Prof. Lorenzo Naranjo Fall 2025

Probability Basics

This document is a brief introduction to probability concepts that will be used throughout the course. For clarity, I present the main results using a discrete probability space. However, all results extend to the general case, where $(\Omega, \mathcal{F}, \mathrm{P})$ consists of an arbitrary sample space Ω, a σ-algebra $\mathcal{F}$ of subsets of Ω, and a probability measure P defined on $\mathcal{F}$.

For a more complete treatment, see introductory textbooks such as Ross (2019) and Blitzstein and Hwang (2019). For more advanced treatments, see Grimmett and Stirzaker (2001) and Durrett (2019). For a classic treatment, see Feller (1968). For a computational treatment with Python, see Unpingco (2019).

Sets

A **set** is a collection of objects. The objects of a set can be anything you want. For example, a set may contain numbers, letters, cars, or pictures. In our case, we will be concerned with sets that contain future possibilities or outcomes that can occur.

Sets are fundamental in probability theory because events are sets of outcomes. Once outcomes are organized as sets, operations such as unions, intersections, and complements translate directly into statements like “or”, “and”, and “not”, which allows us to define probabilities consistently and prove core probability rules.

One way to define a set is to enumerate its elements. For example, the set of all integers from 1 to 10 is

$$A = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}.$$

Once we have defined a set, we can answer if an object is an element of the set or not. For example, the number 3 is an element of A whereas the number 20 is not. We use the symbol $\in$ to denote membership of a set and $\notin$ to denote non-membership. Therefore, we have that $3 \in A$ and $20 \notin A$.

Some sets can have an infinite number of elements. For example, the natural numbers are defined as

$$\mathbb{N} = \{0, 1, 2, 3, \ldots\},$$

where the triple dots mean that if n is in $\mathbb{N}$, then $n + 1$ is also in $\mathbb{N}$.

Since all elements of A are also members of $\mathbb{N}$, we say that A is a subset of $\mathbb{N}$ and write it as $A \subset \mathbb{N}$. Using this terminology, we can redefine the set A defined above in a more *Pythonic* way:

$$A = \{n \in \mathbb{N} : n < 11\}.$$

If we are studying sets of natural numbers, it makes sense to define the **universe** to be $\mathbb{N}$ and sets under study will be subsets of the universe.

Now, define the set B as

$$B = \{6, 7, 8, 9, 10, 11, 12, 13, 14, 15\}.$$

The **intersection** between A and B is the set denoted $A \cap B$ whose members are both in A and B. Using the sets defined above, we have that

$$A \cap B = \{6, 7, 8, 9, 10\}.$$

The **union** of the sets A and B is the set denoted $A \cup B$ whose members are either in A, B, or both. Thus, using our previously defined sets we have that

$$A \cup B = \{1, 2, 3, \ldots, 14, 15\}.$$

The **set difference** of A and B is the set denoted $A \setminus B$ whose members are in A but are not members of B. Thus,

$$A \setminus B = \{1, 2, 3, 4, 5\}$$

and

$$B \setminus A = \{11, 12, 13, 14, 15\}.$$

The **complement** of A is the set denoted by A^C whose members are not in A. Of course this

statement only makes sense if we define a universe where the elements not in A can live. If the universe is $\mathbb{N}$, then

$$A^C = \mathbb{N} \setminus A = \{11, 12, 13, \ldots\}.$$

Similarly,

$$B^C = \{0, 1, 2, 3, 4, 5\} \cup \{16, 17, 18, \ldots\}.$$

Note that if you take all the elements of A out of A you end up with an empty set, that is $A \setminus A = \{\}$. We typically denote the empty set by $\emptyset$, but it is good to keep in mind that $\emptyset = \{\}$. In our universe of natural numbers, no natural number is a member of the empty set. We can write this formally as $n \notin \emptyset, \forall n \in \mathbb{N}$. Thus, the empty set is a subset of any subset of $\mathbb{N}$.

The **cardinality** of the set A, denoted by $|A|$, counts the number of elements in A. We then have that $|A| = |B| = 10$. The empty set has cardinality 0 whereas the cardinality of $\mathbb{N}$ is denoted $\aleph_0$.

The **power set** of a set C, denoted by $\mathcal{P}(C)$, is the set containing all possible subsets of C. For example, if $C = \{1, 2, 3\}$, then

$$\mathcal{P}(C) = \{\{\}, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}, \{1, 3\}, \{1, 2, 3\}\}.$$

Clearly, the power sets of A and B are much bigger. For a given set A, the cardinality of its power set is $2^{|A|}$. Therefore, $\mathcal{P}(A)$ and $\mathcal{P}(B)$ each contain $2^{10} = 1024$ subsets.

