

Chapter One

Introduction to Probability and Random Variables

1.1 INTRODUCTION TO RANDOM VARIABLES

1.1.1 Motivation

Probability theory is an attempt to formalize the notion of uncertainty in the outcome of an experiment. For instance, suppose an urn contains four balls, colored red, blue, white, and green respectively. Suppose we dip our hand in the urn and pull out one of the balls “at random.” What is the likelihood that the ball we pull out will be red? If we make multiple draws, replacing the drawn ball each time and shaking the urn thoroughly before the next draw, what is the likelihood that we have to make at least ten draws before we draw a red ball for the first time? Probability theory provides a mathematical abstraction and a framework where such issues can be addressed.

When there are only finitely many possible outcomes, probability theory becomes relatively simple. For instance, in the above example, when we draw a ball there are only four possible outcomes, namely: $\{R, B, W, G\}$ with the obvious notation. If we draw two balls, after replacing the first ball drawn, then there $4^2 = 16$ possible outcomes, represented as $\{RR, \dots, GG\}$. In such situations, one can get by with simple “counting” arguments. The counting approach can also be made to work when the set of possible outcomes is *countably* infinite.¹ This situation is studied in Section 1.3. However, in probability theory infinity is never very far away, and counting arguments can lead to serious logical inconsistencies if applied to situations where the set of possible outcomes is *uncountably* infinite. The great Russian mathematician A. N. Kolmogorov invented axiomatic probability theory in the 1930s precisely to address the issues thrown up by having uncountably many possible outcomes. Subsequent developments in probability theory have been based on the axiomatic foundation laid out in [81].

Example 1.1 Let us return to the example above. Suppose that all the four balls are identical in size and shape, and differ only in their color. Then it is reasonable to suppose that drawing any one color is as likely as drawing any other color, neither more nor less. This leads to the observation that the likelihood of drawing a red ball (or any other ball) is $1/4 = 0.25$.

Example 1.2 Now suppose that the four balls are all spherical, and that

¹Recall that a set S is said to be **countable** if it can be placed in one-to-one correspondence with the set of natural numbers $\mathbb{N} = \{1, 2, \dots\}$.

their diameters are in the ratio $4 : 3 : 2 : 1$ in the order red, blue, white, and green. We can suppose that the likelihood of our fingers touching and drawing a particular ball is proportional to its surface area. In this case, it follows that the likelihoods of drawing the four balls are in the proportion $4^2 : 3^2 : 2^2 : 1^2$ or $16 : 9 : 4 : 1$ in the order red, blue, white, and green. This leads to the conclusion that

$$P(R) = 16/30, P(B) = 9/30, P(W) = 4/30, P(G) = 1/30.$$

Example 1.3 There can be instances where such analytical reasoning can fail. Suppose that all balls have the same diameter, but the red ball is coated with an adhesive resin that makes it more likely to stick to our fingers when we touch it. The complicated interaction between the surface adhesion of our fingers and the surface of the ball may be too difficult to analyze, so we have no recourse other than to draw balls repeatedly and see how many times the red ball comes out. Suppose we make 1,000 draws, and the outcomes are: 451 red, 187 blue, 174 white, and 188 green. Then we can write

$$\hat{P}(R) = 0.451, \hat{P}(B) = 0.187, \hat{P}(W) = 0.174, \hat{P}(G) = 0.188.$$

The symbol $\hat{P}$ is used instead of P to highlight the fact that these are simply *observed frequencies*, and not the *true but unknown probabilities*. Often the observed frequency of an outcome is referred to as its **empirical probability**, or the empirical estimate of the true but unknown probability based on a particular set of experiments. It is tempting to treat the observed frequencies as true probabilities, but that would not be correct. The reason is that if the experiment is repeated, the outcomes would in general be quite different. The reader can convince himself/herself of the difference between frequencies and probabilities by tossing a coin ten times, and another ten times. It is extremely unlikely that the same set of results will turn up both times. One of the important questions addressed in this book is: Just how close are the observed *frequencies* to the true but unknown *probabilities*, and just how quickly do these observed frequencies converge to the true probabilities? Such questions are addressed in Section 1.3.3.

1.1.2 Definition of a Random Variable and Probability

Suppose we wish to study the behavior of a “random” variable X that can assume one of only a finite set of values belonging to a set $\mathbb{A} = \{a_1, \dots, a_n\}$. The set $\mathbb{A}$ of possible values is often referred to as the “alphabet” of the random variable. For example, in the ball-drawing experiment discussed in the preceding subsection, X can be thought of as the color of the ball drawn, and assumes values in the set $\{R, B, W, G\}$. This example, incidentally, serves to highlight the fact that the set of outcomes can consist of abstract *symbols*, and need not consist of *numbers*. This usage, adopted in this book, is at variance from the convention in many mathematics texts, where it is

assumed that $\mathbb{A}$ is a subset of the real numbers $\mathbb{R}$. However, since biological applications are a prime motivator for this book, it makes no sense to restrict $\mathbb{A}$ in this way. In genomics, for example, $\mathbb{A}$ consists of the four symbol set of nucleic acids, or nucleotides, usually denoted by $\{A, C, G, T\}$. Moreover, by allowing $\mathbb{A}$ to consist of arbitrary symbols, we also allow explicitly the possibility that *there is no natural ordering of these symbols*. For instance, in this book the nucleotides are written in the order A, C, G, T purely to follow the English alphabetical ordering. But there is no consensus on the ordering in biology texts. Thus any method of analysis that is developed here must be *permutation independent*. In other words, if we choose to order the symbols in the set $\mathbb{A}$ in some other fashion, the methods of analysis must give the same answers as before.

Now we give a general definition of the notion of probability, and introduce the notation that is used throughout the book.

Definition 1.1 *Given an integer n , the n -dimensional simplex $\mathbb{S}_n$ is defined as*

$$\mathbb{S}_n = \{\mathbf{v} \in \mathbb{R}^n : v_i \geq 0 \ \forall i, \sum_{i=1}^n v_i = 1\}. \quad (1.1)$$

Thus $\mathbb{S}_n$ consists of all nonnegative vectors whose components add up to one.

Definition 1.2 *Suppose $\mathbb{A} = \{a_1, \dots, a_n\}$ is a finite set. Then a **probability distribution** on the set $\mathbb{A}$ is any vector $\boldsymbol{\mu} \in \mathbb{S}_n$.*

The interpretation of a probability distribution $\boldsymbol{\mu}$ on the set $\mathbb{A}$ is that we say

$$\Pr\{X = a_i\} = \mu_i$$

to be read as “the probability that the random variable X equals x_i is μ_i .” Thus, if $\mathbb{A} = \{R, B, W, G\}$ and $\boldsymbol{\mu} = [0.25 \ 0.25 \ 0.25 \ 0.25]$, then all the four outcomes of drawing the various colored balls are equally likely. This is the case in Example 1.1. If the situation is as in Example 1.2, where the balls have different diameters in the proportion $4 : 3 : 2 : 1$, the probability distribution is

$$\boldsymbol{\mu} = [16/30 \ 9/30 \ 4/30 \ 1/30].$$

If we now choose to reorder the elements of the set $\mathbb{A}$ in the form $\{R, W, G, B\}$, then the probability distribution gets reordered correspondingly, as

$$\boldsymbol{\mu} = [16/30 \ 4/30 \ 1/30 \ 9/30].$$

Thus, when we speak of the probability distribution $\boldsymbol{\mu}$ on the set $\mathbb{A}$, we need to specify the ordering of the elements of the set.

