Chapter 3: Multivariate Random Variables

By Joan Llull[*]

Probability and Statistics.
QEM Erasmus Mundus Master. Fall 2015

Main references:

— Mood: IV: 1.2, I:3.6, IV:3, V:6, IV:4.1, IV:5, IV:4.2, IV:4.6, IV:4.3

— Lindgren: 3.1, 3.4, 2.7, 3.7, 12.1, 12.4, 4.3, 12.2, 12.6, 4.7, 4.8, 4.2, 4.5, 12.3

I. Joint and Marginal Distributions

In this chapter we will work with random vectors, which include a collection of (scalar) random variables. We call these vectors ***multivariate*** random variables. For them, we will define a ***joint*** cumulative density function. Let $X1, ..., X_K$ denote a collection of K random variables. The joint cdf is defined as:

$$F_{X_1 \dots X_K}(x_1, ..., x_K) \equiv P(X_1 \leq x_1, X_2 \leq x_2, ..., X_K \leq x_K). \tag{1}$$

When the random variables are discrete, we can define a joint probability mass function given by:

$$P(X_1 = x_1, X_2 = x_2, ..., X_K = x_K). \tag{2}$$

For example, consider the case of tossing two coins. Let $\Omega = \{head, tail\} \times \{head, tail\}$. Define the following two random variables:

$$X_1 = \begin{cases} 1 & \text{if } \omega = \{(head, head)\} \\ 0 & \text{otherwise} \end{cases} \quad X_2 = \begin{cases} 1 & \text{if } \omega = \{(x, y) : \{x\} = \{y\}\} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

In words, X_1 equals one if we obtain two heads, zero otherwise, and X_2 equals one if we obtain the same outcome with the two coins (either both heads or both tails), zero otherwise. Note that, in this case, the pmf is:

$$P(X_1 = x_1, X_2 = x_2) = \begin{cases} \frac{2}{4} = \frac{1}{2} & \text{if } x_1 = 0, x_2 = 0 \\ \frac{1}{4} & \text{if } x_1 = 0, x_2 = 1 \\ \frac{1}{4} & \text{if } x_1 = 1, x_2 = 1 \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

[*] Departament d'Economia i Història Econòmica. Universitat Autònoma de Barcelona. Facultat d'Economia, Edifici B, Campus de Bellaterra, 08193, Cerdanyola del Vallès, Barcelona (Spain). E-mail: joan.llull[at]movebarcelona[dot]eu. URL: http://pareto.uab.cat/jllull.

(note the connection with the joint relative frequency in Chapter 1) and the cdf is:

$$F_{X_1X_2}(x_1,x_2)=\begin{cases}0 & \text{if } x_1<0 \text{ or } x_2<0\\ \frac{1}{2} & \text{if } x_1\leq 0 \text{ and } 0\leq x_2<1\\ \frac{3}{4} & \text{if } 0\leq x_1<1, x_2\geq 1\\ 1 & \text{if } x_1\geq 1, x_2\geq 1.\end{cases} \tag{5}$$

In the case of continuous variables, we have a joint probability density function, $f_{X_1...X_K}(x_1,...,x_K)$, which is implicitly defined as:

$$F_{X_1...X_K}(x_1,...,x_K)\equiv\int_{-\infty}^{x_1}...\int_{-\infty}^{x_K}f_{X_1...X_K}(z_1,...,z_K)dz_1...dz_K. \tag{6}$$

A joint pdf satisfies the following properties:

- $f_{X_1...X_K}(x_1,...,x_K)\geq 0$ for all $x_1,...,x_K$.
- $F_{X_1...X_K}(\infty,...,\infty)=\int_{-\infty}^{\infty}...\int_{-\infty}^{\infty}f_{X_1...X_K}(z_1,...,z_K)dz_1...dz_K=1$.
- Probabilities:
  - $P(a_1\leq X_1\leq b_1,...,a_K\leq X_K\leq b_k)=\int_{a_1}^{b_1}...\int_{a_K}^{b_K}f_{X_1...X_K}(z_1,...,z_K)dz_1...dz_K$.
  - $P(X_1=a_1,...,X_K=a_K)=0$.
  - $P(X_1=a,a_2\leq X_2\leq b_2,...,a_K\leq X_K\leq b_K)=0$.
- $\dfrac{\partial^K}{\partial x_1...\partial x_K}F_{X_1...X_K}(\cdot)=f(\cdot)$.

For example, the following is a pdf of a bivariate continuous random variable:

$$f_{XY}(x,y)=\begin{cases}\frac{3}{11}(x^2+y) & \text{if } 0\leq x\leq 2, 0\leq y\leq 1\\ 0 & \text{otherwise.}\end{cases} \tag{7}$$

In this example, the cdf is:

$$\begin{aligned}F_{XY}(x,y)&=\begin{cases}\int_0^{\min\{2,x\}}\int_0^{\min\{1,y\}}\frac{3}{11}(x^2+y)dydx & \text{if } x\geq 0, y\geq 0\\ 0 & \text{otherwise}\end{cases}\\ &=\begin{cases}\frac{3}{11}\left[\frac{\min\{1,y\}\min\{8,x^3\}}{3}+\frac{\min\{1,y^2\}}{2}\min\{2,x\}\right] & \text{if } x\geq 0, y\geq 0\\ 0 & \text{otherwise,}\end{cases}\end{aligned} \tag{8}$$

which satisfies $F_{XY}(\infty,\infty)=\frac{3}{11}\left[\frac{8}{3}+\frac{1}{2}2\right]=1$.

