

*Sample space.* The set of all possible outcomes of a random experiment.

*Simple event.* An event that cannot be a union of other events.

*Composite event.* An event that is not a simple event.

**Example 1.**

Random experiment. Tossing a coin twice.

Sample space:  $\{(H, H), (H, T), (T, H), (T, T)\}$ .

The (composite) event that at least one head occurs:  $(H, H) \cup (H, T) \cup (T, H)$ .

**Example 2.**

Random experiment. Reading the temperature (F) at UCSB at noon on November 1.

Sample space: real interval (0,100).

Events of interest are intervals contained in the sample space.

A probability is a nonnegative number we assign to every event. The axioms of probability are the rules we agree to follow when we assign probabilities.

*Axioms of Probability*

- (1)  $P(A) \geq 0$  for any event  $A$ .
- (2)  $P(S) = 1$ , where  $S$  is the sample space.
- (3) If  $\{A_i\}$ ,  $i = 1, 2, \dots$ , are mutually exclusive (that is,  $A_i \cap A_j = \emptyset$  for all  $i \neq j$ ), then  $P(A_1 \cup A_2 \cup \dots) = P(A_1) + P(A_2) + \dots$ .

The first two rules are consistent with everyday use of the word probability. The third rule is consistent with the frequency interpretation of probability, for relative frequency follows the same rule. If, at the roll of a die,  $A$  is the event that the die shows 1 pip and  $B$  is the event that the die shows 2 pips, the relative frequency of  $A \cup B$  is the sum of the relative frequencies of  $A$  and  $B$ . We want probability to follow the same rule.

If the sample space is discrete, as in example 1, it is possible to assign probability to every event (that is, every possible subset of the sample space) in a way that is consistent with the probability axioms. If the sample space is continuous, however, as in example 2, it is not possible to do so. In such a case we restrict attention to a smaller class of events to which we can assign probabilities in a manner consistent with the axioms.

For random experiments with continuous sample spaces there are an infinite number of possible subsets, so we cannot assign positive probability to all of them. Instead, we replace the set of all possible subsets of the sample space with a sigma field.

*Sigma field.* Let  $F$  be a set of subsets of  $S$ . If: (i)  $A \in F$  implies  $A^c \in F$ , where  $A^c$  is the complement of  $A$ ; and (ii)  $\{A_i\} \in F$ ,  $i = 1, 2, \dots$ , implies  $\cup_{i=1}^{\infty} A_i \in F$ , then  $F$  is a sigma-field of  $S$ .

The first condition for a sigma field requires that if a probability is assigned to the occurrence of an event, then a probability must be assigned to the non-occurrence of the event. The second condition requires that if a probability is assigned to two (or more) events, then a probability must be assigned to the outcome that either, or both, events occurs. Of course the outcome that both events occurs is the intersection of the two events. Conditions (i) and (ii) are sufficient because

$$\begin{aligned} \{A_i\} \in F, i = 1, 2, \dots, \text{ implies } \{A_i^c\} \in F, i = 1, 2, \dots, \text{ implies } \cup_{i=1}^{\infty} A_i^c \in F, \\ \text{implies } (\cup_{i=1}^{\infty} A_i^c)^c \in F, \text{ implies } \cap_{i=1}^{\infty} A_i \in F. \end{aligned}$$

In general, there are many sigma fields for a given sample space.

**Example 1.**

Sigma fields:  $\{\emptyset, S\}$  and  $\{\emptyset, \{(H, T)\}, \{(H, H), (T, H), (T, T)\}, S\}$ .

For a discrete sample space we typically work with the sigma field that comprises all the subsets of  $S$ . For a continuous sample space, we wish to work with the smallest sigma field generated by the events of interest. To begin, we choose half-open intervals to be the events of interest. (The result is independent of the subsets initially chosen as events of interest.) The smallest sigma field generated by the events of interest, with a continuous sample space, is the Borel field. More importantly, the Borel field contains all the nicely behaved subsets, that is those to which one can assign probability according to the axioms of probability.

**Example 2.**

Events of interest:  $(0, x]$  where  $x \in (0, 100)$ .

From the definition of a sigma field, the Borel field on  $(0, 100)$  contains:  $(0, x]$ ;  $(x, 100)$ ;  $\bigcap_{i=1}^{\infty} (0, x - \frac{1}{i}] = x$ ;  $\bigcup_{i=1}^{\infty} (0, x - \frac{1}{i}] = (0, x)$ ;  $[x, 100)$ ;  $(0, x] \cup [y, 100) = (x, y)$ .