Finally, the **Cartesian product** of A and B is the set denoted by $A \times B$ whose members are all the pairwise combinations of the elements of A and B.

$A \times B$	6	7	…	15
1	$(1, 6)$	$(1, 7)$	…	$(1, 15)$
2	$(2, 6)$	$(2, 7)$	…	$(2, 15)$
⋮	⋮	⋮	⋱	⋮
10	$(10, 6)$	$(10, 7)$	…	$(10, 15)$

The cardinality of $A \times B$ is equal to the product of the cardinalities of A and B, i.e., $|A \times B| = |A| \times |B|$.

Outcomes and Events

In probability theory, a finite **sample space** is a non-empty finite set denoted by Ω. The sample space includes all possible outcomes that can occur. A *probability measure* is a function that assigns to each element ω of Ω a number in $[0, 1]$ so that

$$\sum_{\omega \in \Omega} \mathrm{P}(\omega) = 1.$$

An event A is a subset of Ω, and we define the probability of that event occurring as

$$\mathrm{P}(A) = \sum_{\omega \in A} \mathrm{P}(\omega). \tag{1}$$

We usually denote the set of all events by $\mathcal{F}$, which for convenience we will take to be the power set of Ω, i.e., $\mathcal{F} = \mathcal{P}(\Omega)$. Thus, we have that $\mathcal{F}$ contains all possible subsets of Ω and P is defined for all those subsets.

The expression $(\Omega, \mathcal{F}, \mathrm{P})$ then defines a finite probability space, where Ω is the sample space, $\mathcal{F}$ is the set of events, and P is the probability measure. Since $\mathcal{F} = \mathcal{P}(\Omega)$, we can simply write (Ω, P) to denote the same probability space.

An immediate consequence of (1) is that $\mathrm{P}(\Omega) = 1$. Furthermore, if A and B are disjoint sets of Ω we have that

$$\begin{aligned} \mathrm{P}(A \cup B) &= \sum_{\omega \in A \cup B} \mathrm{P}(\omega) \\ &= \sum_{\omega \in A} \mathrm{P}(\omega) + \sum_{\omega \in B} \mathrm{P}(\omega) \\ &= \mathrm{P}(A) + \mathrm{P}(B). \end{aligned}$$

If we denote by A^C the complement of A in Ω, the last expression implies that $\mathrm{P}(A) + \mathrm{P}(A^C) = 1$. Also, because $\Omega^C = \emptyset$, we also have that $\mathrm{P}(\Omega) + \mathrm{P}(\emptyset) = 1$, or $\mathrm{P}(\emptyset) = 0$.

Example 1. If $\Omega = \{\omega_1, \omega_2, \omega_3\}$, then

$$\mathcal{P}(\Omega) = \{\emptyset, \{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_1, \omega_2\}, \{\omega_2, \omega_3\}, \{\omega_1, \omega_3\}, \{\omega_1, \omega_2, \omega_3\}\}$$

defines the collection of all possible events that we can measure. As we saw previously, the cardinality of $\mathcal{P}(\Omega)$ grows exponentially with the size of Ω.

The function P such that $\mathrm{P}(\omega_1) = 1/2$, $\mathrm{P}(\omega_2) = 1/4$, and $\mathrm{P}(\omega_3) = 1/4$ defines a probability measure on Ω. For example, we have that $\mathrm{P}(\{\omega_1, \omega_3\}) = 1/2 + 1/4 = 3/4$. □

More generally, if $\{A_i : i \in I\}$ is a collection of pairwise disjoint subsets of Ω, then no outcome ω belongs to more than one A_i. In this case, the probability of their union is simply the sum of their probabilities:

$$\mathrm{P}\left(\bigcup_{i \in I} A_i\right) = \sum_{i \in I} \mathrm{P}(A_i).$$

This property is called countable additivity.

Random Variables

Definition

A **random variable** X is a function that assigns a real value to each outcome: $X(\omega)$ for $\omega \in \Omega$. Several outcomes may have the same value of X.

Example 2. Consider a sample space with four possible outcomes $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$. The table below describes the possible values of three random variables denoted by X, Y and Z.

Outcome	X	Y	Z
ω_1	-10	20	15
ω_2	-5	10	-10
ω_3	5	0	15
ω_4	10	0	-10

Observing the values of X provides perfect information about which event happened. For example, if $X = 5$ then we know that ω_3 occurred.