Similarly, we can define the ***marginal*** cdf, which is given by:

$$\begin{aligned}F_i(x)\equiv P(X_i\leq x)&=P(X_1\leq\infty,...,X_i\leq x,...,X_K\leq\infty)\\ &=F_{X_1...X_K}(\infty,...,x,...,\infty),\end{aligned} \tag{9}$$

and, either the marginal pmf (discrete case), defined as:

$$P(X_i = x) \equiv \sum_{x_1} ... \sum_{x_K} P(X_1 = x_1, ...X_i = x, ..., X_K = x_K), \tag{10}$$

or the marginal pdf (continuous case), defined as:

$$f_i(x) = \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} f_{X_1...X_K}(z_1, ..., x, ..., z_K)dz_1...dx_{i-1}dx_{i+1}...dz_K. \tag{11}$$

Note that the marginal cdf can also be defined as:

$$F_i(x) = \int_{-\infty}^{x} f_i(z)dz. \tag{12}$$

In our discrete example from above, the marginal pmf for X_1 is:

$$P(X_1 = x) = \begin{cases} \frac{3}{4} & \text{if } x = 0 \\ \frac{1}{4} & \text{if } x = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

Note that this is still a well defined probability function for the variable X_1, as it satisfies the three axioms of a probability function.

In the continuous example above, the marginal pdf for X is:

$$f_X(x) = \begin{cases} \int_0^1 \frac{3}{11}(x^2 + y)dy = \frac{3}{11}\left(x^2 + \frac{1}{2}\right) & \text{if } 0 \leq x \leq 2 \\ 0 & \text{otherwise,} \end{cases} \tag{14}$$

which is a well defined pdf, as it integrates to 1, and the marginal cdf is:

$$F_X = \begin{cases} 1 & \text{if } x \geq 2 \\ \int_0^x \frac{3}{11}\left(x^2 + \frac{1}{2}\right) = \frac{3}{11}\left(\frac{x^3}{3} + \frac{x}{2}\right) & \text{if } 0 \leq x \leq 2 \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

II. Conditional Distributions and Independence

A. Conditional probability

Let us first introduce the concept of ***conditional probability***. In probability theory, a conditional probability measures the probability of an event given that another event has occurred. Let $\mathcal{A}$ and $\mathcal{B}$ be two events included in the σ-algebra of the sample space. The probability that $\mathcal{A}$ occurs given that $\mathcal{B}$ occurred, denoted by $P(\mathcal{A} \,|\, \mathcal{B})$ is formally defined as:

$$P(\mathcal{A} \,|\, \mathcal{B}) \equiv \frac{P(\mathcal{A} \cap \mathcal{B})}{P(\mathcal{B})}. \tag{16}$$

To illustrate it, consider the example of tossing two coins. We want to know that is the probability of obtaining two heads, conditional on the fact that the first coin already delivered a head. In this case, the sample space would be $\Omega = \{\{head, head\}, \{head, tail\}, \{tail, head\}, \{tail, tail\}\}$, the set $\mathcal{A}$ would be $\mathcal{A} = \{head, head\}$, the set $\mathcal{B}$ would be $\mathcal{B} = \{\{head, tail\}, \{head, head\}\}$, and the intersection of the two would be $\mathcal{A} \cap \mathcal{B} = \{head, head\}$. Thus, $P(\mathcal{A} \cap \mathcal{B})$, assuming coins are regular and, hence, events are equally likely, would be equal to $\frac{1}{4}$. Likewise, $P(\mathcal{B})$ would be $\frac{2}{4}$. Hence, $P(\mathcal{A} \,|\, \mathcal{B}) = \frac{1}{2}$.

This definition can be reversed to obtain the probability of $\mathcal{B}$ given that $\mathcal{A}$ occur, as they are both connected by $P(\mathcal{A} \cap \mathcal{B})$:

$$P(\mathcal{A} \cap \mathcal{B}) = P(\mathcal{A} \,|\, \mathcal{B})P(\mathcal{B}) = P(\mathcal{B} \,|\, \mathcal{A})P(\mathcal{A}) \Rightarrow P(\mathcal{B} \,|\, \mathcal{A}) = \frac{P(\mathcal{A} \,|\, \mathcal{B})P(\mathcal{B})}{P(\mathcal{A})}. \quad (17)$$

This identity is called the ***Bayes' rule*** (a.k.a. Bayes' law, or Bayes' theorem).

The conditional probability allows us to talk about the ***independence*** of two events. We say that events $\mathcal{A}$ and $\mathcal{B}$ are independent if the conditional and marginal probabilities coincide. That is:

- $P(\mathcal{A} \,|\, \mathcal{B}) = P(\mathcal{A})$
- $P(\mathcal{B} \,|\, \mathcal{A}) = P(\mathcal{B})$
- $P(\mathcal{A} \cap \mathcal{B}) = P(\mathcal{A})P(\mathcal{B})$.

Notice that these three conditions are equivalent, so we only need to check whether one of them holds.