At first glance, it appears that the Borel field is the set of all possible subsets of  $(0, 100)$ . But this is not the case because many unpleasant subsets of  $(0, 100)$  have been omitted. Let me give an example of the kind of unpleasant subset that is omitted. To do so, I call upon the following definition from calculus.

*Rational number.* A number that can be expressed as  $\frac{p}{q}$  where  $p$  and  $q$  are integers,  $q > 0$ , and  $p$  and  $q$  do not have a common divisor greater than 1.

**Example.** Many fractions are rational numbers, pi is an irrational number. (Note  $\frac{3}{4}$  is rational while  $\frac{12}{16}$  is not rational. The common divisor restriction can be thought of as a method to uniquely define fractions and discard all but one member of a set of fractions that are equal.)

**Example 2a.**

Take a value  $x$  from  $(0, 1)$ , that is, draw a real number from  $(0, 1)$ . Draw a second real number from  $(0, 1)$ . If the difference of the two real numbers is a rational number, throw the second real number away, otherwise keep it. Continue in this way and construct a set with the following property: the difference between any 2 distinct elements of the set is not a rational number. For example, let  $C$  be the set of all real numbers on  $(0, 1)$  whose difference is not a rational number. Then  $C$  is clearly a subset of  $(0, 1)$  and is big (with an infinite number of points scattered along the  $(0, 1)$  interval). Let  $\{r_i\}_{i=1}^{\infty}$  be rational numbers, with  $r_0 = 0$ ,  $r_1 = \frac{1}{2}$ ,  $r_2 = \frac{1}{3}$ ,  $r_3 = \frac{2}{3}$ ,  $r_4 = \frac{1}{4}$ ,  $r_5 = \frac{3}{4}$ , ... (there are a countably infinite number of rational numbers on  $(0, 1)$ ). Let  $C + r_i$  be the set in which  $r_i$  is added to every element of  $C$ . (The sum operator is done with wrapping, so that  $.9 + .3 = .2$ .) The set  $C + r_i$  is a translation of  $C$ .

We define two properties of the set  $C$ . First, each distinct translation of  $C$  is disjoint. Second, the union of all translations constructed from  $\{r_i\}_{i=1}^{\infty}$  covers

the interval  $(0, 1)$ . (Note, 0 is a rational number, so 0 cannot be in any such translation of  $C$ .)

Proof of the first property: Method, proof by contradiction. Assume that two distinct translations,  $C + r_i$  and  $C + r_j$  are not disjoint. Then there exists an element  $x$  that is common to both sets, so  $x \in C + r_i$  and  $x \in C + r_j$ . If  $x \in C + r_i$  then  $x = y_1 + r_i$  and if  $x \in C + r_j$  then  $x = y_2 + r_j$ . Thus

$$y_1 + r_i = y_2 + r_j \text{ or } y_1 - y_2 = r_j - r_i. \quad (0.1)$$

If  $i \neq j$ , then (0.1) implies that two distinct elements of  $C$  differ by a rational number (because the difference of two rational numbers is a rational number), which violates the definition of  $C$ , so each distinct translation of  $C$  is disjoint.

Proof of the second property: Let  $x$  be any member of  $[0, 1)$  and let  $y$  be an arbitrary element of  $C$ . Then either:  $x$  does not differ from  $y$  by a rational number, so that  $x \in C$ ; or  $x$  does differ from  $y$  by a rational number,  $r_k$ , so that  $x \in C + r_k$ .

With the two properties of  $C$  established, we now show that it is not possible to assign probability (measure) to the set  $C$ , that is  $C$  is not a measurable set. To assign measure, consider the uniform distribution over  $(0, 1)$ , so that for any interval  $(a, b)$  contained in  $(0, 1)$  the measure assigned to the interval, denoted  $m(a, b)$ , is the distance of the interval  $b - a$ . Because distinct translations,  $C + r_i$  and  $C + r_j$ , are disjoint, the measure we should assign to  $C + r_i \cup C + r_j$  equals  $m(C + r_i) + m(C + r_j)$ . Because the union of all translations covers  $[0, 1)$ :

$$1 = m[0, 1) = \sum_{i=0}^{\infty} m(C + r_i) = \sum_{i=0}^{\infty} mC.$$

If we set  $mC = 0$ , then we assign measure 0 to the entire interval! If we set  $mC > 0$ , then we assign measure  $\infty$  to the entire interval. Yet the measure assigned to the entire interval is clearly 1, so we are not able to assign measure to  $C$  in a way that is consistent with the axioms of probability. The set  $C$  is not a measurable set.