The way we have defined it above, a probability distribution associates a *weight* with each element of the set $\mathbb{A}$ of possible outcomes. Thus $\boldsymbol{\mu}$ can be

thought of as a map from $\mathbb{A}$ into the interval $[0, 1]$. This notion of a weight of *individual elements* can be readily extended to define the weight of *each subset* of $\mathbb{A}$. This is called the probability **measure** P_μ associated with the distribution μ . Suppose $A \subseteq \mathbb{A}$. Then we define

$$P_\mu(A) := \Pr\{X \in A\} = \sum_{i=1}^n \mu_i I_A(x_i), \quad (1.2)$$

where $I_A(\cdot)$ is the so-called **indicator function** of the set A , defined by

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A, \\ 0 & \text{if } x \notin A. \end{cases} \quad (1.3)$$

So (1.2) states that the **probability measure** of the set A , denoted by $P_\mu(A)$, is the sum of the probability weights of the individual elements of the set A . Thus, whereas μ maps the set $\mathbb{A}$ into $[0, 1]$, the corresponding probability measure P_μ maps the “power set” $2^{\mathbb{A}}$ (that is, the collection of all subsets of $\mathbb{A}$) into the interval $[0, 1]$.

In this text, we need to deal with three kinds of objects:

- A probability distribution μ on a finite set $\mathbb{A}$.
- A random variable X assuming values in $\mathbb{A}$, with the probability distribution μ .
- A probability measure P_μ on the power set $2^{\mathbb{A}}$, associated with the probability distribution μ .

We will use whichever interpretation is most convenient and natural in the given context. As for notation, throughout the text, boldface Greek letters such as μ denote probability distributions. The probability measure corresponding to μ is denoted by P_μ . Strictly speaking, we should write P_μ , but for reasons of aesthetics and appearance we prefer to use P_μ . Similar notation applies to all other boldface Greek letters.

From (1.2), it follows readily that the empty set $\emptyset$ has probability measure zero, while the complete set $\mathbb{A}$ has probability measure one. This is true irrespective of what the underlying probability distribution μ is. Moreover, the following additional observations are easy consequences of (1.2):

Theorem 1.3 *Suppose $\mathbb{A}$ is a finite set and μ is a probability distribution on $\mathbb{A}$, and let P_μ denote the corresponding probability measure on $\mathbb{A}$. Then*

1. $0 \leq P_\mu(A) \leq 1 \ \forall A \subseteq \mathbb{A}$.
2. $P_\mu(\emptyset) = 0$ and $P_\mu(\mathbb{A}) = 1$.
3. If A, B are disjoint subsets of $\mathbb{A}$, then

$$P_\mu(A \cup B) = P_\mu(A) + P_\mu(B). \quad (1.4)$$

In the next paragraph we give a brief glimpse of axiomatic probability theory in a general setting, where the set $\mathbb{A}$ of possible outcomes is not necessarily finite. This paragraph is *not needed* to understand the remainder of the book, and therefore the reader can skip it with no aftereffects. In axiomatic probability theory, one actually *begins* with generalizations of the two properties above. One starts with a collection of subsets $\mathcal{S}$ of $\mathbb{A}$ that has three properties:

1. Both the empty set $\emptyset$ and $\mathbb{A}$ itself belong to $\mathcal{S}$.
2. If A belongs to $\mathcal{S}$, so does its complement A^c .
3. If $\{A_1, A_2, \dots\}$ is a *countable* collection of sets belonging to $\mathcal{S}$, then their union $B := \bigcup_{i=1}^{\infty} A_i$ also belongs to $\mathcal{S}$.

Such a collection $\mathcal{S}$ is called a **σ -algebra** of subsets of A , and the pair $(A, \mathcal{S})$ is called a **measurable space**. Note that, on the same set A , it is possible to define different σ -algebras. Given a measurable space $(A, \mathcal{S})$, a **probability measure** P is defined to be a function that maps the σ -algebra $\mathcal{S}$ into $[0, 1]$, or in other words a map that assigns a number $P(A) \in [0, 1]$ to each set A belonging to $\mathcal{S}$, such that two properties hold.

1. $P(\emptyset) = 0$ and $P(\mathbb{A}) = 1$.
2. If $\{A_1, A_2, \dots\}$ is a *countable* collection of *pairwise disjoint* sets belonging to $\mathcal{S}$, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

Starting with just these two simple sets of axioms, together with the notion of independence (introduced later), it is possible to build a tremendously rich edifice of probability theory; see [81].

In the case where the set $\mathbb{A}$ is either finite or countably infinite, by tradition one takes $\mathcal{S}$ to be the collection of *all* subsets of $\mathbb{A}$, because any other σ -algebra $\mathcal{S}'$ of subsets of $\mathcal{A}$ must in fact be a subalgebra of $\mathcal{S}$, in the sense that every set that is contained in the collection $\mathcal{S}'$ must also belong to $\mathcal{S}$. Now suppose P is a probability measure on $\mathcal{S}$. Then P assigns a weight $P(\{a_i\}) =: \mu_i$ to each element $a_i \in \mathbb{A}$. Moreover, if A is a subset of $\mathbb{A}$, then the measure $P(A)$ is just the sum of the weights assigned to individual elements of A . So if we let $\boldsymbol{\mu}$ denote the sequence of nonnegative numbers $\boldsymbol{\mu} := (\mu_i, i = 1, 2, \dots)$, then, in conformity with earlier notation, we can identify the probability measure P with $P_{\boldsymbol{\mu}}$. Conversely, if $\boldsymbol{\mu}$ is any sequence of nonnegative numbers such that $\sum_{i=1}^{\infty} \mu_i = 1$, then the associated probability measure $P_{\boldsymbol{\mu}}$ is defined for every subset $A \subseteq \mathbb{A}$ by

$$P_{\boldsymbol{\mu}}(A) = \sum_{a_i \in A} \mu_i = \sum_{i=1}^{\infty} \mu_i I_A(a_i),$$

where as before $I_A(\cdot)$ denotes the indicator function of the set A .

If $\mathbb{A}$ is finite, and if $\{A_1, A_2, \dots\}$ is a countable collection of pairwise disjoint sets, then the only possibility is that all but finitely many sets are empty. So Property 3 above can be simplified to:

$$P(A \cup B) = P(A) + P(B) \text{ if } A \cap B = \emptyset,$$

which is precisely Property 2 from Theorem 1.3. In other words, the case where $\mathbb{A}$ is countably infinite is not any more complicated than the case where $\mathbb{A}$ is a finite set. This is why most “elementary” books on Markov chains assume that the underlying set $\mathbb{A}$ is countably infinite. But if $\mathbb{A}$ is an uncountable infinite set (such as the real numbers, for example), this approach based on assigning weights to individual elements of the set $\mathbb{A}$ does not work, and one requires the more general version of the theory of probability as introduced in [81].

At this point the reader can well ask: But what does it all *mean*? As with much of mathematics, probability theory exists at many distinct levels. It can be viewed as an exercise in pure reasoning, an intellectual pastime, a challenge to one’s wits. While that may satisfy some persons, the theory would have very little by way of *application* to “real” situations unless the notion of probability is given a little more concrete interpretation. So we can think of the probability distribution μ as arising in one of two ways. First, the distribution can be *postulated*, as in the previous subsection. Thus if we are drawing from an urn containing four balls that are identical in all respects save their color, it makes sense to *postulate* that each of the four outcomes is equally likely. Similarly, if the balls are identical except for their diameter, and if we believe that the likelihood of drawing a ball is proportional to the surface area, then once again we can *postulate* that the four components of μ are in proportion to the surface areas (or equivalently, to the diameter squared) of the four balls. Then the requirement that the components of μ must add up to one gives the normalizing constant. Second, the distribution can be *estimated*, as with the adhesive-coated balls in Example 1.3. In this case there is a true but unknown probability vector μ , and our estimate of μ , based on 1,000 draws of balls, is $\hat{\mu} = [0.451 \ 0.187 \ 0.174 \ 0.188]$. Then we can try to develop theories that allow us to say *how close $\hat{\mu}$ is to μ , and with what confidence we can make this statement*. This question is addressed in Section 1.3.3.

