Chapter 1

Probability

1.1 Introduction

With the advent of big data, computer scientists have come up with a plethora of new algorithms that are aimed at revealing patterns from data. *Machine learning* and *artificial intelligence* become buzz words that attract public attention. They defeated best human Go players, automated manufacturers, powered self-driving vehicles, recognized human faces, and recommended online purchases. Some of these industrial successes are based on statistical theory, and statistical theory is based on probability theory. Although this probabilistic approach is not the only perspective to understand the behavior of machine learning and artificial intelligence, it offers one of the most promising paradigms to rationalize existing algorithms and engineer new ones.

Economics has been an empirical social science since Adam Smith (1723-1790). Many numerical observations and anecdotes were scattered in his *Wealth of Nations* published in 1776. Ragnar Frisch (1895-1973) and Jan Tinbergen (1903-1994), two pioneer econometricians, were awarded in 1969 the first Nobel Prize in economics. Econometrics provides quantitative insights about economic data. It flourishes in real-world management practices, from households and firms up to governance at the global level. Today, the big data revolution is pumping fresh energy into research and exercise of econometric methods, while its very foundation is again modern probability theory.

1.2 Axiomatic Probability

Human beings are awed by uncertainty in daily life. In the old days, Egyptians consulted oracles, Hebrews inquired prophets, and Chinese counted on diviners to interpret tortoise shell or bone cracks. Fortunetellers are abundant in today's Hong Kong.

Probability theory is a philosophy about uncertainty. Over centuries, mathematicians strove to contribute to the understanding of randomness. As measure theory matured in the early 20th century, Russian mathematician Andrey Kolmogorov (1903-1987) laid the foundation of modern probability theory in his monograph published in 1933. The formal mathematical language is a system that allows rigorous explorations that have made fruitful advancements, and is now widely accepted as scientific standard in academic and industrial research.

In this lecture, we will briefly introduce the axiomatic probability theory along with familiar results covered in undergraduate *probability and statistics*. This lecture note is at the level

- Hansen (2020): Introduction to Econometrics, or
- Stachurski (2016): A Primer in Econometric Theory, or
- Casella and Berger (2002): Statistical Inference (second edition)

Interested readers may want to read this textbook for more examples.

1.2.1 Probability Space

A *sample space* Ω is a collection of all possible outcomes. It is a set of things. An *event* A is a subset of Ω. It is something of interest on the sample space. A *σ-field*, denoted by $\mathcal{F}$, is a collection of $(A_i \subseteq \Omega)_{i\in\mathbb{N}}$ events such that

1. $\emptyset \in \mathcal{F}$;
2. if an event $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$;
3. if $A_i \in \mathcal{F}$ for $i \in \mathbb{N}$, then $\bigcup_{i\in\mathbb{N}} A_i \in \mathcal{F}$.

Implications: (a) Since $\Omega = \emptyset^c \in \mathcal{F}$, we have $\Omega \in \mathcal{F}$. (b) If $A_i \in \mathcal{F}$ for $i \in \mathbb{N}$, then $A_i^c \in \mathcal{F}$ for $i \in \mathbb{N}$. Thus, if $\bigcup_{i\in\mathbb{N}} A_i^c \in \mathcal{F}$, then $\bigcap_{i\in\mathbb{N}} A_i = (\bigcup_{i\in\mathbb{N}} A_i^c)^c \in \mathcal{F}$.

Example 1. Let $\Omega = \{1,2,3,4,5,6\}$. Some examples of *σ-fields* include

- $\mathcal{F}_1 = \{\emptyset, \{1,2,3\}, \{4,5,6\}, \Omega\}$;
- $\mathcal{F}_2 = \{\emptyset, \{1,3\}, \{2,4,5,6\}, \Omega\}$.
- Counterexample: $\mathcal{F}_3 = \{\emptyset, \{1,2\}, \{4,6\}, \Omega\}$ is not a *σ-field* since $\{1,2,4,6\} = \{1,2\} \bigcup \{4,6\}$ does not belong to $\mathcal{F}_3$.

The σ-field can be viewed as a well-organized structure built on the ground of the sample space. The pair $(\Omega, \mathcal{F})$ is called a *measure space*.

Let $\mathcal{G} = \{B_1, B_2, \ldots\}$ be an arbitrary collection of sets, not necessarily a σ-field. We say $\mathcal{F}$ is the smallest σ-field generated by $\mathcal{G}$ if $\mathcal{G} \subseteq \mathcal{F}$, and $\mathcal{F} \subseteq \tilde{\mathcal{F}}$ for any $\tilde{\mathcal{F}}$ such that $\mathcal{G} \subseteq \tilde{\mathcal{F}}$. A *Borel σ-field* $\mathcal{R}$ is the smallest σ-field generated by the open sets on the real line $\mathbb{R}$.