We are able to assign measure to the sets contained in the Borel field. For the above example, we assign measure 0 to each individual point  $x$ , and assign measure equal to the length of the interval for all other sets. The assignment of measure is consistent with the axioms of probability. Thus the Borel field excludes all the nasty subsets, such as  $C$  from the previous example. In fact, the Borel field is just the collection of subsets you would likely have come up with if I suggested that you assign measure over the interval  $(0, 1)$ .

Of course, the assignment of measure does not have to be based on a uniform distribution. For example, we could define measure with a tent shaped distribution over the interval. In general, probabilities are assigned through a *set function*, that is, a function that takes a set (such as an element of the Borel field) and returns a number on  $[0, 1]$ . A complete probability measure  $P(\cdot)$  is a set function that maps a sigma field into  $[0, 1]$  in a manner that is consistent with the axioms of probability. (For example,  $P(S) = 1$  and  $P(\emptyset) = 0$ .) From our definition it should be clear that even though  $P(x) = 0$  for any  $x \in (0, 100)$ , this does not mean that  $x$  never occurs. Rather a zero probability simply means that  $\frac{\text{the number of times } x \text{ is drawn}}{\text{the number of draws from } (0,100)} \rightarrow 0$  as the number of draws from  $(0, 100) \rightarrow \infty$ . Similarly, if measure is defined continuously over the sample space  $(0, 100)$ , then  $P((0, 30) \cup (30, 100)) = 1$  does not mean that the value drawn always falls in  $(0, 30) \cup (30, 100)$  (because the value drawn could be 30), so we say the event that the value drawn falls in  $(0, 30) \cup (30, 100)$  occurs *almost* surely. You must keep these concepts distinct from  $P(\emptyset) = 0$  and  $P(S) = 1$ , which are both true because an event always occurs.

Note that the axiomatic approach provides a form for  $P(\cdot)$  but does not provide numeric values for the probabilities, so that either a Bayesian or a frequentist approach to assigning probability is consistent with the axiomatic approach.

We combine the concepts of a sample space, sigma field, and complete probability measure into a

*Probability Space.* A sample space  $S$  endowed with a sigma-field  $F$  and a complete probability measure  $P$  on  $F$  is a probability space, denoted by  $(S, F, P)$ .

### *Random Variables*

Random variables are the building blocks for our statistical procedures. When we speak of a variable, we think of all possible values it can take. When we speak of a random variable, we think in addition of the probability distribution according to which it takes all possible values. As a result, a random variable is misnamed as it is neither random nor a variable.

**Definition.** A *random variable* is a real-valued function defined over a sample space.

The role of abstraction contained in a random variable depends on the structure of the sample space. If the sample space is discrete, as is often the case in classroom examples but rarely the case in practice, the random variable typically maps (or translates) the sample space into real numbers. If the sample space is continuous, as is often the case in practice, the sample space is already defined in

terms of real numbers, so there may be no translation for the random variable to perform.

To fix ideas, we begin with examples with discrete sample spaces.

**Example.**

Roll a six-sided die.

Sample Space:

Random Variable:  $X_1$

Random Variable:  $X_2$

As the example makes clear, there are many possible random variables for each sample space. The key rule that a random variable must satisfy is that the mapping that defines the random variable preserve the event structure of  $F$ . A random variable that preserves the event structure of  $F$ , is said to be measurable with respect to  $F$ , or simply  $F$ -measurable. In general, if the sigma field underlying the random variable contains all subsets of the sample space, a random variable will be measurable with respect to that sigma field.

To understand how a random variable that is not measurable arises in the context of discrete sample spaces, consider the following stylized example.

**Example.**

Flip a coin twice.

Define the sigma field of interest as

$$F_s = \{\emptyset, S, \{(H, H)\}, \{(H, T), (T, H), (T, T)\}, \{(H, H), (T, T)\}, \{(H, T), (T, H)\}\}.$$

The sigma field  $F_s$  defines the events (ordered pairs) of interest to which we assign probabilities. From the axioms of probability, we assign  $P(\emptyset) = 0$  and  $P(S) = 1$ . Suppose the coin is fair, then we assign  $P((H, H)) = \frac{1}{4}$  while the probability that we assign to the set of three ordered pairs  $\{(H, T), (T, H), (T, T)\}$  is  $\frac{3}{4}$ . The key feature of  $F_s$  is that it does not contain all subsets of  $S$ , for example we do not assign probability to the event that consists of only the ordered pair  $\{(T, T)\}$ . In other words, if  $F_s$  is the sigma field of interest, we do not need to know that the coin is fair, we need only know the probabilities we assign to the events in  $F_s$ .

