3 Mathematical Foundations: Part II

3.1 Densities

In this section we go back to the Riemann integrals to define densities since the definition of densities for Lebesgue integrals involves the so called Radon-Nikodym Theorem, which will introduce unnecessary complications for our purpose. Recall that given a random variable X and a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, there is a probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbf{P})$ induced by the random variable X. Therefore, one may calculate the expectation of X in two ways. The first is to use the definition above to integrate in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. That is, $\mathbb{E}X = \int_\Omega X(\omega)\mathbb{P}(\mathrm{d}\omega)$. The other way is to integrate in the induced probability space $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbf{P})$, that is, $\mathbb{E}X = \int_{\mathbb{R}} x\mathbf{P}(\mathrm{d}x) = \int_{\mathbb{R}} xF(\mathrm{d}x)$. Take the last integral as a Riemann integral. If $F(x)$ is differentiable, and $\mathrm{d}F(x) = f(x)\mathrm{d}x$, then we have $\mathbb{E}X = \int_{\mathbb{R}} xf(x)\mathrm{d}x$. We call $f(x)$ the density function of the random variable X.

Similarly, if g is a measurable function from $\mathbb{R}$ to $\mathbb{R}$, we have

$$\mathbb{E}g(X) = \int_\Omega g(X(\omega))\mathbb{P}(\mathrm{d}w) = \int_{\mathbb{R}} g(x)\mathbf{P}(\mathrm{d}x) = \int_{\mathbb{R}} g(x)F(\mathrm{d}x) = \int_{\mathbb{R}} g(x)f(x)\mathrm{d}x.$$

Note that we may view the last integral as a Lebesgue integral with respect to the Lebesgue measure λ:

$$\int_{\mathbb{R}} g(x)f(x)\lambda(\mathrm{d}x), \qquad \text{or} \qquad \int_{\mathbb{R}} gf\mathrm{d}\lambda.$$

That is, $\mathrm{d}\mathbf{P} = f\mathrm{d}\lambda$, or $\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\lambda} = f$. In this sense, f is called the Radon-Nikodym derivative or the density of the measure $\mathbf{P}$ with respect to the measure λ. The Radon-Nikodym theorem gives the conditions under which such densities exist.

Once we have the concept of density for the random variable X, we obviously have that for any $B \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{P}(X \in B) = \int_B f(x)\mathrm{d}x$$

where f is the density function of X. Similarly, for a random vector $X = (X_1, X_2, \ldots, X_n)'$ defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, its density is a function $f(x_1, x_2, \ldots, x_n)$ of n variables such that for any $B \in \mathcal{B}(\mathbb{R}^n)$, we have

$$\mathbb{P}(X \in B) = \int_B f(x_1, x_2, \ldots, x_n) \mathrm{d}x_1 \mathrm{d}x_2 \ldots \mathrm{d}x_n.$$

3.2 σ-algebras and Information

In this section we consider the econometric interpretation of σ-algebras. A probability space describes an experiment that produces a result ω distributed according to $\mathbb{P}$. The probability of ω lying in $A \in \mathcal{F}$ is given by $\mathbb{P}(A)$. Now imagine an observer who does not know which ω has been drawn but does know for each $A \in \mathcal{F}$ whether $\omega \in A$ or $\omega \notin A$. Then we may identify the observer's knowledge about the experiment with the σ-algebra. In this sense, the σ-algebra represents the information the observer has. If there is another σ-algebra $\mathcal{G}$ which is a sub-σ-algebra of $\mathcal{F}$, that is, $\mathcal{G}$ is a σ-algebra and $\mathcal{G} \subset \mathcal{F}$, and if another observer can only tell whether $\omega \in B$ or $\omega \notin B$ for $B \in \mathcal{G}$, then this observer has less information (partial information) about the experiment than the former observer. In this sense, a sequence of σ-algebras $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots \subset \mathcal{F}$ represents different layers of information.