Knowing the values of Y or Z, on the other hand, does not provide the same amount of information. If we learn that $Y = 0$ we only know that either ω_3 or ω_4 occurred. If we denote by $\mathcal{F}_Y$ the set of events that can be generated by Y, we have that

$$\mathcal{F}_Y = \{\emptyset, \{\omega_1\}, \{\omega_2\}, \{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}, \{\omega_1, \omega_3, \omega_4\}, \{\omega_2, \omega_3, \omega_4\}, \Omega\}.$$

The information set provided by Z is even smaller, since

$$\mathcal{F}_Z = \{\emptyset, \{\omega_1, \omega_3\}, \{\omega_2, \omega_4\}, \Omega\}.$$

Thus, a random variable does not necessarily provide all the information generated by the probability space Ω. □

Expectation and Variance

If X is a random variable defined on a finite probability space (Ω, P), the **expectation** or **expected value** of X is defined to be

$$\mathsf{E}\, X = \sum_{\omega \in \Omega} X(\omega)\, \mathsf{P}(\omega),$$

whereas the **variance** of X is

$$\mathsf{V}(X) = \mathsf{E}(X - \mathsf{E}\, X)^2.$$

The **standard deviation** is the square-root of the variance, i.e., $\sigma_X = \sqrt{\mathsf{V}(X)}$.

Example 3. Consider the sample space $\Omega = \{\omega_1, \omega_2, \omega_3\}$ in which we define the probability measure P such that $\mathsf{P}(\omega_1) = 1/2$, $\mathsf{P}(\omega_2) = 1/4$, and $\mathsf{P}(\omega_3) = 1/4$. There are two random variables X and Y that take values in Ω according to the table below.

Outcome	Probability	X	Y
ω_1	1/2	10	2
ω_2	1/4	8	40
ω_3	1/4	4	20

Using this information, we can compute the expectation of each random variable.

$$\mathrm{E}\,X = \frac{1}{2} \times 10 + \frac{1}{4} \times 8 + \frac{1}{4} \times 4 = 8,$$
$$\mathrm{E}\,Y = \frac{1}{2} \times 2 + \frac{1}{4} \times 40 + \frac{1}{4} \times 20 = 16.$$

Having computed the expectations of X and Y, we can compute their variances as

$$\mathrm{V}(X) = \frac{1}{2} \times (10 - 8)^2 + \frac{1}{4} \times (8 - 8)^2 + \frac{1}{4} \times (4 - 8)^2 = 6,$$
$$\mathrm{V}(Y) = \frac{1}{2} \times (2 - 16)^2 + \frac{1}{4} \times (40 - 16)^2 + \frac{1}{4} \times (20 - 16)^2 = 246.$$

Finally, the standard deviations of X and Y are $\sigma_X = \sqrt{6} \approx 2.45$ and $\sigma_Y = \sqrt{246} \approx 15.68$, respectively. □

Covariance

The **covariance** between two random variables X and Y defined on a probability space (Ω, P) is defined as

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

and their **correlation** is

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

The correlation between any two random variables is always between -1 and 1.

Proof

Let σ_X and σ_Y denote the standard deviations of X and Y, respectively. We can then compute

$$\begin{aligned}\mathrm{E}((X - \mathrm{E}\,X)\sigma_Y + (Y - \mathrm{E}\,Y)\sigma_X)^2 &= (\sigma_X^2\sigma_Y^2 + 2\sigma_X\sigma_Y\,\mathrm{Cov}(X, Y) + \sigma_Y^2\sigma_X^2) \\ &= 2\sigma_X\sigma_Y(\sigma_X\sigma_Y + \mathrm{Cov}(X, Y)),\end{aligned}$$

which implies $\sigma_X\sigma_Y + \mathrm{Cov}(X, Y) \geq 0$ or $-\sigma_X\sigma_Y \leq \mathrm{Cov}(X, Y)$.

Similarly,

$$\begin{aligned}\mathrm{E}((X-\mathrm{E}\,X)\sigma_Y-(Y-\mathrm{E}\,Y)\sigma_X)^2 &= (\sigma_X^2\sigma_Y^2-2\sigma_X\sigma_Y\,\mathrm{Cov}(X,Y)+\sigma_Y^2\sigma_X^2)\\ &= 2\sigma_X\sigma_Y(\sigma_X\sigma_Y-\mathrm{Cov}(X,Y)),\end{aligned}$$

which implies $\sigma_X\sigma_Y-\mathrm{Cov}(X,Y)\geq 0$ or $\mathrm{Cov}(X,Y)\leq\sigma_X\sigma_Y$.