B. Conditional distribution

Let X be a random variable, and let $\mathcal{A}$ be an event, with $P(\mathcal{A}) \neq 0$. The ***conditional*** cdf of X given $\mathcal{A}$ occurred is:

$$F_{X|\mathcal{A}}(x) \equiv P(X \leq x|\,\mathcal{A}) = \frac{P(X \leq x \cap \mathcal{A})}{P(\mathcal{A})}. \quad (18)$$

Very often, the event we are conditioning on is represented by a random variable(s), so that both X (which itself could also be a scalar or a random vector) and this random variable(s) form a random vector. In general, let X_1 denote the partition of the random vector that is our outcome of interest, and X_2 be the partition that includes the random variables we are conditioning on. The cdf of X_1 conditional on $X_2 = x_2$ is defined as:

$$F_{X_1|X_2}(x|x_2) \equiv \begin{cases} P(X_1 \leq x|X_2 = x_2) & \text{if } X_2 \text{ is discrete} \\ \lim_{h \to 0} P(X_1 \leq x|x_2 + h \geq X_2 \geq x_2) & \text{if } X_2 \text{ is continuous.} \end{cases} \quad (19)$$

The distinction between continuous and discrete is because we require that the marginal probability of the condition is not equal to zero for it to be defined. In the discrete case, the pmf is:

$$P(X_1 = x|X_2 = x_2) = \frac{P(X_1 = x, X_2 = x_2)}{P(X_2 = x_2)}. \tag{20}$$

Similarly, we can develop an analogous definition for the case of a continuous random vector. The conditional pdf of X_1 conditional on X_2 is:

$$f_{X_1|X_2}(x|x_2) \equiv \frac{f_{X_1X_2}(x, x_2)}{f_{X_2}(x_2)}, \tag{21}$$

where $f_{X_1|X_2}$ denotes the conditional pdf, $f_{X_1X_2}$ is the joint pdf, and f_{X_2} is the marginal pdf for X_2. Note that we can use this expression to ***factorize*** the joint pdf, in a Bayes' rule fashion, as follows:

$$f_{X_1X_2}(x_1, x_2) = f_{X_1|X_2}(x_1|x_2)f_{X_2}(x_2) = f_{X_2|X_1}(x_2|x_1)f_{X_1}(x_1). \tag{22}$$

We use these factorizations very often in econometrics.

We can also use Equation (20) to reformulate the conditional cdf for a continuous random vector as:

$$F_{X_1|X_2}(x|x_2) = \int_{-\infty}^{x} f_{X_1|X_2}(z|x_2)dz \tag{23}$$

Note that the conditional pdf is a well defined pdf:

- $f_{X_1|X_2}(x_1|x_2) \geq 0$.
- $\int_{-\infty}^{\infty} f_{X_1|X_2}(x, x_2)dx = 1$.
- $\frac{\partial}{\partial x}F_{X_1|X_2}(x|x_2) = f_{X_1|X_2}(x|x_2)$.

Also note that, if X_1 is a random vector of size K_1, the above integrals and differentials are K_1-variate.

To illustrate all this, consider the example used in previous section (Equation (7)). The conditional pdf of Y given X is:

$$f_{Y|X}(y|x) = \frac{f_{XY}(x, y)}{f_X(x)} = \frac{\frac{3}{11}(x^2 + y)}{\frac{3}{11}\left(x^2 + \frac{1}{2}\right)} = \frac{x^2 + y}{x^2 + \frac{1}{2}}, \tag{24}$$

for the relevant interval, and zero otherwise. Now we can use this expression to easily compute, for example:

$$P\left(0 \leq y \leq 1/2|x = 1\right) = \int_0^{1/2} \frac{1 + y}{1 + \frac{1}{2}}dy = \frac{2}{3}\left(y + \frac{y^2}{2}\right)\Bigg]_0^{1/2} = \frac{5}{12}, \tag{25}$$

where we are abusing of notation by stating $x = 1$ instead of the limit $\lim_{h\to 0} x \in [1, 1+h]$. Similarly, we can compute:

$$P\left(0 \leq y \leq \frac{1}{2}\middle|\frac{1}{2} \leq x \leq \frac{3}{2}\right) = \frac{\displaystyle\int_{\frac{1}{2}}^{\frac{3}{2}} \int_{0}^{\frac{1}{2}} \frac{3}{11}\left(x^2 + y\right) dydx}{\displaystyle\int_{\frac{1}{2}}^{\frac{3}{2}} \frac{3}{11}\left(x^2 + \frac{1}{2}\right) dx}. \tag{26}$$

C. *Independence*

We say that two random variables X_1 and X_2 are ***independent*** if and only if:

- The conditional distributions $f(X_1|X_2)$ and $f(X_2|X_1)$ do not depend on the conditioning variable, X_2 and X_1 respectively.
- $F_{X_1X_2}(x_1, x_2) = F_{X_1}(x_1)F_{X_2}(x_2)$ for all X_1 and X_2.
- $P(x_1 \in \mathcal{A}_{X_1} \cap x_2 \in \mathcal{A}_{X_2}) = P(x_1 \in \mathcal{A}_{X_1})P(x_2 \in \mathcal{A}_{X_2})$.