1.1.3 Function of a Random Variable, Expected Value

Suppose X is a random variable assuming values in a finite set $\mathbb{A} = \{a_1, \dots, a_n\}$, with the probability measure P_μ and the probability distribution μ . Suppose f is a function mapping the set $\mathbb{A}$ into another set $\mathbb{B}$. Since $\mathbb{A}$ is finite, it is clear that the set $\{f(a_1), \dots, f(a_n)\}$ is finite. So there is no loss of generality in assuming that the set $\mathbb{B}$ (the range of the function f) is also a finite set. Moreover, it is *not* assumed that the values $f(a_1), \dots, f(a_n)$ are distinct. Thus the image of the set $\mathbb{A}$ under the function f can have

fewer than n elements. Now $f(X)$ is itself a random variable. Moreover, the distribution of $f(X)$ can be computed readily from the distribution of X . Suppose $\mu \in \mathcal{S}_n$ is the distribution of X . Thus $\mu_i = \Pr\{X = x_i\}$. To compute the distribution of $f(X)$, we need to address the possibility that $f(a_1), \dots, f(a_n)$ may not be distinct elements. Let $\mathbb{B} = \{b_1, \dots, b_m\}$ denote the set of all possible outcomes of $f(X)$, and note that $m \leq n$. Then

$$\Pr\{f(X) = b_j\} = \sum_{a_i \in f^{-1}(b_j)} \mu_i.$$

In other words, the probability that $f(X) = b_j$ is the sum of the probabilities of all the preimages of b_j under the function f .

Example 1.4 Suppose that in Example 1.1, we receive a payoff of \$2 if we draw a green ball, we pay a penalty of \$1 if we draw a red ball, and we neither pay a penalty nor receive a payment if we draw a white ball or a blue ball. The situation can be represented by defining $X = \{R, B, W, G\}$, $Y = \{-1, 0, 2\}$, and $f(R) = -1$, $f(G) = 2$, and $f(W) = f(B) = 0$. Moreover, if the balls are identical except for color, then

$$\Pr\{Y = -1\} = \Pr\{X = R\} = 0.25,$$

$$\Pr\{Y = 0\} = \Pr\{X = B\} + \Pr\{X = W\} = 0.5,$$

$$\Pr\{Y = 2\} = \Pr\{X = G\} = 0.25.$$

Example 1.5 We give an example from biology, which is discussed in greater detail in Chapter 8. Let $\mathcal{D}$ equal $\{A, C, G, T\}$, the set of DNA nucleotides, and let $\mathcal{R}$ equal $\{A, C, G, U\}$, the set of RNA nucleotides. As explained in Chapter 8, during the transcription phase of DNA replication thymine (T) gets replaced by uracil (U). Then, during the translation phase of DNA replication, each triplet of RNA nucleotides, known as a “codon,” gets converted into one of 20 amino acids, or into the STOP symbol. The map that associates each codon with the corresponding amino acid (or the STOP symbol) is called the “genetic code.” The complete genetic code was discovered by Hargobind mhorana, building upon earlier work by Marshall Nirenberg, and is shown in Figure 1.1. For this pioneering effort, Khorana and Nirenberg were awarded the Nobel Prize in medicine in 1968, along with Robert Holley who discovered tRNA (translation RNA). Further details can be found in Section 8.1.2.

Now let $\mathbb{A} = \mathcal{R}^3$, the set of codons, which has cardinality $4^3 = 64$, and let $\mathbb{B}$ denote the set of amino acids plus the STOP symbol, which has cardinality 21. Then the genetic code can be thought of as a function $f : \mathbb{A} \rightarrow \mathbb{B}$. From Figure 1.1 it is clear that the number of preimages $f^{-1}(b)$ varies quite considerably for each of the 21 elements of $\mathbb{B}$, ranging from a high of 6 for leucine (Leu) down to 1 for tryptophan (Trp). Thus, if we know the frequency of distribution of codons in a particular stretch of genome (a frequency distribution on $\mathbb{A}$), we can convert this into a corresponding frequency distribution of amino acids and stop codons.

		Second letter				
		U	C	A	G	
First letter	U	UUU } Phe UUC } UUA } Leu UUG }	UCU } Ser UCC } UCA } UCG }	UAU } Tyr UAC } UAA Stop UAG Stop	UGU } Cys UGC } UGA Stop UGG Trp	U C A G
	C	CUU } CUC } CUA } CUG }	CCU } CCC } CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } CGC } CGA } CGG }	U C A G
	A	AUU } Ile AUC } AUA } AUG Met	ACU } ACC } ACA } ACG }	AAU } Asn AAC } AAA } AAG } Lys	AGU } Ser AGC } AGA } AGG } Arg	U C A G
	G	GUU } GUC } GUA } GUG }	GCU } GCC } GCA } GCG }	GAU } Asp GAC } GAA } GAG } Glu	GGU } GGC } GGA } GGG }	U C A G

Figure 1.1 The Genetic Code in Tabular Form. Reproduced with permission from Stephen Carr.

Definition 1.4 Suppose X is a real-valued random variable assuming values $\mathbb{A} = \{a_1, \dots, a_n\} \subseteq \mathbb{R}$, with the probability distribution μ , and associated probability measure P_μ . Then the **expected value** of X is denoted by $E[X, P_\mu]$ and is defined by

$$E[X, P_\mu] := \sum_{i=1}^n a_i \mu_i. \quad (1.5)$$

It is important to note that, while the notion of probability can be defined for *any* random variable (for example, the set of nucleotides or the set of amino acids), the notion of an expected value can be defined only for *real-valued* random variables. Also, if the underlying probability distribution P_μ is clear from the context, we sometimes omit it and write $E(X)$ or $E[X]$ instead of $E(X, P_\mu)$ or $E[X, P_\mu]$.

Suppose now that X is a random variable assuming values in some finite set $\mathbb{A}$ (not necessarily a subset of $\mathbb{R}$), and f is a function mapping the set $\mathbb{A}$ into the real numbers $\mathbb{R}$. Thus to each element $a_i \in \mathbb{A}$, the function f assigns a real number $f(a_i)$. Let μ denote the distribution of X and let P_μ denote the associated probability measure.

Definition 1.5 The **expected value** of the function f is denoted by $E[f, P_\mu]$ and is defined by

$$E[f, P_\mu] := \sum_{i=1}^n f(a_i) \mu_i = \sum_{i=1}^n f(a_i) \Pr\{X = a_i\}. \quad (1.6)$$

The point to note in the above definition is that it is permissible for X to be an “abstract” random variable, not necessarily real-valued, so long as

$f(X)$ is real-valued. It is left to the reader to verify that the above equation is the same as the expected value of the real-valued random variable $f(X)$. The point to note is that the formula (1.6) is valid even if the real numbers $f(a_1), \dots, f(a_n)$ are not all distinct.

Observe that the expected value is *linear* in the function f . Thus, if f, g are two functions of a random variable X with probability measure P_μ and the probability distribution μ , and α, β are two real numbers, then

$$E[\alpha f + \beta g, P_\mu] = \alpha E[f, P_\mu] + \beta E[g, P_\mu].$$

Suppose X is a *real-valued* random variable assuming values in $\mathbb{A} = \{a_1, \dots, a_n\}$, with probability measure P_μ and the probability distribution μ . Then the quantity defined earlier as the expected value of X , namely

$$E[X, P_\mu] = \sum_{i=1}^n a_i \mu_i,$$

is also called the **mean value** or just simply the **mean** of X . The quantity $E[(X - E[X])^2, P_\mu]$ is called the **variance** of X . The square root of the variance is called the **standard deviation** of X . In many books, the symbols $\mu(X)$, $V(X)$, and $\sigma(X)$ are commonly used to denote the mean, the variance, and standard deviation respectively. This notation is used only rarely in this book.

Note that we can also define the variance of X as $E[X^2, P_\mu] - (E[X, P_\mu])^2$. This is because, by the linearity of the expected value operation, we have

$$\begin{aligned} E[(X - E[X])^2, P_\mu] &= E[X^2, P_\mu] - 2(E[X, P_\mu])^2 + (E[X, P_\mu])^2 \\ &= E[X^2, P_\mu] - (E[X, P_\mu])^2. \end{aligned} \quad (1.7)$$

The above argument also shows that, for every random variable X , we have

$$E[X^2, P_\mu] \geq (E[X, P_\mu])^2.$$

This is a special case of a very general result known as Schwarz's inequality.