Thus, we conclude that

$$-1\leq\frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}\leq 1,$$

or equivalently $-1\leq\rho_{X,Y}\leq 1$. □

Example 4. Continuing with Example 3, we have that

$$\mathrm{Cov}(X,Y)=\frac{1}{2}\times(10-8)(2-16)+\frac{1}{4}(8-8)(40-16)+\frac{1}{4}(4-8)(20-16)=-18.$$

Thus, $\rho_{X,Y}\approx-0.47$. □

The covariance of X and Y can also be expressed as

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

Proof

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathrm{E}(X-\mathrm{E}(X))(Y-\mathrm{E}(Y))\\ &= \mathrm{E}[X(Y-\mathrm{E}(Y))]-\mathrm{E}[\mathrm{E}(X)(Y-\mathrm{E}(Y))]\\ &= \mathrm{E}(XY)-\mathrm{E}[X\,\mathrm{E}(Y)]-\mathrm{E}(X)\,\mathrm{E}(Y-\mathrm{E}(Y))\\ &= \mathrm{E}(XY)-\mathrm{E}(X)\,\mathrm{E}(Y).\end{aligned}$$

□

Probability Mass Function

For discrete random variables, the **probability mass function** (or pmf) is a real-valued function that specifies the probability that the random variable X is equal to a certain value x, i.e.,

$$p_X(x) = \mathrm{P}(\omega \in \Omega : X(\omega) = x).$$

Example 5. Suppose we define a probability measure P to the random variables X and Y defined in Example 2 according to the table below.

Outcome	P	X	Y
ω_1	0.10	-10	20
ω_2	0.30	-5	10
ω_3	0.40	5	0
ω_4	0.20	10	0

We have that the probability mass function of X is

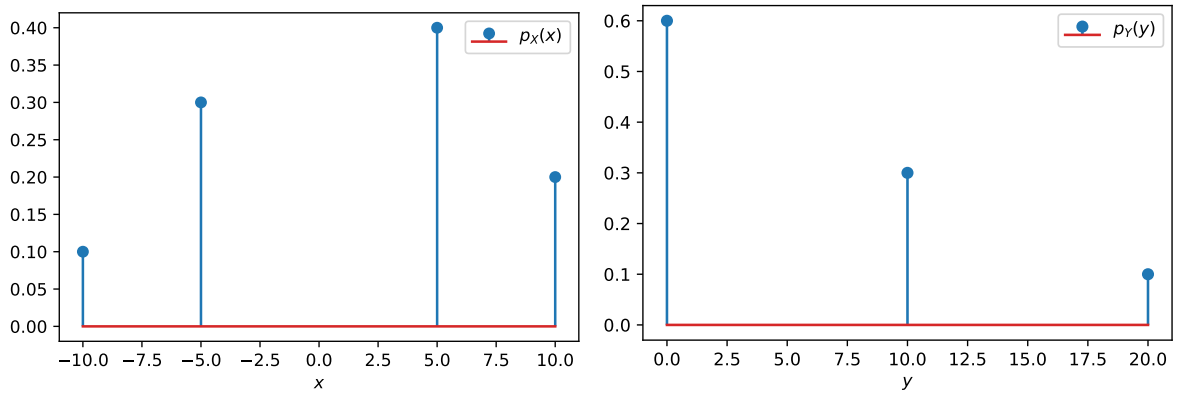
$$p_X(x) = \begin{cases} 0.10 & \text{if } x = -10, \\ 0.30 & \text{if } x = -5, \\ 0.40 & \text{if } x = 5, \\ 0.20 & \text{if } x = 10. \end{cases}$$

The probability mass function of Y takes positive values only at three points.

$$p_Y(y) = \begin{cases} 0.60 & \text{if } y = 0, \\ 0.30 & \text{if } y = 10, \\ 0.10 & \text{if } y = 20. \end{cases}$$

□

It is sometimes easier to visualize the probability mass function by plotting the probability of different values of the random variable.



(a) The function $p_X(x)$ defines the probability of X being equal to $x = \{-10, -5, 5, 10\}$.

(b) The function $p_Y(y)$ defines the probability of Y being equal to $y = \{0, 10, 20\}$.

Figure 1: The figure plots the probability mass function of the random variables X and Y.

It is apparent from the pictures that $p_X(x) = 0$ if $x \notin \{-10, -5, 5, 10\}$. Indeed, the set $\{\omega \in \Omega : X(\omega) = x\}$ is empty for all x not equal to -10, -5, 5, or 10. Similarly, $p_Y(y) = 0$ if $y \notin \{0, 10, 20\}$.

To simplify notation, we will often write $\{X = x\}$ to denote the set $\{\omega \in \Omega : X(\omega) = x\}$. Using this notation, we have that $p_X(x) = \mathrm{P}(X = x)$.

