Introduction to Probability
(Lectures 3-4)

1 Multivariate Distributions, Independence, and Conditional Probability

We have already introduced the formal definition of a *real-valued* random variable. The objective of Lectures 3 and 4 is to provide a framework to think about *collections* of real-valued random variables.

Overview: We will define random vectors and multivariate c.d.f.s. We will define an absolutely continuous random vector, its p.d.f., its mean vector, and its covariance matrix. We will also present the definition of the moment generating function of a random vector $\mathbf{X}$ and state the Cramér-Wold Theorem. Finally, we discuss the notions of independence and conditional probability (technical details are relegated to the appendix). In the appendix, we will also talk about how to think about the distance between probability measures.

1.1 Vector-valued random variables and Multivariate Distributions

The data structures encountered in economics usually involve more than one real-valued random variable. Examples.

1. A time-series of stock prices: If the price of a stock in each period is subject to random fluctuations, it seems natural to think about describing/modeling such price as a random variable $X_t : \Omega \to \mathbb{R}$, where the randomness has been indexed by the time period t. The object of interest is thus the collection:

$$\{X_t\}_{t=1}^{T}.$$

   If all the real-valued random variables X_t are defined in the same probability space Ω then the collection $\{X_t\}_{t=1}^{T}$ induces a vector-valued map given by:

$$\mathbf{X} : \Omega \to \mathbb{R}^T, \quad \text{where} \quad \mathbf{X}(\omega) = (X_1(\omega), \dots X_T(\omega))'$$

2. A cross-section of stock prices: If the price of different stocks in a portfolio is subject to random variations we could describe/model such price as a random variable $X_i : \Omega \to \mathbb{R}$, where the randomness is now indexed by an identity label i. The object of interest is the collection:

$$\{X_i\}_{i=1}^{n}.$$

If all the real-valued random variables X_i are defined in the same probability space Ω then the collection $\{X_i\}_{i=1}^n$ induces a vector-valued map given by:

$$\mathbf{X} : \Omega \to \mathbb{R}^n, \quad \text{where} \quad \mathbf{X}(\omega) = (X_1(\omega), \dots X_n(\omega))'$$

Let $B(\mathbb{R}^S)$ denote the smallest σ-algebra containing the open sets in $\mathbb{R}^S$. Let $\{X_s\}_{s=1}^S$ be a collection of real-valued random variables defined over the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The main definitions of this subsection are as follows:

Definition ($\mathbb{R}^S$-valued random variable or $\mathbb{R}^S$-valued random vector [1])**.** The $\mathbb{R}^S$-valued mapping

$$\mathbf{X}(\omega) \equiv (X_1(\omega), \dots X_S(\omega))'$$

is a random vector if for all A in $B(\mathbb{R}^S)$

$$\mathbf{X}^{-1}(A) \in \mathcal{F}.$$

The c.d.f. of a random vector is defined as follows:

Definition (c.d.f. of an $\mathbb{R}^S$-valued random vector[2], based on **?**, p. 260)**.** The c.d.f. of an $\mathbb{R}^S$-valued random vector defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a function $F : \mathbb{R}^S \to \mathbb{R}$ given by:

$$F(x_1, x_2, \dots x_S) \equiv \mathbb{P}\Big[X^{-1}\Big((-\infty, x_1] \times (-\infty, x_2] \dots (-\infty, x_S]\Big)\Big].$$

The random vector structure allows us to consider "joint" probability statements. The properties of the c.d.f. that we have shown in Problem Set 1 will be verified here (once they are adjusted to the vector set-up). However, in the multivariate case the c.d.f. will satisfy additional properties:

Practice Problem 1 (Optional)**.** Let $S = 2$. Let $(x_1, x_2) \in \mathbb{R}^2$ and let $x_1^* \geq x_1$, $x_2^* \geq x_2$. Show that if F is the c.d.f. of an $\mathbb{R}^2$-valued random vector then:

$$F(x_1^*, x_2^*) - F(x_1^*, x_2) - F(x_1, x_2^*) + F(x_1, x_2) \geq 0.$$

[1] The term random vector is synonymous with measurable vector-valued function

[2] As with single-valued random variable, the *cdf* contains all the information about the distribution of the random variable

1.2 Absolutely Continuous Random Vectors

Just as the real-valued random variables, the $\mathbb{R}^S$-valued random vectors will be classified according to properties of the c.d.f. We will focus on the absolutely continuous case.