In general, for every positive integer l , the quantity

$$E[X^l, P_\mu] = \sum_{i=1}^n a_i^l \mu_i$$

is called the **l -th moment** of the random variable X . Since the random variable assumes only finitely many values, the l -th moment is well-defined for every positive integer l . In this connection, one can define the so-called **moment generating function** $\text{mgf}(\lambda; X)$ as follows:

$$\text{mgf}(\lambda; X) = E[\exp(\lambda X), P_\mu] = \sum_{i=1}^n \exp(\lambda a_i) \mu_i.$$

Again, since X assumes only finitely many values, the moment generating function is well-defined for every real number λ . Moreover, it is a direct consequence of the definition that

$$\left[\frac{d^l \text{mgf}(\lambda; X)}{d\lambda} \right]_{\lambda=0} = \left[\sum_{i=1}^n a_i^l \exp(\lambda a_i) \mu_i \right]_{\lambda=0} = \sum_{i=1}^n a_i^l \mu_i$$

is indeed the l -th moment of the random variable X .

1.1.4 Total Variation Distance

Suppose $\mathbb{A} = \{a_1, \dots, a_n\}$ is a finite set, and μ, ν are two probability distributions on X . Let P_μ, P_ν denote the corresponding probability measures. In this section, we show how to quantify the “difference” between the two measures.

Definition 1.6 Let P_μ, P_ν be two probability measures on a finite set $\mathbb{A} = \{a_1, \dots, a_n\}$, corresponding to the distributions μ, ν respectively. Then the **total variation distance** between P_μ and P_ν (or between μ and ν), denoted by $\rho(P_\mu, P_\nu)$ or $\rho(\mu, \nu)$ is defined as

$$\rho(P_\mu, P_\nu), \rho(\mu, \nu) := \max_{S \subseteq \mathbb{A}} |P_\mu(S) - P_\nu(S)|. \quad (1.8)$$

Now it is shown that $\rho(\cdot, \cdot)$ is indeed a proper metric or “distance” on $\mathbb{S}_n$, which can be identified with the set of all probability distributions on the set $\mathbb{A}$ of cardinality n .

Lemma 1.7 The function $\rho(\cdot, \cdot)$ defined in (1.8) satisfies the following properties:

1. $\rho(P_\mu, P_\nu) \geq 0$ for all $\mu, \nu \in \mathbb{S}_n$.
2. $\rho(P_\mu, P_\nu) = 0$ if and only if $\mu = \nu$.
3. $\rho(P_\mu, P_\nu) = \rho(P_\nu, P_\mu)$ for all $\mu, \nu \in \mathbb{S}_n$.
4. The “triangle inequality” is satisfied, namely:

$$\rho(P_\mu, P_\nu) \leq \rho(P_\mu, P_\phi) + \rho(P_\phi, P_\nu) \quad \forall \mu, \nu, \phi \in \mathbb{S}_n. \quad (1.9)$$

Proof. Property 1 is obvious. To prove Property 2, note that $\rho(P_\mu, P_\nu) = 0$ if $\mu = \nu$. Thus the key observation is the converse, or equivalently, $\rho(P_\mu, P_\nu) > 0$ if $\mu \neq \nu$. Note that if $\mu \neq \nu$, then $\mu_i \neq \nu_i$ for at least one index i (actually for at least two indices). Let $S = \{i\}$, where i is an index such that $\mu_i \neq \nu_i$. Then

$$P_\mu(S) = \mu_i \neq \nu_i = P_\nu(S).$$

Hence $\rho(P_\mu, P_\nu) > 0$. Property 3 is again obvious. Finally, Property 4 follows from the triangle inequality for real numbers, namely:

$$|x - y| \leq |x - z| + |z - y|, \quad \forall x, y, z \in \mathbb{R}.$$

Now suppose $S \subseteq \mathbb{A}$ is arbitrary. Then the triangle inequality for real numbers implies that

$$|P_\mu(S) - P_\nu(S)| \leq |P_\mu(S) - P_\phi(S)| + |P_\phi(S) - P_\nu(S)|, \quad \forall A \subseteq \mathbb{A}.$$

Taking the maximum over all $S \subseteq \mathbb{A}$ proves Property 4. $\square$

As defined in (1.8), $\rho(P_\mu, P_\nu)$ is the maximum difference between $P_\mu(S)$ and $P_\nu(S)$ as S varies over the 2^n subsets of $\mathbb{A}$. Clearly, (1.8) is an impractical formula for actually *computing* the number $\rho(P_\mu, P_\nu)$. The next theorem gives a number of equivalent formulas for computing $\rho(P_\mu, P_\nu)$. Note that, given a real number $x \in \mathbb{R}$, the symbol x_+ denotes the positive part of x , that is, $\max\{x, 0\}$. Similarly x_- denotes $\min\{x, 0\}$.

Theorem 1.8 Suppose $\mathbb{A} = \{a_1, \dots, a_n\}$ is a finite set, and that P_μ, P_ν are two probability measures on $\mathbb{A}$ with associated distributions μ and ν respectively. Then

$$\rho(P_\mu, P_\nu) = \sum_{i=1}^n (\mu_i - \nu_i)_+ \quad (1.10)$$

$$= - \sum_{i=1}^n (\mu_i - \nu_i)_- \quad (1.11)$$

$$= \frac{1}{2} \sum_{i=1}^n |\mu_i - \nu_i|. \quad (1.12)$$

Proof. Define $\delta_i := \mu_i - \nu_i$, for $i = 1, \dots, n$. Then, since $\mu, \nu \in \mathbb{S}_n$, it follows that $\sum_{i=1}^n \delta_i = 0$. Moreover, for any set $S \subseteq \mathbb{A}$, we have that

$$P_\mu(S) - P_\nu(S) = \sum_{a_i \in S} \delta_i = \sum_{i=1}^n I_S(a_i) \delta_i,$$

where $I_S(\cdot)$ is the indicator function of the set S . Now, let us look at the 2^n numbers $P_\mu(S) - P_\nu(S)$ generated by varying S over all subsets of $\mathbb{A}$. (These numbers may not all be distinct.) Let $K \subseteq \mathbb{R}$ denote the set of all these numbers. The first point to note is that

$$P_\mu(S^c) - P_\nu(S^c) = [1 - P_\mu(S)] - [1 - P_\nu(S)] = -[P_\mu(S) - P_\nu(S)].$$

Hence the set K is symmetric: If $x \in K$ (corresponding to a set S), then $-x \in K$ (corresponding to the set S^c). Observe that $\rho(P_\mu, P_\nu)$ is the largest value of $|x|$ for all $x \in K$. However, because of the symmetry of the set K , $\rho(P_\mu, P_\nu)$ also equals the largest value of $x \in K$, and also

$$\rho(P_\mu, P_\nu) = -\min\{x \in K\}.$$

So if we can find the largest or the smallest element in K , then we would have found $\rho(P_\mu, P_\nu)$.