The three conditions are equivalent. Note, for example, that the third condition implies that we can formulate the conditional probability as:

$$\begin{aligned} P(x_1 \in \mathcal{A}_{X_1}|x_2 \in \mathcal{A}_{X_2}) &= \frac{P(x_1 \in \mathcal{A}_{X_1} \cap x_2 \in \mathcal{A}_{X_2})}{P(x_2 \in \mathcal{A}_{X_2})} \\ &= \frac{P(x_1 \in \mathcal{A}_{X_1})P(x_2 \in \mathcal{A}_{X_2})}{P(x_2 \in \mathcal{A}_{X_2})} \\ &= P(x_1 \in \mathcal{A}_{X_1}), \end{aligned} \tag{27}$$

which does not depend on X_2 (as the first condition indicates. Likewise, the second condition implies:

$$f_{X_1X_2}(x_1, x_2) = \frac{\partial^2 F_{X_1X_2}(x_1, x_2)}{\partial X_1 \partial X_2} = \frac{\partial F_{X_1}(x_1)}{\partial X_1}\frac{\partial F_{X_2}(x_2)}{\partial X_2} = f_{X_1}(x_1)f_{X_2}(x_2). \tag{28}$$

Thus, similarly to what we obtained in Equation (27), $f_{X_1|X_2}(x_1|x_2) = f_{X_1}(x_1)$ for any x_1 and x_2. As a corollary, we can state that $(X_1, ..., X_K)$ are independent if and only if $F_{X_1 \dots X_K}(x_1, ..., x_K) = \prod_{i=1}^{K} F_i(x_i)$.

III. Transformations of Random Variables

Let $(X_1, ..., X_K)'$ be a size K vector of independent random variables, and $g_1(\cdot), ..., g_K(\cdot)$ be K functions such that $\{Y_i = g_i(X_i) : j = 1, ..., K\}$ are random variables, then $(Y_1, ..., Y_K)'$ is also a vector of independent random variables.

To see it, note that the cdf of $(Y_1, ..., Y_K)'$ is:

$$\begin{aligned} F_{Y_1 \ldots Y_K}(y_1, ..., y_K) &= P(Y_1 \leq y_1, ..., Y_K \leq y_K) \\ &= P(X_1 \leq g_1^{-1}(y_1), ..., X_K \leq g_K^{-1}(y_K)) \\ &= F_{X_1 \ldots X_K}(g_1^{-1}(y_1), ..., g_K^{-1}(y_K)) \\ &= \prod_{i=1}^{K} F_{X_i}(g_1^{-1}(y_i)), \end{aligned} \tag{29}$$

where the last equality results from the fact that $X_1, ..., X_K$ are independent, and that $g_i(X_i)$ only takes X_i as an argument, and not X_j for $j \neq i$.

Finally, let X be a size K vector of continuous random variables with pdf $f_X(x)$, and let K-dimensional function $Y = g(X)$ with a unique inverse $X = g^{-1}(Y)$, and:

$$\det\left(\frac{\partial g^{-1}(Y)}{\partial Y'}\right) \neq 0. \tag{30}$$

Then, the joint pdf of $Y = g(X)$ is:

$$f_Y(y) = f_X(g^{-1}(y)) \left| \det\left(\frac{\partial g^{-1}(Y)}{\partial Y'}\right) \right|. \tag{31}$$

IV. Multivariate Normal Distribution

The ***multivariate normal distribution*** is defined over a random vector $X = (X_1, ..., X_K)'$ by the following pdf:

$$f_X(x) = (2\pi)^{-\frac{K}{2}} \det(\Sigma_X)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x - \mu_X)'\Sigma_X^{-1}(x - \mu_X)\right), \tag{32}$$

where Σ_X is a $K \times K$ positive definite and symmetric matrix of parameters, and μ_X is a size $K \times 1$ vector of parameters. The fact that a random vector follows a multivariate normal is expressed as $X \sim \mathcal{N}_K(\mu_X, \Sigma_X)$. Thus, the normal distribution is completely characterized by $K + (\frac{1}{2}K(K+1))$ (the second term comes from the fact that Σ_X is symmetric).

The multivariate normal distribution is obtained as a linear transformation of a vector of independent random variables that are (standard) normally distributed. Formally, let $Z = (Z_1, ..., Z_K)'$ be a vector of independent variables such that $\{Z_i \sim \mathcal{N}(0, 1) : j = 1, ..., K\}$. Define the random vector X as $X \equiv \mu_X + \Sigma_X^{\frac{1}{2}} Z$, where μ_X is a size K vector, and $\Sigma_X^{\frac{1}{2}}$ is a $K \times K$ nonsingular matrix that satisfies $\left(\Sigma_X^{\frac{1}{2}}\right)\left(\Sigma_X^{\frac{1}{2}}\right)' = \Sigma_X$. Then, implementing the result in Equation (31), we can see that $X \sim \mathcal{N}(\mu_X, \Sigma_X)$:

$$\left.\begin{array}{l} \phi_k(z) = \prod_{i=1}^{K} \phi(z_i) = (2\pi)^{-\frac{K}{2}} \exp\left(-\frac{1}{2} z' z\right) \\ X = \mu_X + \Sigma_X^{\frac{1}{2}} Z \Leftrightarrow Z = \Sigma_X^{-\frac{1}{2}} (x - \mu_X) \\ \left| \det\left(\Sigma_X^{-\frac{1}{2}}\right) \right| = \det(\Sigma_X)^{-\frac{1}{2}} \end{array}\right\} \Rightarrow \begin{array}{l} f_X(x) = (2\pi)^{-\frac{K}{2}} \det(\Sigma_X)^{-\frac{1}{2}} \times \\ \quad \exp\left(-\frac{1}{2}(x - \mu_X)' \Sigma_X^{-1} (x - \mu_X)\right). \end{array} \tag{33}$$

Using a similar derivation we can prove that $Y = a + BX \sim \mathcal{N}(a + B\mu_X, B\Sigma_X B')$.