Definition (Absolutely continuous random vector)**.** An $\mathbb{R}^S$-valued random vector is said to be absolutely continuous if the c.d.f. can be written as

$$F(x_1, x_2, \dots x_S) \quad = \quad \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_S} f(z_1, \dots z_S) dz_1 \dots dz_S$$

for some nonnegative function $f : \mathbb{R}^S \to \mathbb{R}^+$. The function f is called the p.d.f. of the $\mathbb{R}^s$-valued random vector $\boldsymbol{X}$. This definition is based on the discussion at the bottom of p. 260 in **?**. Note that if F is differentiable $f(x_1, \dots, x_n) = \partial^n F(x_1, \dots, x_n)/\partial x_1 \dots \partial x_n$.

The joint distribution contains the information about the c.d.f. for each of the real-valued random variables in the collection $\{X_s\}_{s \in S}$. The c.d.f. for the real-valued random variable X_s is called the marginal c.d.f. and it is defined as:

Definition (Marginal c.d.f.)**.**

$$F_s(x) \equiv \mathbb{P}\Big[X^{-1}\Big(\mathbb{R} \times \dots (-\infty, x) \dots \times \mathbb{R}\Big)\Big]$$

This definition is based on the discussion at the bottom of p. 260 in **?**. The marginal c.d.f. is obtained obtained by fixing $x_s = x$ and then taking the other arguments of the multivariate c.d.f. to infinity. Hence, in the case of random vectors with a p.d.f. the marginal distributions are given by:

$$F_s(y) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{y} \dots \int_{-\infty}^{\infty} f(z_1, \dots z_s \dots z_S) dz_1 \dots dz_S$$

The real-valued random variable X_s has a *marginal* p.d.f. given by:

$$f_s(x_s) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(z_1, z_2, \dots x_s, \dots z_S) dz_1 \dots dz_{s-1} \dots dz_{s+1} \dots dz_S$$

1.2.1 Expectation of Random Vectors and Covariance Matrix

Let $g : \mathbb{R}^S \to \mathbb{R}^m$. Write

$$g(\mathbf{x}) = (g_1(\mathbf{x}), g_2(\mathbf{x}), \dots g_m(\mathbf{x}))'$$

and let

$$\begin{aligned} \mathbb{E}_F[g(\mathbf{X})] &= \Big(\mathbb{E}_F[g_1(\mathbf{X})], \mathbb{E}_F[g_2(\mathbf{X})], \dots, \mathbb{E}_F[g_m(\mathbf{X})]\Big)' \\ &\equiv \Big(\int_{\mathbb{R}^S} g_1(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}, \int_{\mathbb{R}^S} g_2(\mathbf{x})) f(\mathbf{x}) d\mathbf{x}, \dots, \int_{\mathbb{R}^S} g_m(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}\Big)' \end{aligned}$$

We refer to $\mathbb{E}_F[\mathbf{X}] \in \mathbb{R}^S$ as the expectation (or mean vector) of $\mathbf{X}$.

1. (Covariance i-j) Let X_i, X_j be the ith and jth entry of the random vector $\mathbf{X}$. Consider the mapping:

   $$g(\mathbf{X}) = (X_i - \mathbb{E}_F[X_i])(X_j - \mathbb{E}_F[X_j]).$$

   We will call $\mathbb{E}_F\Big[(X_i - \mathbb{E}_F[X_i])(X_j - \mathbb{E}_F[X_j])\Big]$ the *covariance* between X_i and X_j.

2. (Covariance Matrix) Let X be an $\mathbb{R}^S$-valued random vector. Consider the matrix of expected values given by:

   $$\Sigma \equiv \mathbb{E}_F\Big[(\mathbf{X} - \mathbb{E}_F[\mathbf{X}])(\mathbf{X} - \mathbb{E}_F[\mathbf{X}])'\Big] \in \mathbb{R}^{S \times S}$$

   Note that the j-th diagonal element of the matrix Σ contains the variance of the real-valued random variable X_j. The ith-jth element of the matrix Σ contains the covariance between X_i and X_j. The matrix Σ is usually called the variance matrix of a random vector $\mathbf{X}$. It can also be called *covariance matrix* or *variance-covariance* matrix.

1.2.2 Moment Generating Function of Random Vectors and Cramér-Wold Theorem

Definition (Moment Generating Function)**.** Let $\mathbf{X}$ be an $\mathbb{R}^S$ random vector with c.d.f. F. The moment generating function of $m_{\mathbf{X}} : \mathbb{R}^S \to \mathbb{R}$ is given by:

$$m_{\mathbf{X}}(t) \equiv \mathbb{E}_F[\exp(t'\mathbf{X})] \quad t \in \mathbb{R}^S$$

The usefulness of the m.g.f is the same as in the real-valued case: two random vectors with the same m.g.f. have the same c.d.f. The m.g.f of a random vector is completely determined by the distribution of its linear combinations. This is not a surprise in light of the following theorem:

Theorem 1 (Cramér-Wold Theorem)**.** *Let μ_1, μ_2 be probability measures defined on the Borel σ-algebra of $\mathbb{R}^S$. Suppose that for all $t \in \mathbb{R}^S$ and $z \in \mathbb{R}$*

$$\mu_1\Big\{x \in \mathbb{R}^S \;\Big|\; \sum_{s=1}^{S} t_s x_s \leq z\Big\} = \mu_2\Big\{x \in \mathbb{R}^S \;\Big|\; \sum_{s=1}^{S} t_s x_s \leq z\Big\}.$$

Then $\mu_1(A) = \mu_2(A)$ for all (measurable) sets A.