Next, let $\mathbb{N}_+ \subseteq \mathbb{N} := \{1, \dots, n\}$ denote the set of indices i for which $\delta_i \geq 0$, and let $\mathbb{N}_-$ denote the set of indices i for which $\delta_i < 0$. Then

$$P_\mu(S) - P_\nu(S) = \sum_{i \in \mathbb{N}_+} I_S(a_i) \delta_i + \sum_{i \in \mathbb{N}_-} I_S(a_i) \delta_i.$$

Now the first summation consists of only nonnegative numbers, while the second summation consists only of nonpositive numbers. Therefore the largest possible value of $P_\mu(S) - P_\nu(S)$ is

$$\sum_{i \in \mathbb{N}_+} I_S(a_i) \delta_i = \sum_{i=1}^n (\delta_i)_+,$$

and corresponds to the choice $S = \{a_i : i \in \mathbb{N}_+\}$. By the discussion in the preceding paragraph, it follows that

$$\rho(P_\mu, P_\nu) = \sum_{i=1}^n (\delta_i)_+,$$

which is precisely (1.10). Similarly, the smallest value of $P_\mu(S) - P_\nu(S)$ is

$$\sum_{i \in \mathbb{N}_-} I_S(a_i) \delta_i = \sum_{i=1}^n (\delta_i)_-,$$

corresponding to the choice $S = \{a_i : i \in \mathbb{N}_-\}$. Again from the discussion in the previous paragraph, it follows that

$$\rho(P_\mu, P_\nu) = - \sum_{i=1}^n (\delta_i)_-,$$

which is precisely (1.11). Finally, observe that

$$\sum_{i=1}^n |\delta_i| = \sum_{i=1}^n (\delta_i)_+ - \sum_{i=1}^n (\delta_i)_- = 2\rho(P_\mu, P_\nu),$$

which establishes (1.12). $\square$

From the definition (1.8), it is immediate that $\rho(P_\mu, P_\nu) \in [0, 1]$. This is because both $P_\mu(S)$ and $P_\nu(S)$ lie in the range $[0, 1]$, and so $P_\mu(S) - P_\nu(S) \in [-1, 1]$. Now the proof of Theorem 1.8 shows that, for every pair of probability measures P_μ and P_ν , there actually exists a set S such that $P_\mu(S) - P_\nu(S) = \rho(P_\mu, P_\nu)$; one such choice is $S = \{a_i : i \in \mathbb{N}_+\}$. Now, if $\rho(P_\mu, P_\nu)$ actually equals one, this implies that $P_\mu(S) = 1$ and $P_\nu(S) = 0$ (and also that $P_\mu(S^c) = 0$ and $P_\nu(S^c) = 1$). In such a case the two measures P_μ and P_ν are said to be **mutually singular**, because their weights are supported on disjoint sets: The weights μ_i are concentrated on the set S whereas the weights ν_i are concentrated on the set S^c .

Lemma 1.9 Suppose $\mathbb{A} = \{a_1, \dots, a_n\}$ is a finite set, and P_μ, P_ν are two probability measures on $\mathbb{A}$. Suppose $f : \mathbb{A} \rightarrow [-1, 1]$. Then

$$|E[f, P_\mu] - E[f, P_\nu]| \leq 2\rho(P_\mu, P_\nu). \quad (1.13)$$

Proof. The proof follows by direct substitution. We have that

$$\begin{aligned} |E[f, P_\mu] - E[f, P_\nu]| &= \left| \sum_{i=1}^n f(a_i)(\mu_i - \nu_i) \right| \\ &\leq \sum_{i=1}^n |f(a_i)| \cdot |\mu_i - \nu_i| \\ &\leq \sum_{i=1}^n |\mu_i - \nu_i| \text{ since } |f(a_i)| \leq 1 \\ &= 2\rho(P_\mu, P_\nu). \end{aligned}$$

$\square$

Lemma 1.10 Suppose $\mathbb{A} = \{a_1, \dots, a_n\}$ is a finite set, and P_μ, P_ν are two probability measures on $\mathbb{A}$. Suppose $f : \mathbb{A} \rightarrow [0, 1]$. Then

$$|E[f, P_\mu] - E[f, P_\nu]| \leq \rho(P_\mu, P_\nu). \quad (1.14)$$

Proof. This lemma can be derived as a corollary of Lemma 1.9 by observing that $f(X) - 0.5$ assumes values in $[-0.5, 0.5]$; but we will give an alternate proof. We have that

$$\begin{aligned} E[f, P_\mu] - E[f, P_\nu] &= \sum_{i=1}^n f(a_i)(\mu_i - \nu_i) \\ &\leq \sum_{i=1}^n f(a_i) \cdot (\mu_i - \nu_i)_+ \text{ since } 0 \leq f(a_i) \forall i \\ &\leq \sum_{i=1}^n (\mu_i - \nu_i)_+ \text{ since } f(a_i) \leq 1 \forall i \\ &= \rho(P_\mu, P_\nu). \end{aligned}$$

By entirely analogous reasoning, it follows that

$$E[f, P_\mu] - E[f, P_\nu] \geq \sum_{i=1}^n (\mu_i - \nu_i)_- = -\rho(P_\mu, P_\nu).$$

The desired bound now follows by combining these two inequalities. $\square$

Problem 1.1 Let X be a random variable assuming values in the four-color alphabet $\mathbb{A} = \{R, B, W, G\}$ as in Examples 1.1 and 1.2, and let $f : \mathbb{A} \rightarrow \mathbb{R}$ be the payoff function defined in Example 1.4. First let X have the probability distribution in Example 1.1. Compute the mean, variance, and standard deviation of X . Repeat when X has the probability distribution in Example 1.2.

Problem 1.2 Suppose X is a “binary” random variable, assuming just two real values, namely 0 and 1, with $\Pr\{X = 1\} = \alpha \in (0, 1)$. We denote this by $X = \mathcal{B}(1, \alpha)$. Compute the mean and standard deviation of X .

Problem 1.3 Suppose an urn contains both white and black balls in the proportion α to $1 - \alpha$.² Let X be the associated binary random variable as defined in Problem 1.2 above. Now suppose we draw n balls from the urn, one after the other, replacing the ball drawn after each trial. Let Y_n denote the number of white balls drawn after n trials. Then Y_n is a “binomially distributed” random variable, whereby

$$\Pr\{Y_n = i\} = \binom{n}{i} \alpha^i (1 - \alpha)^{n-i},$$

where

$$\binom{n}{i} = \frac{n!}{i!(n-i)!}$$

²This suggests that α is a rational number, but the problem makes sense even without this assumption.

is called the combinatorial parameter. Y_n is the number of different sequences of n draws containing precisely i white balls. We denote this by $Y_n = \mathcal{B}(n, \alpha)$. Compute the mean and standard deviation of Y_n .

In case α is an irrational number, the urn and balls interpretation is not meaningful. Instead we should think in terms of generating n independent outcomes $X_1, \dots, X_n$ where each X_i is binary with $\Pr\{X_i = 1\} = \alpha$. Such a sequence of random variables is known as a “Bernoulli process” or a set of “Bernoulli trials.” If we equate an outcome of 1 with “success” and 0 with “failure,” then Y_n is the number of successes in n Bernoulli trials.

Problem 1.4 A counterpart of the binomial distribution is the hypergeometric distribution. Suppose an urn contains N balls, out of which M are white. Now suppose we draw n balls one after the other, but this time *without replacing* the ball drawn. Let Z denote the resulting number of white balls. Show that

$$\Pr\{Z = i\} = \frac{\binom{M}{i} \binom{N-M}{n-i}}{\binom{N}{n}},$$

with the convention that

$$\binom{M}{i} = 0 \text{ if } M < i.$$

We denote this by $Z = \mathcal{H}(n, \alpha, N)$ where $\alpha = M/N$ is the fraction of white balls in the urn.

Problem 1.5 Suppose $Z = \mathcal{H}(n, \alpha, N)$ have the hypergeometric distribution. Show that, as $N \rightarrow \infty$, the hypergeometric distribution approaches the binomial distribution. Can you explain why?

Problem 1.6 Suppose $\mu = [0.4 \ 0.6], \nu = [0.6 \ 0.4]$. Compute the total variation distance $\rho(\mu, \nu)$.

Problem 1.7 Let $n = 10$, and let Y, Z be binomially distributed random variables, with $Y_n = \mathcal{B}(10, 0.6), Z = \mathcal{B}(10, 0.4)$. Compute the total variation distance between the probability distributions of Y and Z .

Problem 1.8 Given a finite set $\mathbb{A}$, let $\mathcal{F}(\mathbb{A})$ denote all functions mapping $\mathbb{A}$ into the interval $[0, 1]$. Suppose μ, ν are two probability distributions on $\mathbb{A}$. Show that

$$\max_{f \in \mathcal{F}(\mathbb{A})} |E[f, P_\mu] - E[f, P_\nu]| = \rho(P_\mu, P_\nu).$$

Problem 1.9 Prove the following generalization of Lemma 1.9: Suppose $f : \mathbb{A} \rightarrow [a, b]$ where a, b are real numbers, and that μ, ν are probability distributions on $\mathbb{A}$. Then

$$|E[f, P_\mu] - E[f, P_\nu]| \leq (b - a)\rho(P_\mu, P_\nu).$$

Is this the best possible bound?