V. Covariance, Correlation, and Conditional Expectation

A. Covariance and Correlation between Two Random Variables

Let $(X_1, X_2)'$ be two random variables with expectations $\mu_{X_1} \equiv \mathbb{E}[X_1]$ and $\mu_{X_2} \equiv \mathbb{E}[X_2]$. The ***covariance*** between X_1 and X_2 is defined as:

$$\operatorname{Cov}(X_1, X_2) \equiv \mathbb{E}[(X_1 - \mu_{X_1})(X_2 - \mu_{X_2})]. \tag{34}$$

Note that the variance is a special case of covariance: the one of a variable X with itself. Some properties of the covariance are:

- $\operatorname{Cov}(X_1, X_2) = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1]\,\mathbb{E}[X_2]$.
- $\operatorname{Cov}(X_1, X_2) = \operatorname{Cov}(X_2, X_1)$.
- $\operatorname{Cov}(X, X) = \operatorname{Var}(X)$.
- $\operatorname{Cov}(c, X) = 0$.
- $\operatorname{Cov}(aX_1, bX_2) = ab \operatorname{Cov}(X_1, X_2)$.
- $\operatorname{Cov}(X_1 + X_2, X_3) = \operatorname{Cov}(X_1, X_3) + \operatorname{Cov}(X_2, X_3)$.
- $\operatorname{Var}(X_1 + X_2) = \operatorname{Var}(X_1) + \operatorname{Var}(X_2) + 2\operatorname{Cov}(X_1, X_2)$.

The magnitude of the covariance depends on the units of measure, the same way than the descriptive statistic counterpart seen in Chapter 1 did. That is why we define the ***correlation coefficient*** as:

$$\rho_{X_1 X_2} \equiv \frac{\operatorname{Cov}(X_1, X_2)}{\sqrt{\operatorname{Var}(X_1)\operatorname{Var}(X_2)}}. \tag{35}$$

The ***Cauchy-Schwarz inequality*** implies that $\rho_{X_1 X_2}^2 \leq 1$, or, in other words, that coefficient ranges between -1 and 1. To prove it, define $U \equiv X_1 - \mu_{X_1}$ and $V \equiv X_2 - \mu_{X_2}$, and the function $w(t) = \mathbb{E}[(U - tV)^2] \geq 0$. Because $w(t)$ is a quadratic function in t, $(\mathbb{E}[V^2])t^2 - (2\,\mathbb{E}[UV])t + \mathbb{E}[U^2]$, it either has no roots or one root. This implies the discriminant (for $at^2 + bt + c$, the discriminant is $b^2 - 4ac$) has to be either zero (one root for $w(t)$) or negative (no roots).

Thus, $(2\,\mathbb{E}[UV])^2 - 4\,\mathbb{E}[V^2]\,\mathbb{E}[U^2] \leq 0$, which implies $\mathbb{E}[UV]^2 \leq \mathbb{E}[V^2]\,\mathbb{E}[U^2]$, from which the result follows trivially. In the case when $\rho^2_{X_1X_2} = 1$, there exists a value for t^* such that $w(t^*) = 0$, which is equivalent to say that $U = t^*V$, i.e. $X_1 - \mu_{X_1} = t^*(X_2 - \mu_{X_2})$, or, in words, X_1 is a linear transformation of X_2.

B. Expectation and Variance of Random Vectors

Let $X = (X_1, ..., X_K)'$ be a size K vector of random variables. The ***expectation*** of the random vector X is defined as:

$$\mathbb{E}[X] \equiv \int_{-\infty}^{\infty} x dF_{X_1 \dots X_K}(x_1, ..., x_K) = \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_K] \end{pmatrix}, \tag{36}$$

where we make use of the Rienman-Stiljes integral. Similarly, we can define the expectation of a transformation of X as:

$$\mathbb{E}[g(X)] \equiv \int_{-\infty}^{\infty} g(x) dF_{X_1 \dots X_K}(x_1, ..., x_K). \tag{37}$$

A corollary of this is that, since $g(X) = (c_1, ..., c_K)X = c_1X_1 + ... + c_KX_K$, we can see that $\mathbb{E}[c_1X_1 + ... + c_KX_K] = c_1\,\mathbb{E}[X_1] + ... + c_K\,\mathbb{E}[X_K]$. Moreover, even though we cannot derive a general result for the expectation of the product of random variables, in the special case where the random variables are independent, we can establish that:

$$\begin{aligned} \mathbb{E}[X_1X_2] &= \int_{-\infty}^{\infty} x_1x_2 dF_{X_1X_2}(x_1, x_2) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2 f_{X_1X_2}(x_1, x_2) dx_1 dx_2 \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_1x_2 f_{X_1}(x_1) f_{X_2}(x_2) dx_1 dx_2 \\ &= \int_{-\infty}^{\infty} x_1 f_{X_1}(x_1) dx_1 \int_{-\infty}^{\infty} x_2 f_{X_2}(x_2) dx_2 \\ &= \mathbb{E}(X_1)\,\mathbb{E}(X_2). \end{aligned} \tag{38}$$

Thus, note that the fact that two variables are independent imply that the expectation of the product is the product of the expectations. However, the reverse implication is not true.