The **support** of X is the set

$$R_X = \{x \in \mathbb{R} : p_X(x) > 0\},$$

which is countable because Ω is countable. Similarly, the support of Y is $R_Y = \{y \in \mathbb{R} : p_Y(y) > 0\}$. Using this notation we can rewrite the expectation of X as

$$\mathrm{E}(X) = \sum_{x \in R_X} x\, p_X(x). \tag{2}$$

which is commonly used in statistics. The variance of X then becomes

$$\mathrm{V}(X) = \sum_{x \in R_X} (x - \mathrm{E}(X))^2\, p_X(x).$$

Note that we have

$$\begin{aligned}
\mathrm{V}(X) &= \sum_{x \in R_X} (x - \mathrm{E}(X))^2 \, p_X(x) \\
&= \sum_{x \in R_X} (x^2 - 2x \, \mathrm{E}(X) + \mathrm{E}(X)^2) \, p_X(x) \\
&= \sum_{x \in R_X} x^2 \, p_X(x) - 2 \, \mathrm{E}(X) \sum_{x \in R_X} x \, p_X(x) + \mathrm{E}(X)^2 \\
&= \mathrm{E}(X^2) - 2 \, \mathrm{E}(X)^2 + \mathrm{E}(X)^2 \\
&= \mathrm{E}(X^2) - \mathrm{E}(X)^2.
\end{aligned}$$

For two random variables X and Y defined in (Ω, P), the set $\{X = x, Y = y\}$ denotes all outcomes in Ω that satisfy $\{X = x\}$ and $\{Y = y\}$. Therefore, we have that

$$\{X = x, Y = y\} = \{X = x\} \cap \{Y = y\}.$$

The function

$$p_{X,Y}(x, y) = \mathrm{P}(X = x, Y = y)$$

is called the joint probability mass function of X and Y.

Example 6. The joint pmf of the random variables defined in Example 5 is given in the table below.

$X \setminus Y$	0	10	20
-10	0	0	0.1
-5	0	0.3	0
5	0.4	0	0
10	0.2	0	0

The function $p_{X,Y}(x, y)$ has many zeros since in Example 5 there are only four outcomes. Any other outcome then has probability zero of occurring. □

Example 7. We can generate any joint pmf for two random variables as long as the sum of all probabilities is equal to one. The table below reports the joint probabilities of a random variable X taking values in $[-1, 0, 1]$ and a random variable Y taking values in $[0, 1, 2, 3]$.

$X \setminus Y$	0	1	2	3
-1	0.12500	0.09375	0.06250	0.03125
0	0.06250	0.12500	0.12500	0.06250
1	0.03125	0.06250	0.09375	0.12500

In this case the underlying probability space has at least $3 \times 4 = 12$ possible outcomes. The figure below plots the joint pmf of X and Y. □

To plot the joint pmf of two random variables we need a three-dimensional graph.

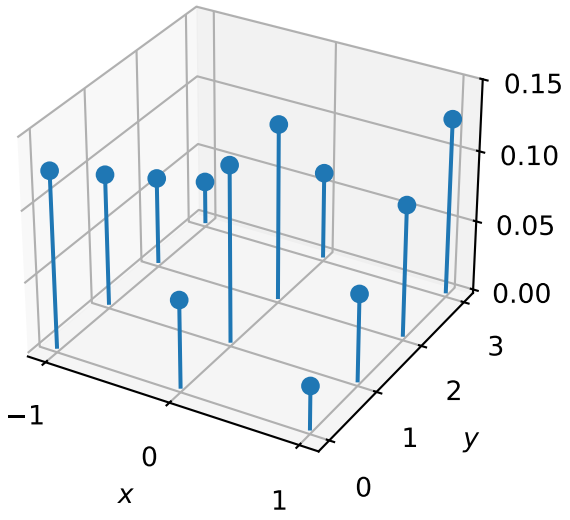


Figure 2: The figure plots the joint probability mass function of X and Y in Example 7.

We can use the joint pmf to compute the expectation of a function of two random variables. Indeed, we have that

$$\mathrm{E}(f(X,Y)) = \sum_{x \in R_X} \sum_{y \in R_Y} f(x,y)\, p_{X,Y}(x,y). \tag{3}$$

If we write $\mu_X = \mathrm{E}(X)$ and $\mu_Y = \mathrm{E}(Y)$, equation (3) implies that the covariance of X and Y can be computed as

$$\mathrm{Cov}(X, Y) = \sum_{x \in R_X} \sum_{y \in R_Y} (x - \mu_X)(y - \mu_Y) p_{X,Y}(x, y).$$

The joint pmf contains all the information of X and Y since we can recover the individual pmfs of X and Y from it. To find the probability that Y equals a specific value y, we sum over all possible values of X:

$$p_Y(y) = \sum_{x \in R_X} p_{X,Y}(x, y).$$

This works because the events $\{X = x, Y = y\}$ for different x are disjoint and together cover all ways Y can be y. Similarly,

$$p_X(x) = \sum_{y \in R_Y} p_{X,Y}(x, y).$$

Therefore, we can marginalize out Y to obtain the probability mass function of X, in the same way that we can marginalize out X to obtain the probability mass function of Y. It is important to note that the joint pmf not only contains the individual information of two random variables but also captures their mutual dependence.

Independence

We say that two events A and B are **independent** if $\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B)$.