Practice Problem 2 (Optional)**.** The second exercise in this week's problem set will ask you to go over the proof of this result. In doing so, you will see the definition of the *characteristic function* of a random vector.

1.2.3 A common example of an absolutely continuous random vector: Bivariate Normal Distribution

To make ideas more concrete, this subsection introduces the bivariate normal distribution.

The following definition of a Bivariate Normal Distribution is based on **?** pg. 173 and uses the moment generating function (other definitions, for example **?** and **?**, use the characteristic function instead).

Definition (Bivariate Normal Distribution with parameters μ and Σ)**.** Let $\mu \in \mathbb{R}^2$ and let Σ be a positive semi-definite matrix of dimension 2×2. A $\mathbb{R}^2$-valued random vector X is said to have a bivariate normal distribution, denoted $\mathbf{X} \sim \mathcal{N}_2(\mu, \Sigma)$, if:

$$\mathbb{E}_F[\exp(t'\mathbf{X})] = \exp\Big(t'\mu + \frac{1}{2}t'\Sigma t\Big).$$

As you probably know if $\mathbf{X} \sim \mathcal{N}_2(\mu, \Sigma)$ then

$$\mathbb{E}_F[\mathbf{X}] = \mu \quad \text{and} \quad \mathbb{E}_F[(\mathbf{X}-\mu)(\mathbf{X}-\mu)'] = \Sigma.$$

Here are three important results that follow from the definition of a bivariate normal vector:

Result 1. *Let $\mu \in \mathbb{R}^2, \Sigma \in \mathbb{R}^{2\times 2}$. Let A be a 2×2 matrix such that $\Sigma = AA'$ and suppose $\mathbf{Z} \sim \mathcal{N}_2(\boldsymbol{0}, \mathbb{I}_2)$. Then*

$$\mathbf{Y} = \mu + A\boldsymbol{Z} \sim \mathcal{N}_2(\mu, \Sigma).$$

Result 2 (Linear Combinations of a Bivariate Normal characterize Univariate Normal)**.** $\mathbf{X}$ *is a bivariate normal distribution with parameters μ and Σ if and only if for all $c \in \mathbb{R}^2$*

$$c'\mathbf{X} \sim \mathcal{N}_1(c'\mu, c'\Sigma c).$$

Result 3. *If Σ is invertible and $\mathbf{X}$ is bivariate normal with parameters μ and Σ then the random vector $\mathbf{X}$ has density*

$$f(\mathbf{x}) = \frac{1}{(det\, 2\pi\Sigma)^{1/2}} \exp\Big(-\frac{1}{2}(\mathbf{x}-\mu)'\boldsymbol{\Sigma^{-1}}(\mathbf{x}-\mu)\Big).$$

Bivariate and multivariate normal distributions will be important for this course. In this week's problem set I will ask you to work out some specific properties of this distribution. But before doing so, consider the following problem.

Practice Problem 3 (Best Linear Predictor)**.** Let X and Y be two real-valued random variables. Let

$$\mu \equiv \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \text{ and } \Sigma,$$

denote the mean and covariance of $(X, Y)'$. Both μ and Σ are assumed to be known. Suppose that we are interested in linearly "predicting" Y using the deviations of X from its mean. That is, we would like to predict Y using a function of the form

$$\alpha + \beta(X - \mu_x).$$

This is a purely probabilistic problem. There is no data, no sample, and no econometrics: just two-random variables. To assess the quality of the prediction, it seems natural to report the (squared)-distance between Y and its prediction (typically referred to as the *squared error*)

$$(Y - \alpha - \beta(X - \mu_x)).^2$$

The squared error is a random variable, which means that sometimes it will be big and sometimes it will be small. To summarize the performance of the linear predictor we can compute the average or mean squared-error:

$$\mathbb{E}[(Y - \alpha - \beta(X - \mu_x))^2].$$

The "best" linear predictor of Y in terms of $(X-\mu_x)$ is defined as the linear function $\alpha^*+\beta^*(X-\mu_x)$ that minimizes the mean squared-error; i.e., the function with parameters (α^*, β^*) such that:

$$(\alpha^*, \beta^*) \in \text{argmin}_{(\alpha,\beta)} \mathbb{E}[(Y - \alpha - \beta(X - \mu_x))^2].$$

1. Show first that

   $$\alpha^* = \mu_y.$$

   This result implies that if X happens to be exactly equal to its mean, then the best linear predictor of Y is its expectation.