The ***variance-covariance matrix*** is defined as:

$$\begin{aligned}\operatorname{Var}(X) \equiv \mathbb{E}[(X-\mu_X)(X-\mu_X)'] = & \qquad (39)\\ = \begin{pmatrix} \operatorname{Var}(X_1) & \operatorname{Cov}(X_1,X_2) & \dots & \operatorname{Cov}(X_1,X_K) \\ \operatorname{Cov}(X_2,X_1) & \operatorname{Var}(X_2) & \dots & \operatorname{Cov}(X_2,X_K) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}(X_K,X_1) & \operatorname{Cov}(X_K,X_2) & \dots & \operatorname{Var}(X_K) \end{pmatrix},\end{aligned}$$

where $\mu_X \equiv \mathbb{E}[X]$. This matrix is symmetric (because $\operatorname{Cov}(X_i, X_j) = \operatorname{Cov}(X_j, X_i)$), and positive-semidefinite (which is the equivalent to say that the variance is non-negative). To prove the last, recall that a matrix M is positive-semidefinite if, for all non-zero vectors $c \in \mathbb{R}^K$, $c'Mc \geq 0$. In the case of the variance-covariance matrix:

$$c' \operatorname{Var}(X)c = c' \mathbb{E}[(X-\mu_X)(X-\mu_X)']c = \mathbb{E}[c'(X-\mu_X)(X-\mu_X)'c] = \mathbb{E}[Y^2] \geq 0, \qquad (40)$$

where we make use of the fact that the linear combination $c'(X - \mu_X)$ delivers a scalar random variable.

Retaking the example of the multivariate normal distribution, $\mathbb{E}[X] = \mu_X$, and $\operatorname{Var}(X) = \Sigma_X$. We can write Σ_X as:

$$\Sigma_X = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \dots & \rho_{1K}\sigma_1\sigma_K \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \dots & \rho_{2K}\sigma_2\sigma_K \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1K}\sigma_1\sigma_K & \rho_{2K}\sigma_2\sigma_K & \dots & \sigma_K^2 \end{pmatrix}. \qquad (41)$$

It can be proved that $\det \Sigma > 0 \Leftrightarrow -1 < \rho < 1$. Thus, if two variables are perfectly correlated, we cannot write its joint normal density (which makes sense, given that one variable is a linear transformation of the other).

C. Conditional Expectation

The ***conditional expectation*** of a continuous random variable X_1 given X_2 is defined as:

$$\mathbb{E}[X_1|X_2 = x_2] \equiv \int_{-\infty}^{\infty} x_1 f_{X_1|X_2}(x_1|x_2)dx_1. \qquad (42)$$

If X_1 is discrete, it is analogously defined as:

$$\mathbb{E}[X_1|X_2 = x_2] \equiv \sum_{x_1 \in (-\infty,\infty)} x_1 P(X_1 = x_1|X_2 = x_2). \qquad (43)$$

In general, using the Rienman-Stiltjes integral, we can write:

$$\mathbb{E}[X_1|X_2 = x_2] \equiv \int_{-\infty}^{\infty} x_1 dF_{X_1|X_2}(x_1|x_2). \tag{44}$$

The ***conditional variance*** is defined as:

$$\text{Var}[X_1|X_2 = x_2] \equiv \int_{-\infty}^{\infty} (x_1 - \mathbb{E}[X_1|X_2])^2 dF_{X_1|X_2}(x_1|x_2), \tag{45}$$

which can be expressed as $\mathbb{E}[X_1^2|X_2 = x_2] - \mathbb{E}[X_1|X_2 = x_2]^2$ (the derivation is analogous to the one for the unconditional variance).

Since we can compute the conditional expectation for every possible value that X_2 can take, we can simply denote $\mathbb{E}[X_1|X_2]$, which is a function of X_2. Let $h(X_2) \equiv \mathbb{E}[X_1|X_2]$ denote this function. We can compute $\mathbb{E}[h(X_2)]$ (i.e. $\mathbb{E}[\mathbb{E}[X_1|X_2]]$) integrating over the marginal distribution of X_2:

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[X_1|X_2]] &= \int_{-\infty}^{\infty} h(X_2) f_{X_2}(x_2) dx_2 && (46)\\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} x_1 f_{X_1|X_2}(x_1|x_2) dx_1 \right) f_{X_2}(x_2) dx_2 \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 f_{X_1|X_2}(x_1|x_2) f_{X_2}(x_2) dx_1 dx_2 \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x_1 f_{X_1X_2}(x_1, x_2) dx_1 dx_2 \\
&= E[X_1]. && (47)
\end{aligned}$$

This result is known as the ***law of iterated expectations***, and, even though we derived it for continuous variables, it also applies to discrete ones.