Example 8. Suppose that the weather tomorrow can be either sunny, fair or rainy. In addition, a certain stock tomorrow can either go up or down in price.

We can define

$$W = \{\text{sunny, fair, rainy}\}$$

and

$$S = \{\text{up, down}\}.$$

The set of outcomes can be described as all possible pairwise combinations of weather tomorrow and the stock price movement. The sample space Ω is then the Cartesian product of W and S, i.e., $\Omega = W \times S$.

We can then define the weather events

$$\begin{aligned}\text{Sunny} &= \{(\text{sunny}, \text{up}), (\text{sunny}, \text{down})\},\\ \text{Fair} &= \{(\text{fair}, \text{up}), (\text{fair}, \text{down})\},\\ \text{Rainy} &= \{(\text{rainy}, \text{up}), (\text{rainy}, \text{down})\}.\end{aligned}$$

The table below describes the probabilities for tomorrow's weather.

Weather	Sunny	Fair	Rainy
Probability	0.3	0.5	0.2

Similarly, the stock events can be defined as

$$\begin{aligned}\text{Up} &= \{(\text{sunny}, \text{up}), (\text{fair}, \text{up}), (\text{rainy}, \text{up})\},\\ \text{Down} &= \{(\text{sunny}, \text{down}), (\text{fair}, \text{down}), (\text{rainy}, \text{down})\}.\end{aligned}$$

The probabilities of the stock price going up or down are described in the table below.

Stock	Up	Down
Probability	0.6	0.4

If the weather does not affect the likelihood of the stock going up or down, we should expect to see on sunny days 60% of the time the stock going up and 40% of those days the stock going down.

That is, if the weather tomorrow and the stock price movement are independent events, we should expect

$$\text{P}(\text{Stock} \cap \text{Weather}) = \text{P}(\text{Stock})\,\text{P}(\text{Weather}),$$

where Stock is either Up or Down, and Weather is either Sunny, Fair, or Rainy.

The table below describes the combined probabilities of the stock price movement and the weather tomorrow that are consistent with the independence of those events.

Stock\Weather	Sunny	Fair	Rainy
Up	0.18	0.30	0.12
Down	0.12	0.20	0.08

In the table, the weather does not change the relative proportions of the probabilities for the stock price. □

The previous example shows how to generate independent events out of two finite probability spaces (Ω_1, P_1) and (Ω_2, P_2). If we define $\Omega = \Omega_1 \times \Omega_2$ and let $\mathsf{P}(\omega_1, \omega_2) = \mathsf{P}_1(\omega_1)\,\mathsf{P}_2(\omega_2)$ for each $\omega_1 \in \Omega_1$ and $\omega_2 \in \Omega_2$, the pair (Ω, P) is a well-defined finite probability space. In this new probability space, the events $A = \{\omega_1\} \times \Omega_2$ and $B = \Omega_1 \times \{\omega_2\}$ are independent for any $\omega_1 \in \Omega_1$ and $\omega_2 \in \Omega_2$.

Proof

We have that

$$\begin{aligned}\mathsf{P}(A) &= \sum_{\omega_2 \in \Omega_2} \mathsf{P}(\omega_1, \omega_2) \\ &= \sum_{\omega_2 \in \Omega_2} \mathsf{P}_1(\omega_1)\,\mathsf{P}_2(\omega_2) \\ &= \mathsf{P}_1(\omega_1) \sum_{\omega_2 \in \Omega_2} \mathsf{P}_2(\omega_2) \\ &= \mathsf{P}_1(\omega_1).\end{aligned}$$

Similarly, $\mathsf{P}(B) = \mathsf{P}_2(\omega_2)$. Since $A \cap B = \{(\omega_1, \omega_2)\}$, we have that $\mathsf{P}(A \cap B) = \mathsf{P}(A)\,\mathsf{P}(B)$, proving that A and B are independent. □

Example 9. The sample space Ω is always independent from any event $A \subset \Omega$ since $\mathsf{P}(A \cap \Omega) = \mathsf{P}(A) = \mathsf{P}(A)\,\mathsf{P}(\Omega)$. Intuitively, conditioning on Ω does not change the probability of A. □