2. Show then that β^* solves the problem

   $$\min_{\beta} \mathbb{E}[((Y - \mu_y) - \beta(X - \mu_x))^2],$$

   and therefore

   $$\beta^* = Cov(X, Y)/V(X).$$

HINT: The objective function is quadratic in β and the choice set for β is convex. The problem is thus convex and the first order conditions are necessary and sufficient. Moreover, $\frac{\partial}{\partial\beta}\mathbb{E}[((Y-\mu_y)-\beta(X-\mu_x))^2] = \mathbb{E}[\frac{\partial}{\partial\beta}((Y-\mu_y)-\beta(X-\mu_x))^2]$.

3. Finally, define the error of the best linear predictor as:

$$\epsilon = Y - \alpha^* - \beta^*(X-\mu).$$

Show that $\mathbb{E}[\epsilon] = 0$ and $Cov(\epsilon, X) = 0$.

4. (OPTIONAL) Suppose now that instead of being real-valued, X is a random vector of dimension k. Suppose that we look for the coefficients α and $\beta = (\beta_1, \ldots, \beta_k)$ of the best linear predictor

$$\alpha + \beta_1(X_1 - \mu_1) + \beta_2(X_2 - \mu_2) + \ldots + \beta_k(X_k - \mu_k).$$

Is it true that the coefficients of the best linear predictor are:

$$\alpha^* = \mu_y, \quad \beta = V(X)^{-1}Cov(X,Y)?$$

1.3 Independence

This section discusses the notion of *independence*. In the past, you have probably been exposed to the notion of "independence of events" and "independence of random variables". The appendix connects these definitions with the more general notion of independence of σ-algebras. The material therein presented is loosely based on **?**, Chapter 2, pgs. 41-47 (there is no need to take a look at it). Here we focus on presenting some common and useful characterizations of independence.

1.3.1 Useful characterizations of independence

Below we provide useful characterizations of independence. We will not provide proofs: instead we give some references.

1. **Independence of finite collections of random variables and the c.d.f.**: A collection of real-valued random variables $(X_1, X_2, \ldots X_S)$ is independent if and only if

$$F(x_1, x_2, \ldots, x_S) = F_1(x_1)F_2(x_2)\ldots F_S(x_S).$$

See Theorem 2.1.4 in **?**, pg. 44 (which connects to the definition of independence provided in the Appendix)

2. **Independence of finite collections of random variables and the p.d.f.s**: A collection of real-valued random variables $(X_1, X_2, \dots X_S)$ with joint p.d.f. f is independent if and only if

$$f(x_1, x_2, \dots, x_S) = f_1(x_1) f_2(x_2) \dots f_S(x_S)$$

The second characterization follows from the first one using the fact that the $f(x_1, \dots, x_n) = \partial^n F(x_1, \dots, x_n)/\partial x_1 \dots \partial x_n$. Another version of this result follows from Exercise 2.1.4. in **?**, pg. 44.

3. **Independence of finite collections of random variables and products of expectations:** A collection of real-valued random variables $(X_1, X_2, \dots X_S)$ is independent if and only if for any measurable real-valued functions $g_1, g_2, \dots g_n$, (with finite expectation)

$$\mathbb{E}_F\Big[g_1(X_1) g_2(X_2) \dots g_n(X_S)\Big] = \mathbb{E}_{F_1}\Big[g_1(X_1)\Big]\mathbb{E}_{F_2}\Big[g_2(X_2)\Big] \dots \mathbb{E}_{F_S}\Big[g_n(X_S)\Big]$$

One side of this implication (independence implies product of expectations) follows from 1). See also Theorem 2.1.8 in **?**. For the other side take indicator functions for sets of the form $(-\infty, x_1]$ and note that:

$$\mathbb{E}_{F_1}\Big[\mathbf{1}_{(-\infty, x_1]}(X_1)\Big] = F_1(x_1)$$

4. **Independence and moment generating functions:** A collection of real-valued random variables $(X_1, X_2, \dots X_S)$ with well-defined MGFs is independent if and only if

$$\mathbb{E}_F\Big[\exp\Big(t'\mathbf{X}\Big)\Big] = \mathbb{E}_F[\exp(t_1 X_1)] \dots \mathbb{E}_F[\exp(t_2 X_2)]$$

See Theorem 2.5.5 pg. 112 in **?** and the discussion in pg. 119 Section 2.6.