This partial-information interpretation can be most clearly understood in the context of σ-algebras generated by random variables. The σ-algebra $\sigma(X)$ generated by a random variable X is the σ-algebra generated by the sets of the form

$$\{X \leq c\}, \qquad c \in \mathbb{R}.$$

Once we know the realization of X, for any $A \in \sigma(X)$, we know whether the drawn ω is in

A or not. Now consider transforming X by a function g defined by

$$g(x) = \begin{cases} 1, & \text{if } x \geq 0, \\ -1, & \text{otherwise.} \end{cases}$$

The transformation still yield a random variable $Y = g(X)$, and the corresponding generated σ-algebra is

$$\sigma(Y) = \Big\{\varnothing, \{Y = 1\}, \{Y = -1\}, \Omega\Big\}.$$

Now knowing the value of Y corresponds to knowing whether ω is in each of the four sets in $\sigma(Y)$. By transforming X with a non-one-to-one mapping, we reduce the σ-algebra and therefore the information contained.

3.3 Conditional Probabilities and Conditional Expectations

In introductory probability, the conditional probability of B given A is defined to be

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)},$$

given that $\mathbb{P}(A) \neq 0$. Now the question is, what if the event A has probability zero? Consider an experiment in which we first draw a number from the interval $[0, 1]$ and then toss a fair coin. If we ask, what is the probability of getting a heads-up conditioning on that we draw the number $1/3$ in the first step? If we calculate according to the definition above, we know that the probability of drawing $1/3$ in the first step is zero. That is, we have a zero in the denominator if we want to calculate the conditional probability using the formula above. The conditional probability is not well defined! However, intuitively we know that the conditional probability should be $1/2$ because the probability of getting a heads-up in the second step is $1/2$ no matter what number you have drawn in the first step.

The above example indicates that we need a better definition of conditional probability

if we want to work with a probabilistic model in which there are events that has probability zero. Before we start to define conditional probabilities, we define conditional expectations first.

Let the probability space be $(\Omega, \mathcal{F}, \mathbb{P})$, X be a random variable, and $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$ (that is, $\mathcal{G} \subset \mathcal{F}$). The conditional expectation $\mathbb{E}(X|\mathcal{G})$ of X with respect to the σ-algebra $\mathcal{G}$ is a $\mathcal{G}$-measurable function (or in other words, a $\mathcal{G}$-measurable random variable) such that

$$\int_A X \mathrm{d}\mathbb{P} = \int_A \mathbb{E}(X|\mathcal{G}) \mathrm{d}\mathbb{P}$$

for all $A \in \mathcal{G}$. If Y is another random variable, then the expectation of X conditional on Y, denoted as $\mathbb{E}(X|Y)$ is just the conditional expectation of X with respect to $\sigma(Y)$.

Intuitively we may think conditional expectation as local average. To better see this, suppose that the probability space is $(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{F}$ is generated by a countable partition of Ω, that is, $\Omega = \bigcup_{i=1}^{\infty} D_i$, D_i's are disjoint, and $\mathcal{F} = \sigma(D_1, D_2, \ldots)$. Let X be a random variable. Since X is $\mathcal{F}$-measurable, it is constant on each piece of D_i. Let $\{C_i\}$ be a coarser partition of Ω, that is, each C_i is the union of some D_i's, $\Omega = \bigcup C_i$, and the C_i's are disjoint. Let $\mathcal{G} = \sigma(C_1, C_2, \ldots)$. Then $\mathbb{E}(X|\mathcal{G})$ is a $\mathcal{G}$-measurable function. That is, $\mathbb{E}(X|\mathcal{G})$ is constant on each piece of C_i. However, by the definition of conditional expectation, the value of $\mathbb{E}(X|\mathcal{G})$ on a piece of C_i must be the average (weighted by the probability) of values of X on the pieces that constitute C_i. For example, if $C_i = \bigcup_{j=1}^{\infty} D_{ij}$, and the value of X on D_{ij} is x_{ij}, then by definition,