Example 2. Let $\Omega = \{1,2,3,4,5,6\}$ and $A = \{\{1\}, \{1,3\}\}$. Then the smallest σ-field generated by A is

$$\sigma(A) = \{\emptyset, \{1\}, \{1,3\}, \{3\}, \{2,4,5,6\}, \{2,3,4,5,6\}, \{1,2,4,5,6\}, \Omega\}.$$

A function $\mu : (\Omega, \mathcal{F}) \mapsto [0, \infty]$ is called a *measure* if it satisfies

1. (positiveness) $\mu(A) \geq 0$ for all $A \in \mathcal{F}$;
2. (countable additivity) if $A_i \in \mathcal{F}$, $i \in \mathbb{N}$, are mutually disjoint, then

$$\mu\left(\bigcup_{i \in \mathbb{N}} A_i\right) = \sum_{i \in \mathbb{N}} \mu(A_i).$$

Measure can be understand as weight or length. In particular, we call μ a *probability measure* if $\mu(\Omega) = 1$. A probability measure is often denoted as P. The triple $(\Omega, \mathcal{F}, P)$ is called a *probability space*.

So far we have answered the question: "What is a well-defined probability?", but we have not yet answered "How to assign the probability?"

There are two major schools of thinking on probability assignment. One is *frequentist*, who considers probability as the average chance of occurrence if a large number of experiments are carried out. The other is *Bayesian*, who deems probability as a subjective brief. The principles of these two schools are largely incompatible, while each school has peculiar merit under different context.

1.2.2 Random Variable

The terminology *random variable* is a historic relic which belies its modern definition of a deterministic mapping. It is a link between two measurable spaces such that any event in the σ-field installed on the range can be traced back to an event in the σ-field installed on the domain.

Formally, a function $X : \Omega \mapsto \mathbb{R}$ is $(\Omega, \mathcal{F}) \setminus (\mathbb{R}, \mathcal{R})$ *measurable* if

$$X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$$

for any $B \in \mathcal{R}$. *Random variable* is an alternative, and somewhat romantic, name for a measurable function. We say a measurable is a *discrete random* *variable* if the set $\{X(\omega) : \omega \in \Omega\}$ is finite or countable. We say it is a *continuous random variable* if the set $\{X(\omega) : \omega \in \Omega\}$ is uncountable.

A measurable function connects two measurable spaces. No probability is involved in its definition yet. While if a probability measure P is installed on $(\Omega, \mathcal{F})$, the measurable function X will induce a probability measure on $(\mathbb{R}, \mathcal{R})$. It is easy to verify that $P_X : (\mathbb{R}, \mathcal{R}) \mapsto [0, 1]$ is also a probability measure if defined as

$$P_X(B) = P\left(X^{-1}(B)\right)$$

for any $B \in \mathcal{R}$. This P_X is called the probability measure *induced* by the measurable function X. The induced probability measure P_X is an offspring of the parent probability measure P though the channel of X.

1.2.3 Distribution Function

We go back to some terms that we have learned in a undergraduate probability course. A *(cumulative) distribution function* $F : \mathbb{R} \mapsto [0, 1]$ is defined as

$$F(x) = P(X \le x) = P(\{X \le x\}) = P(\{\omega \in \Omega : X(\omega) \le x\}).$$

It is often abbreviated as CDF, and it has the following properties.

(i) $\lim_{x \to -\infty} F(x) = 0$,

(ii) $\lim_{x\to\infty} F(x) = 1$,

(iii) non-decreasing,

(iv) right-continuous $\lim_{y\to x^+} F(y) = F(x)$.

For continuous distribution, if there exists a function f such that for all x,

$$F(x) = \int_{-\infty}^{x} f(y)\,\mathrm{d}y,$$

then f is called the *probability density function* of X, often abbreviated as PDF. It is easy to show that $f(x) \geq 0$ and $\int_a^b f(x)\,dx = F(b) - F(a)$.

Example 3. We have learned many parametric distributions like the binary distribution, the Poisson distribution, the uniform distribution, the normal distribution, χ^2, t, F and so on. They are parametric distributions, meaning that the CDF or PDF can be completely characterized by a few parameters.

1.3 Expected Value

1.3.1 Integration

Integration is one of the most fundamental operations in mathematical analysis. We have studied Riemann's integral in the undergraduate calculus. Riemann's integral is intuitive, but Lebesgue integral is a more general approach to defining integration.