5. **Independence and covariance**: If X and Y are two independent real-valued random variables then:

$$Cov(X, Y) = \mathbb{E}_F[(X - \mu_X)(Y - \mu_X)] = 0$$

The converse, in general, is not true. You will work this out in the next practice problem:

Practice Problem 4.

1. Show that if $(X, Y)'$ are bivariate normal random variables $Cov(X, Y) = 0$ implies independence. HINT: Use the definition of bivariate normal and any of the characterizations of independence we presented above.

2. Let $X \sim \mathcal{N}(0, 1)$. Let $Y = X^2$. Show that the $Cov(X, Y) = 0$. Are (X, Y) independent?

1.4 Conditional Probability and Conditional Expectation

Given a pair of real-valued random variables (X, Y) we have learnt how to compute and how to interpret a joint probability statement like:

$$\mathbb{P}_{XY}(X \leq x, Y \leq y)$$

In this section we will introduce the conditional probability function and the conditional expectation function. There are at least two different ways of presenting the definitions of conditional probability and conditional expectation.

- **Conditioning with respect to a σ-algebra**: The formal textbook definition of both conditional probability and conditional expectation uses σ-algebras as conditioning structures. For instance, if you go to **?** [pg. 430, Equation 33.8] you will find the definition of the conditional probability of A given a σ-algebra $\mathcal{G}$ denoted:

  $$\mathbb{P}(A|\mathcal{G})$$

  Likewise, the conditional expectation of a random variable X given a σ-algebra $\mathcal{G}$ is defined in [pg. 445, Equation 34.1] and denoted:

  $$\mathbb{E}[X|\mathcal{G}]$$

  $\mathbb{P}(A|\mathcal{G})$ and $\mathbb{E}[A|\mathcal{G}]$ are defined as $\mathcal{G}$-measurable random variables that satisfy an "integral equation". Defining such integral equations usually requires the notion of Lebesgue integration on product spaces, which we did not even mention in class. Therefore, we will not present conditional probability and conditional expectation in all its generality. Instead, we present a simple and more applied definition. Billingsley writes:

  > *"The concepts of conditional probability and expected value with respect to a σ-field underlie much of modern probability theory. The difficulty in understanding these ideas has to do not with mathematical detail so much as with probabilistic meaning, and the way to get at this meaning is through calculations and examples ...*

  So, we will try to go over some calculations and examples.

- **Conditional expectation as a projection:** An alternative way of introducing the definition of conditioning is through the idea of an orthogonal projection (over a space of functions). Such definition of conditioning is, perhaps, closer to what we have seen before. The main

idea is as follows. Given two random variables (X, Y) with finite second moments, $E[Y|X]$ is defined as a $\sigma(X)$-measurable function $h(X)$ such that:

$$h(\cdot) \in \arg\min_h \mathbb{E}_{\mathbb{P}}[(Y - h(X))^2]$$

We will work this out as a practice problem, but we will not use this as the definition of conditional expectation.

1.4.1 The conditional probability of Y given X

For the sake of exposition consider an $\mathbb{R}^2$-valued random vector (X, Y). We would like to define the conditional probability of the event $\{\omega \,|Y(\omega) \leq y\}$ as a function of the realizations of x.[3]

1. **Conditional Probability function for continuous** (X, Y): Let (X, Y) be an $\mathbb{R}^2$ random vector with p.d.f. $f(x, y)$. For any $y^* \in \mathbb{R}$, the conditional probability of the event $\{Y \leq y^*\}$ given x, denoted

$$\mathbb{P}_{Y|X}(Y \leq y^*|x),$$

is defined as a function satisfying the following integral equation: for any $x^* \in \mathbb{R}$

$$\int_{-\infty}^{x^*} \mathbb{P}_{Y|X}(Y \leq y^*|x) f_X(x) dx = \mathbb{P}_{XY}(X \leq x^*, Y \leq y^*). \tag{1.1}$$

The conditional probability is thus defined in reference to joint probability statements. Interestingly, there is (almost surely) a unique function satisfying the restriction (1.1). The function is given by:

$$\mathbb{P}_{Y|X}(Y \leq y^*|x) \equiv \int_{-\infty}^{y^*} \frac{f(x, y)}{f_X(x)} dy \tag{1.2}$$

This function is called the conditional c.d.f. of Y given X and we call

$$\frac{f(x, y)}{f_X(x)} \tag{1.3}$$

the conditional density of $Y|X$ and we denote it by $f(y|x)$. By definition, we can always write the joint density as

$$f(x, y) = f(y|x) f_X(x).$$

[3]Note that to give these explicit formulas, we need to treat separately absolutely continuous random variables (which have pdfs) and discrete random variables (which do not).

Thus, if X and Y are independent $f(y|x) = f_Y(y)$.