$$\sum_{j=1}^{\infty} x_{ij} \mathbb{P}(D_{ij}) = \int_{C_i} X \mathrm{d}\mathbb{P} = \int_{C_i} \mathbb{E}(X|\mathcal{G}) \mathrm{d}\mathbb{P} = c\mathbb{P}(C_i),$$

where c is the constant value of $\mathbb{E}(X|\mathcal{G})$ on C_i. That is,

$$c = \frac{\sum_{j=1}^{\infty} x_{ij} \mathbb{P}(D_{ij})}{\mathbb{P}(C_i)}.$$

The followings are some properties of conditional expectations. Suppose that X and Y are random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$, $\mathcal{H} \subset \mathcal{G} \subset \mathcal{F}$, and a, b are two real numbers. We have

1. $\mathbb{E}(aX + bY|\mathcal{G}) = a\mathbb{E}(X|\mathcal{G}) + b\mathbb{E}(Y|\mathcal{G})$ a.s..
2. $\mathbb{E}[\mathbb{E}(X|\mathcal{G})|\mathcal{H}] = \mathbb{E}(X|\mathcal{H})$ a.s..
3. $\mathbb{E}[\mathbb{E}(X|\mathcal{H})|\mathcal{G}] = \mathbb{E}(X|\mathcal{H})$ a.s..
4. If X and Y are independent, then $\mathbb{E}(X|Y) = \mathbb{E}X$ a.s..
5. If Y is $\mathcal{G}$-measurable, $\mathbb{E}\,|X| < \infty$ and $\mathbb{E}\,|XY| < \infty$, then $\mathbb{E}(XY|\mathcal{G}) = Y\mathbb{E}(X|\mathcal{G})$ a.s..

Now we may use conditional expectations to define conditional probabilities. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\mathcal{G}$ be a sub-σ-algebra and $B \in \mathcal{F}$. The conditional probability $\mathbb{P}(B|\mathcal{G})$ is defined to be $\mathbb{E}(I_B|\mathcal{G})$. Under this definition, we have that for every $A \in \mathcal{G}$,

$$\int_A \mathbb{P}(B|\mathcal{G})\mathrm{d}\mathbb{P} = \int_A \mathbb{E}(I_B|\mathcal{G})\mathrm{d}\mathbb{P} = \int_A I_B \mathrm{d}\mathbb{P} = \mathbb{P}(A \cap B).$$

Loosely speaking, for each $\omega \in \Omega$ we may define a probability measure μ_ω on $\mathcal{B}(\mathbb{R})$ such that $\mu_\omega(A) = \mathbb{P}(X \in A|\mathcal{G})$ a.s. for every $A \in \mathcal{B}(\mathbb{R})$. Each μ_ω is called a conditional (probability) distribution of X given $\mathcal{G}$. In this case, we have

$$\mathbb{E}(X|\mathcal{G}) = \int x \mathrm{d}\mu_\omega.$$

3.4 Gaussian Systems

Let X be a random variable. We say that X is normal, or normally distributed, or Gaussian, if X has density function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

where μ and $\sigma \geq 0$ are two real numbers called respectively the mean and the standard deviation of the normal distribution. σ^2 is called the variance of the distribution.

Let X and Y be two random variables. The covariance of the two random variables is defined to be

$$\mathrm{Cov}(X,Y) = \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y).$$

It is known that $\mathrm{Cov}(X,Y) = \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y$. The variance $\mathrm{Var}(X)$ of a random variable X could be viewed as the covariance of X with itself. We have $\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2$.

If $\mathrm{Cov}(X,Y) = 0$, we say that X and Y are uncorrelated. If X and Y are independent with finite second moments, then X and Y are uncorrelated. However, if X and Y are uncorrelated, they are not necessarily independent.

There is a special case. If both X and Y are normal random variables, then they are independent if and only if they are uncorrelated.