Two random variables X and Y are independent if the events $\{X = x\}$ and $\{Y = y\}$ are independent. Thus, if X and Y are independent we have that

$$\mathrm{P}(X = x, Y = y) = \mathrm{P}(X = x)\,\mathrm{P}(Y = y),$$

or equivalently

$$p_{X,Y}(x, y) = p_X(x)p_Y(y).$$

An important consequence of independence is that if X and Y are two independent random variables, then

$$\mathrm{E}(XY) = \mathrm{E}(X)\,\mathrm{E}(Y). \tag{4}$$

Proof

$$\begin{aligned}
\mathrm{E}(XY) &= \sum_{x \in R_X} \sum_{y \in R_Y} xy\, p_{X,Y}(x, y) \\
&= \sum_{x \in R_X} \sum_{y \in R_Y} xy\, p_X(x) p_Y(y) \\
&= \sum_{y \in R_Y} y\, p_Y(y) \sum_{x \in R_X} x\, p_X(x) \\
&= \sum_{y \in R_Y} y\, p_Y(y)\, \mathrm{E}(X) \\
&= \mathrm{E}(X) \sum_{y \in R_Y} y\, p_Y(y) \\
&= \mathrm{E}(X)\,\mathrm{E}(Y).
\end{aligned}$$

□

Equation (4) implies that if X and Y are independent, their covariance is equal to zero. Indeed,

$$\begin{aligned}
\mathrm{Cov}(X, Y) &= \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y) \\
&= \mathrm{E}(X)\,\mathrm{E}(Y) - \mathrm{E}(X)\,\mathrm{E}(Y) \\
&= 0.
\end{aligned}$$

However, the opposite statement is not true.

Example 10. Consider a random variable X that takes the values $\{1, 0, -1\}$, each with probability $1/3$. We compute:

$$\mathrm{E}(X) = \frac{1}{3}(1) + \frac{1}{3}(0) + \frac{1}{3}(-1) = 0,$$

and

$$\mathrm{E}(X^3) = \frac{1}{3}(1^3) + \frac{1}{3}(0^3) + \frac{1}{3}((-1)^3) = 0.$$

Now, define $Y = X^2$. The covariance between X and Y is

$$\mathrm{Cov}(X, Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y) = \mathrm{E}(X^3) - \mathrm{E}(X)\,\mathrm{E}(X^2) = 0.$$

This example shows that X and Y are uncorrelated (zero covariance), but they are not independent—Y is completely determined by X.

Linear Combinations

In portfolio theory, we usually study linear combinations of random variables of the form $Z = \alpha X + \beta Y$. The expectation of Z is just a linear combination of the expectations of X and Y,

$$\mathrm{E}\,Z = \alpha\,\mathrm{E}\,X + \beta\,\mathrm{E}\,Y. \tag{5}$$

The variance of Z, though, includes not only the variances of X and Y but also their covariances,

$$\mathrm{V}(Z) = \alpha^2\,\mathrm{V}(X) + \beta^2\,\mathrm{V}(Y) + 2\alpha\beta\,\mathrm{Cov}(X, Y). \tag{6}$$

This is an important result which is at the heart of **portfolio diversification**.

Proof

The expectation of Z is computed as

$$\begin{aligned}\mathrm{E}\, Z &= \mathrm{E}(\alpha X + \beta Y) \\ &= \sum_{\omega \in \Omega} (\alpha X(\omega) + \beta Y(\omega))\, \mathrm{P}(\omega) \\ &= \alpha \sum_{\omega \in \Omega} X(\omega)\, \mathrm{P}(\omega) + \beta \sum_{\omega \in \Omega} Y(\omega)\, \mathrm{P}(\omega) \\ &= \alpha\, \mathrm{E}\, X + \beta\, \mathrm{E}\, Y.\end{aligned}$$

The variance of Z is computed as

$$\begin{aligned}\mathrm{V}(Z) &= \mathrm{V}(\alpha X + \beta Y) \\ &= \mathrm{E}(\alpha X + \beta Y - \mathrm{E}(\alpha X + \beta Y))^2 \\ &= \mathrm{E}(\alpha(X - \mathrm{E}\, X) + \beta(Y - \mathrm{E}\, Y))^2 \\ &= \mathrm{E}(\alpha^2 (X - \mathrm{E}\, X)^2 + \beta^2 (Y - \mathrm{E}\, Y)^2 + 2\alpha\beta (X - \mathrm{E}\, X)(Y - \mathrm{E}\, Y)) \\ &= \alpha^2\, \mathrm{E}(X - \mathrm{E}\, X)^2 + \beta^2\, \mathrm{E}(Y - \mathrm{E}\, Y)^2 + 2\alpha\beta\, \mathrm{E}(X - \mathrm{E}\, X)(Y - \mathrm{E}\, Y) \\ &= \alpha^2\, \mathrm{V}(X) + \beta^2\, \mathrm{V}(Y) + 2\alpha\beta\, \mathrm{Cov}(X, Y).\end{aligned}$$

□

More generally, consider the random variables $X_1, X_2, \ldots, X_n$, and form a new random variable X such that

$$X = \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_n X_n,$$

where $\alpha_i \in \mathbb{R}$ for all $i \in \{1, 2, \ldots, n\}$.