Practice Problem 5. Consider the same set-up as in the practice problem 3 where we introduced the notion of best linear predictor. We make the additional assumption that $(X, Y)' \sim \mathcal{N}_2((\mu_x, \mu_y)', \Sigma)$. What is the conditional distribution of $Y|X$?

HINT: There are two ways of doing this problem. One of them is long: use the formula of conditional density in (1.3) and do some algebra. The other one is short(er). Write:

$$Y = \alpha^* + \beta^*(X - \mu_x) + \epsilon, \quad \epsilon \equiv Y - \alpha^* - \beta^*(X - \mu_x),$$

and apply the results in practice problem 3.

1.4.2 The Conditional Expectation of $g(X)$ given Y

To close this section we will define the conditional expectation function of the transformation $g(X)$ given Y.[4] The conditional expectation of $g(Y)$ given X is a function mapping the values of X into the real line:

$$\mathbb{E}_{\mathbb{P}}[g(Y)|\cdot] : \mathbf{X} \to \mathbb{R}.$$

1. **Conditional Expectation function with a p.d.f. for** (X, Y): Let (X, Y) be an $\mathbb{R}^2$ random vector with p.d.f. $f(x, y)$. For any $y \in \mathbb{R}$, the conditional expectation function is defined as:

$$\mathbb{E}_f[g(Y) \mid x] = \int_{-\infty}^{\infty} g(x) \frac{f(x, y)}{f_Y(y)} dx \tag{1.4}$$

   One of the properties of conditional expectations that you will use quite often is the Law of Iterated Expectation (L.I.E.). Broadly speaking, the L.I.E. relates the expectation of the random variable $\mathbb{E}_{\mathbb{P}}[g(Y)|X]$ with the expectation of $\mathbb{E}_{\mathbb{P}}[g(Y)]$. The L.I.E. states that:

$$\mathbb{E}[\mathbb{E}_{\mathbb{P}}[g(Y) \mid x]] = \mathbb{E}_{\mathbb{P}}[g(Y)]. \tag{1.5}$$

   We can verify this property using (1.4).

[4] The general definition of conditional expectation requires a σ-algebra as a conditioning structure. We sidestep this construction by defining conditional expectation as an integral based on the conditional probability functions we have computed before.

OPTIONAL: Take a look at the definition of conditional expectation in pg. 445, Section 34 of **?**. Is it true that the L.I.E:

$$\mathbb{E}[\mathbb{E}[X|\mathcal{F}]] = \mathbb{E}[X]$$

holds by definition?

A Kolmogorov's independence, Independence of Events, and Independence of Random Variables

Definition. (**Independence of σ-algebras**) Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We say that a collection $\{\mathcal{F}_i\}_{i\in I}$ $\mathcal{F}_i \subseteq \mathcal{F}$ of σ-algebras is independent if for any finite set $\{i_1, i_2, \dots i_N\}$ contained in I, the following holds:

$$\mathbb{P}(A_{i1} \cap A_{i2} \dots \cap A_{1N}) = \mathbb{P}(A_{i1})P(A_{i2}) \dots P(A_{iN})$$

for any $(A_{i1}, A_{i2}, \dots A_{iN}) \in \times_{n=1}^{N} \mathcal{F}_{in}$.

The notion of independence refers to a collection of events: σ-algebras. The formal definition require these σ-algebras to be defined on the sample probability space.

We would like to relate the abstract definition of independence with other notions that are more common (and more useful!).

Definition. (**Independence of events**) We say that a collection of events $\{A_i\}_{i\in I}$ is independent if the collection of σ-algebras $\{\sigma(A_i)\}_{i\in I}$ is independent.

As you already know

$$\sigma(\mathcal{F}_i) = \{\Omega, \emptyset, A_i, A_i^c\} \quad \text{(i.e., the smallest } \sigma\text{-algebra containing } A_i)$$

Proposition 1 (Characterization of independence for pairs of events)**.** *Let $A, B \in \mathcal{F}$. Event A is independent of event B (in the sense of the definition above) if and only if*

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

Proof. We only show that if $\mathbb{P}(A\cap B) = \mathbb{P}(A)\mathbb{P}(B)$, then the events are independent in the sense of the definition above. Take A and B^c. $\mathbb{P}(A\cap B^c)$ equals $\mathbb{P}(A\backslash(A\cap B))$, which equals $\mathbb{P}(A)-\mathbb{P}(A\cap B) = \mathbb{P}(A) - \mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(A)(1-\mathbb{P}(B))$. This implies that $\mathbb{P}(A \cap B^c) = \mathbb{P}(A)(1-\mathbb{P}(B))$. Analogously, $\mathbb{P}(A^c \cap B) = (1-\mathbb{P}(A))\mathbb{P}(B)$. □