Let $(X_1, X_2, \ldots, X_n)'$ be a random vector. The mean of the random vector is defined by

$$\mu = \begin{bmatrix} \mathbb{E}X_1 \\ \mathbb{E}X_2 \\ \vdots \\ \mathbb{E}X_n \end{bmatrix}.$$

The covariance matrix of the random vector is defined by

$$\Sigma = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1,X_2) & \cdots & \mathrm{Cov}(X_1,X_n) \\ \mathrm{Cov}(X_2,X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2,X_n) \\ \vdots & \vdots & & \vdots \\ \mathrm{Cov}(X_n,X_1) & \mathrm{Cov}(X_n,X_2) & \cdots & \mathrm{Var}(X_n) \end{bmatrix}.$$

A random vector X with mean μ and variance Σ is called normal, or Gaussian, if it has the joint density function

$$f(x) = \frac{1}{(2\pi)^{n/2}\det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right), \qquad x \in \mathbb{R}^n.$$

If a random vector $(X_1, \ldots, X_n)'$ is normal, then each X_i is normal. However, if a set of random variables $\{Y_1, \ldots, Y_n\}$ are normal, the random vector $(Y_1, Y_2, \ldots, Y_n)'$ is not necessarily normal. Once again, there is a special case. If $\{Y_1, \ldots, Y_n\}$ are normal and independent, then $(Y_1, Y_2, \ldots, Y_n)'$ is normal.

If $X = (X_1, \ldots, X_n)'$ is a normal random vector and A is a matrix from $\mathbb{R}^n$ to $\mathbb{R}^m$, then AX is a normal random vector that takes values in $\mathbb{R}^m$.

We denote a normal distribution with mean μ and variance Σ by $\mathbb{N}(\mu, \Sigma)$.

3.5 Convergence in Probability and Convergence in Distribution

Let X be a random variable and $X_1, X_2, \ldots$ be a sequence of random variables defined on a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say that the sequence converges to X in probability if for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(\{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \epsilon\}) = 0.$$

In this case, we write $X_n \to_p X$.

Let X be a random variable and $X_1, X_2, \ldots$ be a sequence of random variables. Let F and $F_1, F_2, \ldots$ be their corresponding distribution functions. We say that the sequence of random variables $X_1, X_2, \ldots$ converges in distribution to X if $F_n(x) \to F(x)$ for all x where F is continuous. In this case, we write $X_n \to_d X$.

Note that for a sequence of random variables $X_1, X_2, \ldots$ to converge in distribution to a random variable X, we do not need the random variables to be defined on the same probability space. This is because to determine $F, F_1, F_2, \ldots$, we only need to know the induced probability measures $\mathbf{P}, \mathbf{P}_1, \mathbf{P}_2, \ldots$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ instead of the original probability measures $\mathbb{P}, \mathbb{P}_1, \mathbb{P}_2, \ldots$.

It is well know that convergence in probability implies convergence in distribution. The other direction does not hold in general since a sequence of random variables that converges in distribution could be defined on different sample spaces.

3.6 The (Weak) Law of Large Numbers

If you would like to see whether a coin is fair, you may want to toss the coin many times, and see whether the number of heads is close to 1/2. And as you toss more and more times, you expect that your test becomes more and more precise.

This is the idea behind the law of large numbers. To state formally, let $X_1, X_2, \ldots$ be independent and identically distributed random variables defined on the same probability space. Then

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \to_p \mathbb{E}X.$$

3.7 The Central Limit Theorem

Suppose we have a sequence of random variables $X_1, X_2, \ldots$ that is independently and identically distributed with mean zero. Then the law of large numbers says that

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to_p 0.$$

Note that the sum shrunken by $1/n$ converges in probability to zero. We ask what if we shrunken by a slower speed? The central limit theorem says that if we normalize by $1/\sqrt{n}$, the sum converges in distribution to a normal distribution:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \to_d \mathbb{N}(0, \sigma^2)$$

where $\sigma^2 = \mathrm{Var}(X_1)$.