Lebesgue integral is constructed by the following steps. X is called a *simple function* on a measurable space $(\Omega, \mathcal{F})$ if $X = \sum_i a_i \cdot 1\{A_i\}$ and this summation is finite, where $a_i \in \mathbb{R}$ and $\{A_i \in \mathcal{F}\}_{i \in \mathbb{N}}$ is a partition of Ω. A simple function is measurable.

1. Let $(\Omega, \mathcal{F}, \mu)$ be a measure space. The integral of the simple function X with respect to μ is

$$\int X \mathrm{d}\mu = \sum_i a_i \mu(A_i).$$

   Unlike the Rieman integral, this definition of integration does not partition the domain into splines of equal length. Instead, it tracks the distinctive values of the function and the corresponding measure.

2. Let X be a non-negative measurable function. The integral of X with respect to μ is

$$\int X \mathrm{d}\mu = \sup\left\{\int Y \mathrm{d}\mu : 0 \leq Y \leq X,\ Y \text{ is simple}\right\}.$$

3. Let X be a measurable function. Define $X^+ = \max\{X, 0\}$ and $X^- = -\min\{X, 0\}$. Both X^+ and X^- are non-negative functions. The integral of X with respect to μ is

$$\int X \mathrm{d}\mu = \int X^+ \mathrm{d}\mu - \int X^- \mathrm{d}\mu.$$

The Step 1 above defines the integral of a simple function. Step 2 defines the integral of a non-negative function as the approximation of steps functions from below. Step 3 defines the integral of a general function as the difference of the integral of two non-negative parts.

If the measure μ is a probability measure P, then the integral $\int X \mathrm{d}P$ is called the *expected value*, or *expectation*, of X. We often use the notation $E[X]$, instead of $\int X \mathrm{d}P$, for convenience.

Expectation provides the average of a random variable, despite that we cannot foresee the realization of a random variable in a particular trial (otherwise the study of uncertainty is trivial). In the frequentist's view, the expectation is the average outcome if we carry out a large number of independent trials.

If we know the probability mass function of a discrete random variable, its expectation is calculated as $E[X] = \sum_x xP(X = x)$, which is the integral of a simple function. If a continuous random variable has a PDF $f(x)$, its expectation can be computed as $E[X] = \int xf(x)\,\mathrm{d}x$. These two expressions are unified as $E[X] = \int X\mathrm{d}P$ by the Lebesgue integral.

1.3.2 Properties of Expectations

Here are some properties of mathematical expectations.

- The probability of an event A is the expectation of an indicator function. $E[1\{A\}] = 1 \times P(A) + 0 \times P(A^c) = P(A)$.

- $E\left[X^r\right]$ is call the r-moment of X. The *mean* of a random variable is the first moment $\mu = E\left[X\right]$, and the second *centered* moment is called the *variance* $\text{var}\left[X\right] = E\left[(X-\mu)^2\right]$. The third centered moment $E\left[(X-\mu)^3\right]$, called *skewness*, is a measurement of the symmetry of a random variable, and the fourth centered moment $E\left[(X-\mu)^4\right]$, called *kurtosis*, is a measurement of the tail thickness.

- Moments do not always exist. For example, the mean of the Cauchy distribution does not exist, and the variance of the $t(2)$ distribution does not exist.

- $E[\cdot]$ is a linear operation. If $\phi(\cdot)$ is a linear function, then $E[\phi(X)] = \phi(E[X])$.

- *Jensen's inequality* is an important fact. A function $\varphi(\cdot)$ is convex if $\varphi(ax_1 + (1-a)x_2) \leq a\varphi(x_1) + (1-a)\varphi(x_2)$ for all x_1, x_2 in the domain and $a \in [0,1]$. For instance, x^2 is a convex function. Jensen's inequality says that if $\varphi\left(\cdot\right)$ is a convex function, then $\varphi\left(E\left[X\right]\right) \leq E\left[\varphi\left(X\right)\right]$.

- *Markov inequality* is another simple but important fact. If $E\left[|X|^r\right]$ exists, then $P\left(|X| > \epsilon\right) \leq E\left[|X|^r\right]/\epsilon^r$ for all $r \geq 1$. *Chebyshev inequality* $P\left(|X| > \epsilon\right) \leq E\left[X^2\right]/\epsilon^2$ is a special case of the Markov inequality when $r = 2$.

- The distribution of a random variable is completely characterized by its CDF or PDF. Moment is a function of the distribution. To back out the underlying distribution from moments, we need to know the moment-generating function (mgf) $M_X(t) = E[e^{tX}]$ for $t \in \mathbb{R}$ whenever the expectation exists. The rth moment can be computed from mgf as

$$E[X^r] = \frac{\mathrm{d}^r M_X(t)}{\mathrm{d}t^r}\Big|_{t=0}.$$

1.4 Multivariate Random Variable

A bivariate random variable is a measurable function $X : \Omega \mapsto \mathbb{R}^2$, and more generally a multivariate random variable is a measurable function $X : \Omega \mapsto \mathbb{R}^n$. We can define the *joint CDF* as $F(x_1, \ldots, x_n) = P(X_1 \le x_1, \ldots, X_n \le x_n)$. Joint PDF is defined similarly.