An important observation: Let A, B, C be three events. Suppose that (A, B), (A, C), (B, C) are pairwise independent. The events (A, B, C) need not be independent. Consider the following example:

$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}, \quad P(\omega_i) = \frac{1}{4}$$

$$A = \{\omega_1, \omega_2\}, \quad B = \{\omega_2, \omega_3\}, \quad C = \{\omega_1, \omega_3\}$$

Note that the collection of events A, B, C is pairwise independent. In order to see this, note that

$$\mathbb{P}(A \cap B) = \frac{1}{4} = \mathbb{P}(A)\mathbb{P}(B)$$
$$\mathbb{P}(B \cap C) = \frac{1}{4} = \mathbb{P}(B)\mathbb{P}(C)$$
$$\mathbb{P}(A \cap C) = \frac{1}{4} = \mathbb{P}(A)\mathbb{P}(C)$$

However, $\mathbb{P}(A \cap B \cap C) = 0 \neq \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$.

The example above might or might not surprise you. In the latter case, you are probably thinking that a condition like

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$$

should deliver independence. However, such definition does not even imply that the events are pairwise independent. We will give a very simple example. First, modify the probability measure in the previous example to get:

$$\mathbb{P}(\omega_1) = \frac{1}{2} = \mathbb{P}(\omega_4), \quad \mathbb{P}(\omega_2) = \mathbb{P}(\omega_3) = 0$$

Note that $\mathbb{P}(A \cap B \cap C) = 0 = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$, yet:

$$\mathbb{P}(A \cap C) = \mathbb{P}(\{\omega_1\}) = \frac{1}{2} \neq \mathbb{P}(A)\mathbb{P}(C) = \frac{1}{4}$$

The following practice exercise asks you to prove (or disprove) the following claim:

Practice Problem 6. If $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$ and the events (A, B, C) are pairwise independent then the events (A, B, C) are independent.

He have presented the definition of independent events. In the remaining part of this section we focus on the relation between the independence of σ-algebras and the independence of random variables. Let $X_1, X_2, \ldots, X_S$ be a collection of random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Define:

$$\sigma(X) \equiv \sigma\{F \in \mathcal{F} \mid X^{-1}(A) = F \quad \text{for some } A \in B(\mathbb{R}^S)\}$$

The collection $\sigma(X)$ is called the σ-algebra generated by a random variable X.

Definition (Independence of Random Variables)**.** Let $(X_1, X_2, \ldots)$ be a collection of real-valued

random variables. We say that $(X_1, X_2, \ldots)$ are independent (or jointly independent) if the collection of σ-algebras $\sigma(X_1), \sigma(X_2) \ldots$ are independent.

B Conditional Probability

B.1 Preliminaries

Definition (Conditional probability given an event)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $B \in \mathcal{F}$, $\mathbb{P}(B) > 0$. For a fixed event $A \in \mathcal{F}$ define the conditional probability of A given B as:

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

As you can imagine, we can interpret conditional probability as a new measure over $(\Omega, \mathcal{F})$. For instance, for a fixed $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$ we could define the set function:

$$\mathbb{P}_B : \mathcal{F} \to [0, 1]$$

as $\mathbb{P}_B(F) \equiv \mathbb{P}(F|B)$ for all $F \in \mathcal{F}$. It is true that the set function $\mathbb{P}_B$ is a probability measure over $(\Omega, \mathcal{F})$ (remember that all we need to show is normalization and σ-additivity). However, this interpretation is not going to take us very far. Conditional probability will be better described as a random variable and not as probability measure.

Definition (Conditional probability given the σ-algebra generated by a partition[5])**.** Let $\mathcal{B} \equiv \{B_1, B_2, \ldots B_K\}$ be a finite partition of Ω, with $B_k \in \mathcal{F}$, $\mathbb{P}(B_k) > 0$. For a fixed event $A \in \Omega$ define:

$$\mathbb{P}(A|\, \sigma(\mathcal{B})\,) : \Omega \to [0, 1]$$

as a random variable satisfying:

1. $P(A \mid \sigma(\mathcal{B})\,)(\omega) = P_k \quad \forall \quad \omega \in B_k$
2. For any $\{k_1, k_2, \ldots k_n\} \subseteq \{1, 2 \ldots K\}$

$$\sum_{i=1}^{n} \mathbb{P}(B_{k_i}) P_{k_i} = \mathbb{P}(A \cap (\cup_{i=1}^{n} B_{k_i}))$$

where $\mathbb{P}_{k_i} = \mathbb{P}(A|B_{k_i})$

[5]For further discussion of conditional probability and expectation, please see **?**

The first restriction in the definition above is equivalent to requiring the function $\mathbb{P}(A|\ \sigma(\mathcal{B})\ )$ to be measurable with respect to $\sigma(\mathcal{B})$. In a sense, whenever we condition, we restrict ourselves to make probability statements dependent only on the information available on the conditioning set. In this definition the information available is given by the partition $\mathcal{B}$.