D. Revisiting Independence

In Section II.C above we defined the concept of independence, based on the distributions of the variables. Now, given the inputs in this section, we can revise the notion of independence, and define alternative degrees of independence. In particular, from strongest to weakest notion of independence (i.e. the first implies the second, which implies the third, but the reverse is not true), we define:

- ***Independence***: $F_{X_1|X_2}(x_1|x_2) = F_{X_1}(x_1)$ (or any of the equivalent specifications defined in Section II.C).
- ***Mean independence***: X_1 is mean independent of X_2 if $\mathbb{E}[X_1|X_2] = \mathbb{E}[X_1]$ for all values of X_2. Unlike the other two, this relation is not symmetric, as X_2 being mean independent of X_1 does not necessarily imply that X_1 is mean independent of X_2.

- ***Absence of correlation***: $\text{Cov}(X_1, X_2) = 0 = \rho_{X_1 X_2}$.

To illustrate all this, consider a simple example. Let X_1 and X_2 be two discrete random variables, with pmf defined by the following table:

$X_1 \backslash X_2$	1	2	3
2	1/12	1/6	1/12
3	1/6	0	1/6
4	0	1/3	0

In this example:

$$F(X_1 = x_1 | X_2 = 1) = \begin{cases} \frac{1/12}{1/12+1/6} = \frac{1}{3} & \text{if } x_1 = 2 \\ \frac{1/6}{1/12+1/6} = \frac{2}{3} & \text{if } x_1 = 3 \\ 0 & \text{if } x_1 = 4 \end{cases} \tag{48}$$

$$F(X_1 = x | X_2 = 2) = \begin{cases} \frac{1/6}{1/6+1/3} = \frac{1}{3} & \text{if } x_1 = 2 \\ 0 & \text{if } x_1 = 3 \\ \frac{1/3}{1/6+1/3} = \frac{2}{3} & \text{if } x_1 = 4 \end{cases}$$

$$F(X_1 = x | X_2 = 3) = \begin{cases} \frac{1}{3} & \text{if } x_1 = 2 \\ \frac{2}{3} & \text{if } x_1 = 3 \\ 0 & \text{if } x_1 = 4, \end{cases}$$

so clearly there is no independence, as $F(X_1|X_2)$ depends on X_2. Now, to check whether there is mean independence, we have to compute the conditional expectations:

$$\mathbb{E}[X_1 | X_2 = x_2] = \begin{cases} 2 * \frac{1}{3} + 3 * \frac{2}{3} = \frac{8}{3} & \text{if } x_2 = 1 \\ 2 * \frac{1}{3} + 4 * \frac{2}{3} = \frac{10}{3} & \text{if } x_2 = 2 \\ 2 * \frac{1}{3} + 3 * \frac{2}{3} = \frac{8}{3} & \text{if } x_2 = 3 \end{cases} \tag{49}$$

$$\mathbb{E}[X_2 | X_1 = x_1] = \begin{cases} 1 * \frac{1}{4} + 2 * \frac{1}{2} + 3 * \frac{1}{4} = 2 & \text{if } x_1 = 2 \\ 1 * \frac{1}{2} + 3 * \frac{1}{2} = 2 & \text{if } x_1 = 3 \\ 2 & \text{if } x_1 = 4, \end{cases}$$

so X_2 is mean independent of X_1, but X_1 is not mean independent of X_2. This implies that $\text{Cov}(X_1, X_2) = 0$. It is easy to show. Applying the law of iterated expectations, one can trivially see that $\text{Cov}(X_1, X_2) = \text{Cov}(X_1, \mathbb{E}[X_2|X_1]) = \text{Cov}(X_1, \mathbb{E}[X_2]) = 0$.

Another example is the multivariate normal distribution. In this case, independence, mean independence, and absence of correlation are equivalent, and all three occur (for the relation between X_i and X_j) if and only if $\rho_{ij} = 0$. To illustrate it, we can think of the bivariate normal, but it is trivially extended generally for the multivariate normal of any dimension. Independence is proved by checking that, when ρ is equal to zero, the joint density can be factorized as the product of the two marginals. Additionally, we can prove that the conditional distribution of X_1 given X_2 is:

$$X_1|X_2 \sim \mathcal{N}\left(\mu_1 + \rho_{12}\frac{\sigma_1}{\sigma_2}(X_2 - \mu_2), \sigma_1^2(1 - \rho_{12}^2)\right), \tag{50}$$

which implies mean independence if and only if $\rho_{12} = 0$ (provided that X_1 is not a degenerate random variable, and thus $\sigma_1 \neq 0$). Finally, we prove that absence of correlation is equivalent to $\rho_{12} = 0$ from the definition of ρ_{12}.

VI. Linear Prediction

A. Expectations and Prediction

The conditional expectation (and the expectation in general) can be written as the result of a minimization process. We already pointed out something similar for the sample mean in Chapter 1. The expectation is the ***optimal predictor*** in the sense that it minimizes the expected quadratic loss. More formally, let $h(X)$ denote a prediction of a variable Y based on the information in X, and define $U \equiv Y - h(X)$ the prediction error. The conditional expectation satisfies:

$$\mathbb{E}[Y|X] = \arg\min_{h(X)} \mathbb{E}[(Y - h(X))^2] = \arg\min_{h(X)} \mathbb{E}[U^2]. \tag{51}$$

This property is trivial to prove by checking that, for any other function $m(X)$, $\mathbb{E}[(Y - m(X))^2] = \mathbb{E}\left[\{(y - \mathbb{E}[y|X]) + (\mathbb{E}[Y|X] - m(x))\}^2\right] \geq \mathbb{E}[(y - \mathbb{E}[y|X])^2]$. By extension, $\mathbb{E}[Y]$ is the ***optimal constant predictor***.