It is sufficient to introduce the joint distribution, conditional distribution and marginal distribution in the simple bivariate case, and these definitions can be extended to multivariate distributions. Suppose a bivariate random variable (X, Y) has a joint density $f(\cdot, \cdot)$. The *conditional density* can be roughly written as $f(y|x) = f(x, y) / f(x)$ if we do not formally deal with the case $f(x) = 0$. The *marginal density* $f(y) = \int f(x, y)\, dx$ integrates out the coordinate that is not interested.

1.4.1 Independence

In a probability space $(\Omega, \mathcal{F}, P)$, for two events $A_1, A_2 \in \mathcal{F}$ the *conditional probability* is

$$P(A_1|A_2) = \frac{P(A_1 A_2)}{P(A_2)}$$

if $P(A_2) \neq 0$. If $P(A_2) = 0$, the conditional probability can still be valid in some cases, but we need to introduce the *dominance* between two measures, which we do not elaborate here. In the definition of conditional probability, A_2 plays the role of the outcome space so that $P(A_1 A_2)$ is standardized by the total mass $P(A_2)$.

Since A_1 and A_2 are symmetric, we also have $P(A_1 A_2) = P(A_2|A_1)P(A_1)$. It implies

$$P(A_1|A_2) = \frac{P(A_2|A_1)\, P(A_1)}{P(A_2)}$$

This formula is the well-known *Bayes' Theorem*.

We say two events A_1 and A_2 are *independent* if $P(A_1 A_2) = P(A_1)P(A_2)$. If $P(A_2) \neq 0$, it is equivalent to $P(A_1|A_2) = P(A_1)$. In words, knowing A_2 does not change the probability of A_1.

Regarding the independence of two random variables, X and Y are *independent* if $P(X \in B_1, Y \in B_2) = P(X \in B_1)\, P(Y \in B_2)$ for any two Borel sets B_1 and B_2.

If X and Y are independent, $E[XY] = E[X]E[Y]$.

1.4.2 Law of Iterated Expectations

Given a probability space $(\Omega, \mathcal{F}, P)$, a sub σ-algebra $\mathcal{G} \subseteq \mathcal{F}$ and a $\mathcal{F}$-measurable function X with $E|X| < \infty$, the *conditional expectation* $E[X|\mathcal{G}]$ is defined as a $\mathcal{G}$-measurable function such that $\int_A X \mathrm{d}P = \int_A E[X|\mathcal{G}] \mathrm{d}P$ for all $A \in \mathcal{G}$. *Law of iterated expectation* is a trivial fact if we take $A = \Omega$.

In the bivariate case, if the conditional density exists, the conditional expectation can be computed as $E[Y|X] = \int y f(y|X) \mathrm{d}y$. The law of iterated expectation implies $E[E[Y|X]] = E[Y]$.

Below are some properties of conditional expectations

1. $E[E[Y|X_1, X_2]|X_1] = E[Y|X_1]$;
2. $E[E[Y|X_1]|X_1, X_2] = E[Y|X_1]$;
3. $E[h(X) Y|X] = h(X) E[Y|X]$.

1.5 Summary

If it is your first encounter of measure theory, the new definitions here may seem overwhelmingly abstract. A natural question is that: "I earned high grade in my undergraduate probability and statistics; do I really need the fancy mathematics in this lecture to do well in econometrics?" The answer is yes and no. *No* is in the sense that if you want to consume econometric methods, instead of grasp the underlying theory, then the axiomatic probability does not add much to your weaponry. You can be an excellent economist or even applied econometrician without knowing measure theory. *Yes* is in the sense that without measure theory, we cannot even formally define conditional expectation, which will be the subject of our next lecture and is a core concept of econometrics. The pillars of asymptotic theory — law of large numbers and central limit theorem — can only be made accurate with this foundation. If you are aspired to work on econometric theory, you will meet and use measure theory so often in your future study and finally it becomes part of your muscle memory.

In this course, we try to keep a balance manner. On the one hand, many econometrics topics can be presented with elementary mathematics. Whenever possible, econometrics should reach wider audience with a plain appearance. On the other hand, we introduce these concepts in this lecture and will invoke them in the discussion of asymptotic theory later. Your investment in advanced mathematics will not be wasted in vain.

`Zhentao Shi. September 5, 2020`