The second restriction relates the values of the conditional probability (P_{k_i}) with joint probability statements by "integrating" over different collections of elements of the partition. In particular, note that one can select a singleton $\{k\}$ and show that by 1 and 2:

$$P_k = \mathbb{P}(A \cap B_k)/\mathbb{P}(B_k),$$

Also, by selecting the set the whole set $\{1, 2, \dots K\}$ we get the formula

$$\sum_{k=1}^{K} \mathbb{P}(B_k)P_k = \mathbb{P}(A \cap \Omega) = \mathbb{P}(A)$$

C Distance between probability measures

Every now and then it will be convenient to think about "how far" a probability measure P is from a probability measure Q. An extremely popular and useful way to think about $d(P, Q)$ is the Bounded Lipschitz metric.

C.1 Bounded Lipschitz Distance

Let X be a real-valued random variable with induced probability F_X, and let Y and F_Y be defined analogously. Define the class of "Bounded Lipschitz functions" with Lipschitz constant 1 as

$$BL(1) := \left\{ h : \mathbb{R} \to \mathbb{R} \mid |h(x) - h(y)| \leq |x - y| \text{ and } \sup_{x \in \mathbb{R}} |h(x)| \leq 1 \right\} \tag{C.1}$$

Definition: (Bounded Lipschitz Distance, p.395 Dudley (RAP))

$$d_{\mathrm{BL}}(F_X, F_Y) := \sup_{h \in BL(1)} |\mathbb{E}_{F_X}[h(X)] - \mathbb{E}_{F_Y}[h(Y)]| \tag{C.2}$$

Thus we will say that F_X and F_Y are close to each other (in the BL sense) if $d_{\mathrm{BL}}(F_X, F_Y)$ is small.

C.1.1 Examples

Example: Let $X \sim \text{Bernoulli}(p)$ and $Y \sim \text{Bernoulli}(q)$. When are F_X and F_Y close to each other?

$$\begin{aligned}|\mathbb{E}_{F_X}[h(X)] - \mathbb{E}_{F_Y}[h(Y)]| &= |ph(1) + (1-p)h(0) - qh(1) - (1-q)h(0)| \\ &= |(p-q)h(1) - (p-q)h(0)| \\ &= |(p-q)(h(1) - h(0))| \\ &\leq |p-q|\end{aligned}$$

Optional: Let $X \sim N(\mu_X, 1)$ and $Y \sim N(\mu_Y, 1)$. When are F_X and F_Y close to each other (in the BL sense)?

C.1.2 Main Result

We now show that whenever P and Q are close to each other in the BL sense, then $P(A)$ and $Q(A)$ are close to each other provided that A is a "continuity set" under either P or Q. Define the 'δ-expansion of A' as

$$A^\delta := \left\{ x \in \mathbb{R} \mid \inf_{y \in A} |x - y| < \delta \right\} \tag{C.3}$$

Say that A is a continuity set over a probability measure A defined on the real line if, for every $\varepsilon > 0$, there exists $\delta_\varepsilon > 0$ such that

$$Q(A^{\delta_\varepsilon} \backslash A) < \varepsilon \quad \text{and} \quad Q((A^c)^{\delta_\varepsilon} \backslash A^c) < \varepsilon \tag{C.4}$$

Define also the function

$$f_{A,\delta} := \max\{0, 1 - [d(x, A)/\delta]\} \tag{C.5}$$

where

$$d(x, A) := \inf_{y \in A} d(x, y) \tag{C.6}$$

Claim: (Homework, Optional) If $0 < \delta < 1$, then $(\delta)(f_{A,\delta}) : \mathbb{R} \to \mathbb{R} \in BL(1)$.

We use the fact that $\delta f_{A,\delta}$ is Lipschitz to prove the following.

Claim: (Homework, Optional)

$$-Q((A^c)^\delta \backslash A^c) - \frac{1}{\delta} d_{\text{BL}}(P, Q) \leq P(A) - Q(A) \leq Q(A^\delta \backslash A) + \frac{1}{\delta} d_{\text{BL}}(P, Q) \tag{C.7}$$

Thus, if A is a continuity set of Q, then, for all $\varepsilon > 0$, there exists a $\delta > 0$ such that

$$-\frac{\varepsilon}{2} - \frac{1}{\delta_{\varepsilon/2}} d_{\mathrm{BL}}(P, Q) \leq P(A) - Q(A) \leq \frac{1}{\delta_{\varepsilon/2}} d_{\mathrm{BL}}(P, Q) + \frac{\varepsilon}{2} \tag{C.8}$$