This notion of prediction is interesting because it allows us to separate the variation in Y that can be explained by X from the one that cannot. A way of quantifying to what extent a variable Y is explained by X compared to other factors is through the variance. The variance of Y can be decomposed as:

$$\text{Var}(Y) = \text{Var}(\mathbb{E}[Y|X]) + \text{Var}(U) + 2\,\text{Cov}(\mathbb{E}[Y|X], U), \tag{52}$$

where we used the fact that $h(X) = \mathbb{E}[Y|X]$ is the optimal predictor, and, thus, $Y \equiv \mathbb{E}[Y|X] + U$. To compute the $\text{Var}(U)$, we need $\mathbb{E}(U^2)$ and $\mathbb{E}[U]^2$. For the second term, we can use the law of iterated expectations:

$$\mathbb{E}[U|X] = \mathbb{E}[Y - \mathbb{E}[Y|X]|X] = \mathbb{E}[Y|X] - \mathbb{E}[Y|X] = 0 \quad \Rightarrow \quad \mathbb{E}[U] = 0. \tag{53}$$

Thus:

$$\text{Var}(U) = \mathbb{E}[U^2] = \mathbb{E}[\mathbb{E}[U^2|X]] = \mathbb{E}[\mathbb{E}[(Y - \mathbb{E}[Y|X])^2|X]] = \mathbb{E}[\text{Var}(Y|X)]. \tag{54}$$

Finally, using the result in (53), the last term in Equation (52) is:

$$\mathrm{Cov}(\mathbb{E}[Y|X], U) = \mathbb{E}(\mathbb{E}[Y|X]U) = \mathbb{E}(\mathbb{E}[Y|X]\,\mathbb{E}[U|X]) = 0. \tag{55}$$

Hence, we can write:

$$\mathrm{Var}(Y) = \mathrm{Var}(\mathbb{E}[Y|X]) + \mathbb{E}[\mathrm{Var}(Y|X)]. \tag{56}$$

The first term of the right-hand-side of the above expression is the variation of Y that is explained by X. The second term gives the expected variation of Y for a given value of X. Hence, we can introduce the ***population*** R^2:

$$R^2 \equiv \frac{\mathrm{Var}(\mathbb{E}[Y|X])}{\mathrm{Var}(Y)} = 1 - \frac{\mathbb{E}[\mathrm{Var}(Y|X)]}{\mathrm{Var}(Y)}. \tag{57}$$

This coefficient ranges between 0 and 1, and its interpretation is the fraction of the variation in Y that is explained by the variation in the prediction of Y given X. Thus, it is a measure of the ***goodness of fit*** of the model.

B. Optimal Linear Predictor

Now we focus on the ***optimal linear predictor***. Given a random vector (Y, X), the optimal linear predictor of Y given X is the function $\mathbb{E}^*[Y|X] \equiv \alpha + \beta X$ that satisfies:

$$(\alpha, \beta) = \arg\min_{(a,b)} \mathbb{E}[(Y - a - bX)^2]. \tag{58}$$

Solving for the first order conditions:

$$\beta = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}, \qquad \alpha = \mathbb{E}[Y] - \beta\,\mathbb{E}[X], \tag{59}$$

and, hence, it is equal to:

$$\mathbb{E}^*[Y|X] = \mathbb{E}[Y] + \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}(X - \mathbb{E}[X]), \tag{60}$$

The optimal linear predictor is optimal (in the sense of minimal quadratic error) in the class of linear predictors. Thus, when the conditional expectation function is linear, the optimal linear predictor is equal to the conditional expectation.

Using the optimal linear predictor we can compute a goodness of fit statistic that is analogous to the population R^2, defined in Equation (57):

$$\rho^2_{XY} \equiv \frac{\mathrm{Var}(\mathbb{E}^*[Y|X])}{\mathrm{Var}(Y)} = \beta^2 \frac{\mathrm{Var}(X)}{\mathrm{Var}(Y)} = \frac{\mathrm{Cov}(X, Y)^2}{\mathrm{Var}(X)\,\mathrm{Var}(Y)}, \tag{61}$$

and, hence, the notation ρ^2_{XY}. Notice that $0 \leq \rho^2_{XY} \leq R^2$. Also note that, if X is itself a random vector, then:

$$\beta = \operatorname{Var}(X)^{-1} \operatorname{Cov}(X, Y), \qquad \alpha = \mathbb{E}[Y] - \beta' \, \mathbb{E}[X]. \tag{62}$$

Let us introduce some properties of the optimal linear predictor:

- $\mathbb{E}^*[c|X] = c$.
- $\mathbb{E}^*[cX|X] = cX$.
- $\mathbb{E}^*[Y + Z|X] = \mathbb{E}^*[Y|X] + \mathbb{E}^*[Z|X]$.
- $\mathbb{E}^*[Y|X_1] = \mathbb{E}^*[\mathbb{E}^*[Y|X_1, X_2]|X_1]$.

An interesting case in which the optimal predictor is linear (i.e. the conditional expectation function of Y given X is linear in X) is the multivariate normal distribution (e.g. see the bivariate case in Equation (50)).