1.2 MULTIPLE RANDOM VARIABLES

1.2.1 Joint and Marginal Distributions

Up to now we have discussed only one random variable. It is also possible to have more than one random variable, each assuming values in its own set. Suppose $\mathbb{A} = \{a_1, \dots, a_n\}$ and $\mathbb{B} = \{b_1, \dots, b_m\}$ are finite sets. Then the Cartesian product $\mathbb{A} \times \mathbb{B}$ consists of all pairs of elements from the two sets; specifically

$$\mathbb{A} \times \mathbb{B} = \{(a_i, b_j) : a_i \in \mathbb{A}, b_j \in \mathbb{B}\}.$$

Moreover, the “joint random variable” $Z = (X, Y)$ assumes values in $\mathbb{A} \times \mathbb{B}$. Define

$$\phi_{ij} = \Pr\{X = a_i \wedge Y = b_j\},$$

where the wedge symbol $\wedge$ denotes “and.” Then clearly the vector $\phi \in \mathbb{S}_{nm}$. We can think of ϕ as the probability distribution of some random variable that assumes values in a set of cardinality nm . But the fact that the range of values of Z is a cartesian product of two sets allows us to carry out a refined analysis. The probability distribution ϕ on the set $\mathbb{A} \times \mathbb{B}$ is called the **joint distribution** of the variables X and Y . Thus

$$\Pr\{Z = (a_i, b_j)\} = \Pr\{X = a_i \wedge Y = b_j\} = \phi_{ij}, \quad \forall a_i \in X, b_j \in Y.$$

So we can arrange the nm elements of ϕ in an array, as shown below.

$$\Phi = \begin{bmatrix} \phi_{11} & \dots & \phi_{1m} \\ \vdots & \vdots & \vdots \\ \phi_{n1} & \dots & \phi_{nm} \end{bmatrix}.$$

Up to now we have gained nothing by arranging the nm values of the probability distribution in an array. But now can take the analysis to another level.

Let us define the vectors ϕ_X and ϕ_Y as follows.

$$(\phi_X)_i := \sum_{j=1}^m \phi_{ij}, \quad i = 1, \dots, n, \quad (1.15)$$

$$(\phi_Y)_j := \sum_{i=1}^n \phi_{ij}, \quad j = 1, \dots, m. \quad (1.16)$$

Then it follows that $\phi_X \in \mathbb{S}_n$ and $\phi_Y \in \mathbb{S}_m$. This is because ϕ is a probability distribution and as a result we have

$$\left(\sum_{i=1}^n \sum_{j=1}^m \phi_{ij} = 1 \right) \Rightarrow \left(\sum_{i=1}^n \left[\sum_{j=1}^m \phi_{ij} \right] = 1 \right) \Rightarrow \left(\sum_{i=1}^n (\phi_X)_i = 1 \right).$$

Similarly

$$\left(\sum_{i=1}^n \sum_{j=1}^m \phi_{ij} = 1 \right) \Rightarrow \left(\sum_{j=1}^m (\phi_Y)_j = 1 \right).$$

So $\phi_X \in \mathbb{S}_n$ and $\phi_Y \in \mathbb{S}_m$. The distribution ϕ_X is called the **marginal distribution** of the random variable X . Similarly ϕ_Y is called the marginal distribution of the random variable Y . Depending on the context, it may be more natural to write $\phi_{\mathbb{A}}$ in the place of ϕ_X and $\phi_{\mathbb{B}}$ in the place of ϕ_Y . Mimicking earlier notation, we refer to the measure corresponding to the distribution $\phi \in \mathbb{S}_{nm}$ as the **joint measure** P_ϕ of the joint random variable (X, Y) . We refer to the measure corresponding to the distribution $\phi_X \in \mathbb{S}_n$ as the **marginal measure** of X (or the marginal measure on the set $\mathbb{A}$), and denote it by $P_{\phi, X}$ or $P_{\phi, \mathbb{A}}$. The symbols $P_{\phi, Y}$ and $P_{\phi, \mathbb{B}}$ are defined analogously.

Now we proceed to show that indeed it is the case that

$$\Pr\{X = a_i\} = (\phi_X)_i, \forall i = 1, \dots, n.$$

To see this, fix an index i and observe that the singleton sets $\{(a_i, b_1)\}$ through $\{(a_i, b_m)\}$ are all pairwise disjoint subsets of $\mathbb{A} \times \mathbb{B}$. Moreover, it is clear that

$$\{(X, Y) \in \mathbb{A} \times \mathbb{B} : X = a_i\} = \bigcup_{j=1}^m \{(a_i, b_j)\}.$$

Hence from Property 2 of Theorem 1.3, we can conclude that

$$\begin{aligned} \Pr\{X = a_i\} &= \Pr\{(X, Y) \in \mathbb{A} \times \mathbb{B} : X = a_i\} \\ &= P_\phi \left(\bigcup_{j=1}^m \{(a_i, b_j)\} \right) \\ &= \sum_{j=1}^m P_\phi\{(a_i, b_j)\} \\ &= \sum_{j=1}^m \phi_{ij} = (\phi_X)_i. \end{aligned}$$

By entirely analogous reasoning, it follows that

$$\Pr\{Y = b_j\} = \sum_{i=1}^n \phi_{ij} = (\phi_Y)_j, \forall j = 1, \dots, m.$$

1.2.2 Independence and Conditional Distributions

Up to now we have introduced the notion of a joint distribution of the two random variables X and Y , as well as their individual distributions, which can be obtained as the marginal distributions of the joint distribution. The next notion is among the most fundamental notions of probability theory.

Definition 1.11 Suppose X, Y are random variables assuming values in finite sets $\mathbb{A}$ and $\mathbb{B}$ respectively. Let P_ϕ denote their joint probability measure, and let $\phi \in \mathbb{S}_{nm}$ denote their joint distribution. Then the two random variables are said to be **independent** under the measure P_ϕ (or the distribution ϕ) if

$$\phi_{ij} = (\phi_X)_i \cdot (\phi_Y)_j, \forall i = 1, \dots, n, j = 1, \dots, m. \quad (1.17)$$

An equivalent way of stating (1.17) is:

$$\Pr\{X = a_i \wedge Y = b_j\} = \Pr\{X = a_i\} \cdot \Pr\{Y = b_j\}, \quad \forall a_i \in \mathbb{A}, b_j \in \mathbb{B}. \quad (1.18)$$

The above definition can be made a little more intuitive by introducing the notion of a “product” distribution. Suppose $\mu \in \mathbb{S}_n, \nu \in \mathbb{S}_m$ are distributions on the sets $\mathbb{A}, \mathbb{B}$ respectively. Then their **product distribution** $\mu \times \nu$ on the set $\mathbb{A} \times \mathbb{B}$ is defined by

$$(\mu \times \nu)_{ij} = \mu_i \cdot \nu_j, \quad \forall i, j. \quad (1.19)$$

With this definition, it follows that the two random variables X, Y are independent under the distribution ϕ if and only if $\phi = \phi_X \times \phi_Y$.

In the sequel we will often have occasion to speak about “independent and identically distributed” random variables. This notion can be made to fit into the above frame work by using product distributions where each of the marginals is the same. Thus if $\mu \in \mathbb{S}_n$, then the product distribution $\mu^2 \in \mathbb{S}_{n^2}$ is defined by

$$(\mu^2)_{ij} = \mu_i \cdot \mu_j, \quad \forall i, j.$$

The associated probability measure is often denoted by P_μ^2 . Higher “powers” of μ and P_μ are defined in an entirely analogous fashion.