If the random variable  $Y$  is defined as

$$\begin{aligned} Y((H, H)) &= Y((H, T)) = 1, \\ Y((T, T)) &= Y((T, H)) = 0, \end{aligned}$$

then  $Y$  is not measurable with respect to  $F_s$ . To see this, note that  $Y^{-1}(0) = \{(T, T), (T, H)\} \notin F_s$ , so probability is not assigned to this set of ordered pairs and probability cannot be assigned to  $Y(\cdot) = 0$ .

If the random variable  $X$  is defined as

$$\begin{aligned} X((H, H)) &= 0, \\ X((T, T)) &= X((H, T)) = X((T, H)) = 1, \end{aligned}$$

then  $X$  is measurable with respect to  $F_s$ . To see this, note that  $X^{-1}(0) = \{(H, H)\} \in F_s$  and  $X^{-1}(1) = \{(H, T), (T, H), (T, T)\} \in F_s$ , so probability can be assigned to both values of  $X$ . Of course,  $Y$  is measurable with respect to alternative sigma fields. The problem of measurability arose from the fact that  $F_s$  did not contain enough sets. If we define the sigma field to be the set of all subsets of  $S$ , then both  $X$  and  $Y$  are measurable with respect to this sigma field.

For continuous sample spaces, the sigma field is generally a Borel field. Because a Borel field is defined over the real line, a random variable simply maps the elements of the Borel field onto the real line. In general, if the sigma field is a Borel field, then the sigma field contains enough members so that a random variable is typically continuous with respect to the Borel field. Except for “contrived” examples, the problem of measurability does not arise if the sigma field is a Borel field and random variables are generally Borel-measurable.

If we extend the concept of a random variable to a random process  $\{X_t\}_{0 \leq t \leq T}$ , then information and measurability have much in common. Note that the random variable is now a function of two quantities:  $s \in S$  (again we assume  $S$  is continuous); and  $t \in [0, T]$ . The sample space differs for each value of  $t$ . The sample space for  $X_t$  is  $\{(s, r) \in S \times [0, t]\}$ , so the associated sigma field is  $B(S) \otimes B([0, t])$ , where  $B(S)$  is the Borel field generated by  $S$ . Clearly,  $X_{t+1}$  is not measurable with respect to the sigma field  $B(S) \otimes B([0, t])$ , rather  $X_{t+1}$  is measurable with respect to the sigma field  $B(S) \otimes B([0, t+1])$ . For such cases, the sigma field stands in for the information set, so the statement that a random variable is measurable with respect to the sigma field is simply the statement that the random variable is contained in the information set and is considered known.

**Example.**

Let the sigma field consist of past values of  $Y_t$ . If  $X_t$  is a random variable that consists of past values of  $Y_t$  (say  $Y_{t-1} - Y_{t-2}$ ), then  $X_t$  is measurable with respect to the sigma field. That is, if we know the past history of  $Y_t$ , we know the value of  $X_t$ .

The probability (measure) assigned to a random variable is a (set) function that takes the values of the random variable (elements of the Borel field) and assigns a number on  $[0, 1]$ . Let  $P(\cdot)$  be the function (probability measure) that assigns probability to the elements of the underlying sample space  $S$ . Let  $P_X(\cdot)$  be the function (probability measure) that assigns probability to the outcomes of the random variable  $X$ . To understand the relation between  $P(\cdot)$  and  $P_X(\cdot)$ , reconsider

**Example 1.** (Tossing a *fair* coin twice.)

Define the random variable  $X$  to take the value 1 if at least one  $T$  occurs and 0 otherwise

$$\begin{aligned} X((H, H)) &= 0, \\ X((T, T)) &= X((H, T)) = X((T, H)) = 1. \end{aligned}$$

The random variable  $X$  has mapped the sample space into the real line. The probability space that accords with the underlying problem is  $(S, F, P)$  where  $S = \{(H, H), (H, T), (T, H), (T, T)\}$ ,  $F$  is the power set of  $S$ , and  $P$  assigns probability  $\frac{1}{4}$  to each of the elements of  $S$  (and so assigns probability to each of the elements of  $F$ ). The appropriate sample space for  $X$  is not  $S$ , rather it is the real line  $R$ . Similarly, the appropriate sigma field for  $X$  is not  $F$ , rather it is the Borel field  $B$  defined on the real line. Finally, the appropriate probability for  $X$  is not  $P$ , rather it is the function  $P_X$  that maps the Borel field into  $[0, 1]$ . The probability space  $(R, B, P_X)$  is the probability space *induced* by the random variable  $X$ . If the underlying sample space  $S$  were the real line, then the underlying sample space and the induced sample space would be identical.