The expectation of X is a linear combination of the expectations of $X_1, X_2, \ldots, X_n$. The variance of X, though, takes into account all covariances between X_i and X_j, for $i, j = 1, 2, \ldots, n$. Indeed, we have that

$$\mathrm{V}(X) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j\, \mathrm{Cov}(X_i, X_j). \tag{7}$$

The previous expression can be simplified if the random variables $X_1, X_2, \ldots, X_n$ are independent from each other. In such case, we have that $\mathrm{Cov}(X_i, X_j) = 0$ for all $i \neq j$. Recognizing that

$\mathrm{Cov}(X_i, X_i) = \mathrm{V}(X_i)$, equation (7) implies that

$$\mathrm{V}(X) = \sum_{i=1}^{n} \alpha_i^2 \, \mathrm{V}(X_i). \tag{8}$$

Example 11. Suppose that $X_1, X_2, \ldots, X_n$ are independent random variables with the same variance denoted by σ^2. Define X to be the sum of these random variables so that

$$X = X_1 + X_2 + \ldots + X_n.$$

Equation (8) implies that

$$\mathrm{V}(X) = \sum_{i=1}^{n} \mathrm{V}(X_i) = n\sigma^2.$$

□

Conditional Probability

For two events A and B with $\mathrm{P}(B) > 0$, the *conditional probability* of A given B is

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}.$$

Similarly, for x such that $p_X(x) > 0$, the conditional probability mass function of Y given $X = x$ is

$$p_{Y|X}(y \mid x) = \frac{p_{X,Y}(x, y)}{p_X(x)}.$$

Example 12. Suppose we roll a fair six-sided die and define the events

$$A = \{2, 4, 6\} \quad \text{and} \quad B = \{4, 5, 6\}.$$

Then

$$A \cap B = \{4, 6\}, \qquad \mathrm{P}(B) = \frac{3}{6}, \qquad \mathrm{P}(A \cap B) = \frac{2}{6}.$$

Therefore, the conditional probability of A given B is

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{2/6}{3/6} = \frac{2}{3}.$$

□

The *conditional expectation* of Y given $X = x$ is

$$\mathrm{E}(Y \mid X = x) = \sum_{y \in R_Y} y\, p_{Y|X}(y \mid x).$$

This is a function of x, and we can define the random variable $\mathrm{E}(Y \mid X)$ by assigning to each outcome ω the value $\mathrm{E}(Y \mid X = X(\omega))$.

A key result in probability theory is the *law of iterated expectations*:

$$\begin{aligned}
\mathrm{E}(\mathrm{E}(Y \mid X)) &= \sum_{x \in R_X} \mathrm{E}(Y \mid X = x)\, p_X(x) \\
&= \sum_{x \in R_X} \sum_{y \in R_Y} y\, p_{Y|X}(y \mid x)\, p_X(x) \\
&= \sum_{x \in R_X} \sum_{y \in R_Y} y\, p_{X,Y}(x, y) \\
&= \sum_{y \in R_Y} y\, p_Y(y) \\
&= \mathrm{E}(Y).
\end{aligned}$$

This means that the expected value of the conditional expectation equals the expected value of Y itself.

Example 13. Using the joint pmf in Example 7, we compute the conditional expectation of Y given $X = 1$. From the row $X = 1$, we have

$$p_{X,Y}(1,0) = 0.03125,\ p_{X,Y}(1,1) = 0.06250,\ p_{X,Y}(1,2) = 0.09375,\ p_{X,Y}(1,3) = 0.12500.$$

Hence

$$p_X(1) = 0.03125 + 0.06250 + 0.09375 + 0.12500 = 0.3125.$$

So the conditional pmf is

$$p_{Y|X}(y \mid 1) = \{0.1,\ 0.2,\ 0.3,\ 0.4\} \quad \text{for } y = \{0, 1, 2, 3\}.$$

Therefore,

$$\mathrm{E}(Y \mid X = 1) = \sum_{y \in R_Y} y\, p_{Y|X}(y \mid 1) = 0(0.1) + 1(0.2) + 2(0.3) + 3(0.4) = 2.$$

□

References

Blitzstein, Joseph K., and Jessica Hwang. 2019. *Introduction to Probability*. 2nd ed. Chapman; Hall/CRC.

Durrett, Rick. 2019. *Probability: Theory and Examples*. 5th ed. Cambridge University Press.

Feller, William. 1968. *An Introduction to Probability Theory and Its Applications. Volume 1*. 3rd ed. Wiley.

Grimmett, Geoffrey, and David Stirzaker. 2001. *Probability and Random Processes*. 3rd ed. Oxford University Press.

Ross, Sheldon M. 2019. *A First Course in Probability*. 11th ed. Pearson.

Unpingco, Jose C. 2019. *Python for Probability, Statistics, and Machine Learning*. 2nd ed. Springer.