Theorem 1.12 *Suppose X, Y are random variables assuming values in finite sets $\mathbb{A}$ and $\mathbb{B}$ respectively. Suppose $\phi \in \mathbb{S}_{nm}$ is their joint distribution, that P_ϕ is their joint measure, and that X, Y are independent under the measure P_ϕ . Suppose further that $f : \mathbb{A} \rightarrow \mathbb{R}, g : \mathbb{B} \rightarrow \mathbb{R}$ are functions on $\mathbb{A}$ and $\mathbb{B}$ respectively. Then*

$$E[f(X)g(Y), P_\phi] = E[f(X), P_{\phi, X}] \cdot E[g(Y), P_{\phi, Y}]. \quad (1.20)$$

The point of the theorem is this: Even if f is a function of X alone and g is a function of Y alone, the function fg depends on both X and Y , and the pair (X, Y) has the joint measure P_ϕ . For an arbitrary joint probability measure P_ϕ , we cannot say anything about the expected value of the function $f(X)g(Y)$ under the measure P_ϕ . However, if the two random variables are *independent* under the measure P_ϕ , then the expected value factors neatly into the product of two different expected values, namely the expected value of f under the marginal measure $P_{\phi, X}$, and the expected value of g under the marginal measure $P_{\phi, Y}$.

Proof. The proof is a ready consequence of (1.17). We have

$$\begin{aligned} E[f(X)g(Y), P_\phi] &= \sum_{i=1}^n \sum_{j=1}^m f(a_i)g(b_j)\phi_{ij} \\ &= \sum_{i=1}^n \sum_{j=1}^m f(a_i)g(b_j)(\phi_X)_i(\phi_Y)_j \\ &= \left[\sum_{i=1}^n f(a_i)(\phi_X)_i \right] \cdot \left[\sum_{j=1}^m g(b_j)(\phi_Y)_j \right] \\ &= E[f(X), P_{\phi, X}] \cdot E[g(Y), P_{\phi, Y}]. \end{aligned}$$

This is precisely the desired conclusion. $\square$

The above observation motivates the notion of the correlation coefficient between two real-valued random variables.

Definition 1.13 Suppose X, Y are real-valued random variables assuming values in finite sets $\mathbb{A}, \mathbb{B} \subseteq \mathbb{R}$ respectively. Let ϕ denote their joint distribution, and ϕ_X, ϕ_Y the two marginal distributions. Let $E[XY, \phi]$, $E[X, \phi_X]$, $E[Y, \phi_Y]$ denote expectations, and let $\sigma(X), \sigma(Y)$ denote the standard deviations of X, Y under their respective marginal distributions. Then the quantity

$$C(X, Y) := \frac{E[XY, \phi] - E[X, \phi_X]E[Y, \phi_Y]}{\sigma(X)\sigma(Y)} \quad (1.21)$$

is called the **correlation coefficient** between X and Y .

Note that some authors refer to $C(X, Y)$ as the “Pearson” correlation coefficient after its inventor. It can be shown that the correlation coefficient $C(X, Y)$ always lies in the interval $[-1, 1]$. It is often said that X, Y are **uncorrelated** if $C(X, Y) = 0$, **positively correlated** if $C(X, Y) > 0$, and **negatively correlated** if $C(X, Y) < 0$. One of the advantages of the correlation coefficient is that it is invariant under both scaling and centering. In other words, if $\alpha, \beta, \gamma, \delta$ are any real numbers, then

$$C(\alpha X + \beta, \gamma Y + \delta) = C(X, Y). \quad (1.22)$$

If two random variables are independent, then they are uncorrelated. However, the converse statement is most definitely not true; see Problem 1.12.

The next definition is almost as important as the notion of independence.

Definition 1.14 Suppose X, Y are random variables assuming values in finite sets $\mathbb{A}$ and $\mathbb{B}$ respectively, and let $\phi \in \mathbb{S}_{nm}$ denote their joint distribution. The **conditional probability** of X given an observation $Y = b_j$ is defined as

$$\Pr\{X = a_i | Y = b_j\} := \frac{\Pr\{X = a_i \wedge Y = b_j\}}{\Pr\{Y = b_j\}} = \frac{\phi_{ij}}{\sum_{i'=1}^n \phi_{i'j}}. \quad (1.23)$$

In case $\Pr\{Y = b_j\} = 0$, we define $\Pr\{X = a_i | Y = b_j\} = \Pr\{X = a_i\} = (\phi_X)_i$.

Let us use the notation $\phi_{\{a_i|b_j\}}$ as a shorthand for $\Pr\{X = a_i | Y = b_j\}$. Then the vector

$$\phi_{\{X|Y=b_j\}} := [\phi_{\{a_1|b_j\}} \cdots \phi_{\{a_n|b_j\}}] \in \mathbb{S}_n. \quad (1.24)$$

This is obvious from (1.23). So $\phi_{\{X|Y=b_j\}}$ is a probability distribution on the set $\mathbb{A}$; it is referred to as the **conditional distribution** of X , given that $Y = b_j$. The corresponding probability measure is denoted by $P_{\phi, \{X|Y=b_j\}}$ and is referred to as the **conditional measure** of X , given that $Y = b_j$.

We also use the simplified notation $\phi_{X|b_j}$ and $P_{\phi, X|b_j}$ if the variable Y is clear from the context.

Now we briefly introduce the notion of convex combinations of vectors; we will discuss this idea in greater detail in Section 2.1. If $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ are n -dimensional vectors and $\lambda \in [0, 1]$, then the vector $\lambda\mathbf{x} + (1 - \lambda)\mathbf{y}$ is called a **convex combination** of $\mathbf{x}$ and $\mathbf{y}$. More generally, if $\mathbf{x}_1, \dots, \mathbf{x}_l \in \mathbb{R}^n$ and $\boldsymbol{\lambda} \in \mathbb{S}_l$, then the vector $\sum_{i=1}^l \lambda_i \mathbf{x}_i$ is called a **convex combination** of the vectors $\mathbf{x}_1$ through $\mathbf{x}_l$. In the present context, it is easy to see that

$$(\phi_X)_i = \sum_{j=1}^m (\phi_Y)_j \phi_{\{X|Y=b_j\}}. \quad (1.25)$$

Thus, the marginal distribution ϕ_X is a convex combination of the m conditional distributions $\phi_{\{X|Y=b_j\}}, j = 1, \dots, m$. The proof of (1.25) is a straightforward consequence of the definitions and is left as an exercise.

Thus far we have introduced a lot of terminology and notation, so let us recapitulate. Suppose X and Y are random variables, assuming values in finite sets $\mathbb{A} = \{a_1, \dots, a_n\}$ and $\mathbb{B} = \{b_1, \dots, b_m\}$ respectively. Then they have a *joint probability measure* P_ϕ , defined on the product set $\mathbb{A} \times \mathbb{B}$. Associated with P_ϕ is a *marginal probability* $P_{\phi, X}$, which is a measure on $\mathbb{A}$, and a *marginal probability* $P_{\phi, Y}$, which is a measure on $\mathbb{B}$. Finally, for each of the m possible values of Y , there is an associated *conditional probability* $P_{\phi, \{X|Y=b_j\}}$, which is a measure on $\mathbb{A}$. Similarly, for each of the n possible values of X , there is an associated *conditional probability* $P_{\phi, \{Y|X=a_i\}}$, which is a measure on $\mathbb{B}$.

Example 1.6 Let us return to the problem studied earlier of an urn containing four uniform balls with colors red, blue, white and green. Suppose we draw two balls from the urn, one after the other, but *without* replacing the first ball before drawing the second ball. Let X denote the color of the first ball, and Y the color of the second ball. We can ask: What is the probability of drawing a red ball the second time? The answer is somewhat counter-intuitive because, as shown below, the answer is 0.25. We know that, when we make the second draw, there are only three balls in the urn, and which three colors they represent depends on X , the outcome of the first draw. Nevertheless, the probability of drawing a red ball (or any other colored ball) turns out to be 0.25, as is shown next.