Just as  $P$  assigns probability to the elements of  $S$ ,  $P_X$  assigns probability to the elements of  $R$ . The most convenient elements to work with are half-closed intervals of the form  $(-\infty, x]$ . Because the coin is fair

$$P_X((-\infty, x]) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{4} & \text{if } 0 \leq x < 1 \\ 1 & \text{if } 1 \leq x \end{cases} .$$

Note that the probability assigned to the interval  $(-\infty, x]$  by  $P_X$  equals the probability assigned to the set  $(s : X(s) \leq x, s \in S)$  by  $P$ . Clearly,  $X$  is a discrete random variable taking only the values 0 and 1. We write  $P(s : X(s) = 1, s \in S)$  as  $P(X = 1)$ , although you must remember that  $s$  is random,  $X$  is a mapping and so is not random.

The formal definition of  $P_X(\cdot)$  is

$$P_X(\cdot) : B \rightarrow [0, 1] \text{ such that } P_X(b) = P(X^{-1}(b)) \forall b \in B.$$

The set  $P(X^{-1}(b)) = P(s : X(s) \in b, s \in S)$  is the set of all elements of  $S$  that result in the outcome  $b$ .

### *Distributions and Density Functions*

A discrete random variable takes a countable (finite or countably infinite) number of real numbers with preassigned probabilities. A continuous random variable takes a continuum of values in the real line according to the rule determined by a density function. A third type of random variable is formed as a mixture of discrete and continuous random variables. The term probability distribution captures a broad concept that refers either to a set of discrete probabilities or a density function or a mixture of both.

Distribution functions arise from the need to make our life simpler mathematically. Because  $P_X(\cdot)$  is a set function (the argument to  $P_X$  is a set of points of the form  $(-\infty, x]$ ), we cannot represent  $P_X(\cdot)$  with an algebraic formula. To allow us to use an algebraic formula, we need a point function. A distribution function is simply a point function that accords with  $P_X$ . To derive a distribution function  $F_X$ , we simply use the endpoints of the interval

$$P_X((-\infty, x]) = F_X(x) - F_X(-\infty). \quad (0.2)$$

To present a formal definition, I call upon the following definition from calculus.

*Right continuous.* A function  $F_X(\cdot)$  is right continuous if

$$\lim_{h \downarrow 0} F_X(x + h) = F_X(x) \text{ for all } x \in R.$$

**Definition.** The point function  $F_X : R \rightarrow [0, 1]$  defined by (0.2) is the distribution function for the random variable  $X$  if: (i)  $F_X(\cdot)$  is a nondecreasing function of  $x$ ; (ii)  $F_X(-\infty) = \lim_{x \rightarrow -\infty} F_X(x) = 0$ ; (iii)  $F_X(\infty) = \lim_{x \rightarrow \infty} F_X(x) = 1$ ; (iv)  $F_X(x)$  is right continuous.

From the definition it follows that

$$F_X(x) = P_X((-\infty, x]) = P(X \leq x).$$

For a discrete random variable, the probability distribution is completely characterized by the equation

$$P(X = x_i) = p_i, i = 1, 2, \dots, n.$$

For a continuous random variable, if there is a nonnegative function  $f_X(x)$  defined over the whole line such that

$$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f_X(x) dx$$

for any  $x_1, x_2$  satisfying  $x_1 \leq x_2$ , then  $f_X(x)$  is the density function of  $X$ . Note  $f_X(\cdot) : R \rightarrow [0, \infty)$ . To understand the density function in more detail: For every point at which  $F_X(x)$  is continuous,  $f_X(x) = \frac{d}{dx} F_X(x) = \lim_{h \downarrow 0} \frac{F(x+h) - F(x)}{h} = \lim_{h \downarrow 0} h^{-1} P(x < X \leq x+h)$ , so the density at  $x$  is approximately  $\frac{1}{h}$  multiplying the probability that the random variable takes values between  $x$  and  $x+h$  (for small values of  $h$ ). The distribution function relates to the density function as

$$F_X(x_1) = \int_{-\infty}^{x_1} f_X(x) dx.$$

To use algebraic formulas, we typically parameterize the distribution function as  $F_X(\cdot; \theta)$ , where  $\theta$  is a finite dimensional parameter vector, and correspondingly parameterize  $f_X(\cdot; \theta)$ . A parametric family of distributions corresponds to a family of probability measures indexed by  $\theta$ :  $P(\cdot; \theta)$  and  $P_X(\cdot; \theta)$  (although the probability measures do not have algebraic representations).