Let us first compute the marginal or “unconditional” distribution of X , the outcome of the first draw. Since the balls are assumed to be uniform and there are four balls when we draw for the first time, we can define $\mathbb{A} = \{R, B, W, G\}$ and with this definition the distribution ϕ_X of X is given by

$$\phi_X = [0.25 \ 0.25 \ 0.25 \ 0.25].$$

Now let us compute the conditional probability of Y given X . If $X = R$, then at the second draw there are only B, G, Y in the urn. So we can say that

$$\phi_{\{Y|X=R\}} = [0 \ 1/3 \ 1/3 \ 1/3].$$

Similarly,

$$\phi_{\{Y|X=B\}} = [1/3 \ 0 \ 1/3 \ 1/3],$$

$$\phi_{\{Y|X=W\}} = [1/3 \ 1/3 \ 0 \ 1/3],$$

$$\phi_{\{Y|X=G\}} = [1/3 \ 1/3 \ 1/3 \ 0].$$

Therefore

$$\begin{aligned} \Pr\{Y = R\} &= \Pr\{Y = R|X = R\} \cdot \Pr\{X = R\} + \cdots \\ &\quad + \Pr\{Y = R|X = G\} \cdot \Pr\{X = G\}, \end{aligned}$$

summing over all possible outcomes for Y . Doing this routine calculation shows that

$$\phi_Y = [0.25 \ 0.25 \ 0.25 \ 0.25].$$

This somewhat counter-intuitive result can be explained as follows: When we make the second draw to determine Y , there are indeed only three balls in the urn. However, *which three* they are depends on X , the outcome of the first draw. There are four possible sets of three colors, and each of them is equally likely. Hence the probability of getting a red ball the first time is exactly the same as the probability of getting a red ball the second time, even though we are not replacing the first ball drawn.

Example 1.7 The purpose of this example is to show that it is necessary to verify the condition (1.26) for *every* possible value b_j . Suppose $\mathbb{A} = \{a_1, a_2\}$, $\mathbb{B} = \{b_1, b_2, b_3\}$, and that the joint probability distribution is

$$[\phi_{ij}] = \begin{bmatrix} 0.12 & 0.08 & 0.20 \\ 0.20 & 0.10 & 0.30 \end{bmatrix}.$$

Then it follows from (1.15) and (1.16) that

$$\phi_X = [0.4 \ 0.6] \text{ and } \phi_Y = [0.32 \ 0.18 \ 0.50].$$

It can be readily checked that the condition (1.26) holds when $j = 3$ but not when $j = 1$ or $j = 2$. Hence the variables X and Y are not independent.

Lemma 1.15 *Suppose X, Y are random variables assuming values in finite sets $\mathbb{A}$ and $\mathbb{B}$ respectively, and let $\phi \in \mathbb{S}_{nm}$ denote their joint distribution. Then X and Y are independent if and only if*

$$\phi_{\{X|Y=b_j\}} = \phi_X, \quad \forall b_j \in Y. \quad (1.26)$$

There is an apparent asymmetry in the statement of Lemma 1.15. It appears as though we should say “ X is independent of Y if (1.26) holds” as opposed to “ X and Y are independent if (1.26) holds.” It is left as an exercise to show that (1.26) is equivalent to the statement

$$\phi_{\{Y|X=a_i\}} = \phi_Y, \quad \forall a_i \in X. \quad (1.27)$$

The following observation follows readily from Definition 1.11.

Lemma 1.16 *Suppose X, Y are random variables assuming values in finite sets $\mathbb{A}$ and $\mathbb{B}$ respectively, and let $\phi \in \mathbb{S}_{nm}$ denote their joint distribution. Then X and Y are independent if and only if the matrix*

$$\Phi := \begin{bmatrix} \phi_{11} & \cdots & \phi_{1m} \\ \vdots & \vdots & \vdots \\ \phi_{n1} & \cdots & \phi_{nm} \end{bmatrix}$$

has rank one.

The proof is easy and is left as an exercise.

In the preceding discussion, there is nothing special about having *two* random variables—we can have any finite number of them. We can also condition the probability distribution on multiple events, and the results are consistent. To illustrate, suppose X, Y, Z are random variables assuming values the finite sets $\mathbb{A} = \{a_1, \dots, a_n\}$, $\mathbb{B} = \{b_1, \dots, b_m\}$, $\mathbb{C} = \{c_1, \dots, c_l\}$ respectively. Let P_θ denote their joint probability measure and $\theta = [\theta_{ijk}] \in \mathbb{S}_{nml}$ their joint probability distribution. Then

$$\Pr\{X = a_i | Y = b_j \wedge Z = c_k\} = \frac{\Pr\{X = a_i \wedge Y = b_j \wedge Z = c_k\}}{\Pr\{Y = b_j \wedge Z = c_k\}}. \quad (1.28)$$

In the shorthand notation introduced earlier, this becomes

$$\theta_{\{a_i | b_j \wedge c_k\}} = \left[\frac{\theta_{ijk}}{\sum_{i'=1}^n \theta_{i'jk}}, i = 1, \dots, n \right] \in \mathbb{S}_n. \quad (1.29)$$

When there are three random variables, the “law of iterated conditioning” applies, namely:

$$\theta_{\{X | Y=b_j \wedge Z=c_k\}} = \theta_{\{\{X \wedge Y | Z=c_k\} | Y=b_j\}}. \quad (1.30)$$

In other words, in order to compute the conditional distribution of X given that $Y = b_j$ and $Z = c_k$, we can think of two distinct approaches. First, we can directly apply (1.28). Second, we can begin by computing the joint conditional distribution of $X \wedge Y$ given that $Z = c_k$, and then condition this distribution of $Y = b_j$. Both approaches give the same answer.

The proof of (1.30) is straightforward. To begin with, we have

$$\theta_{\{X \wedge Y | Z=c_k\}} = \left[\frac{\theta_{ijk}}{\sum_{i'=1}^n \sum_{j'=1}^m \theta_{i'j'k}}, i = 1, \dots, n, j = 1, \dots, m \right] \in \mathbb{S}_{nm}. \quad (1.31)$$

To make this formula less messy, let us define

$$\zeta_k := \sum_{i'=1}^n \sum_{j'=1}^m \theta_{i'j'k}.$$

Then

$$\theta_{\{X \wedge Y | Z=c_k\}} = \left[\frac{\theta_{ijk}}{\zeta_k}, i = 1, \dots, n, j = 1, \dots, m \right].$$

If we now condition this joint distribution of $X \wedge Y$ on $Y = b_j$, we get

$$\begin{aligned} \theta_{\{X \wedge Y | Z=c_k\} | Y=b_j} &= \left[\frac{\theta_{ijk}/\zeta_k}{\sum_{i'=1}^n \theta_{i'jk}/\zeta_k}, \mathbf{i} = 1, \dots, n \right] \\ &= \left[\frac{\theta_{ijk}}{\sum_{i'=1}^n \theta_{i'jk}}, i = 1, \dots, n \right], \end{aligned}$$

which is the same as (1.29).

The next notion introduced is conditional independence, which is very important in the case of hidden Markov processes, which are a central theme of this book.

Definition 1.17 Suppose X, Y, Z are random variables assuming values in finite sets $\mathbb{A} = \{a_1, \dots, a_m\}$, $\mathbb{B} = \{b_1, \dots, b_m\}$ and $\mathbb{C} = \{c_1, \dots, c_l\}$ respectively. Then we say that X, Y are **conditionally independent given Z** if, for all $c_k \in \mathbb{C}, b_j \in \mathbb{B}, a_i \in \mathbb{A}$, we have

$$\Pr\{X = a_i \wedge Y = b_j | Z = c_k\} = \Pr\{X = a_i | Z = c_k\} \cdot \Pr\{Y = b_j | Z = c_k\} \quad (1.32)$$

Example 1.8 Consider three random variables X, Y, Z , each assuming values in $\{0, 1\}$. Suppose the joint distribution of the three variables is given by

$$\begin{bmatrix} \theta_{000} & \theta_{001} \\ \theta_{010} & \theta_{011} \end{bmatrix} = \begin{bmatrix} 0.018 & 0.056 \\ 0.042 & 0.084 \end{bmatrix}, \begin{bmatrix} \theta_{100} & \theta_{101} \\ \theta_{110} & \theta_{111} \end{bmatrix} = \begin{bmatrix} 0.096 & 0.192 \\ 0.224 & 0.288 \end{bmatrix},$$

where θ_{ijk} denotes $\Pr\{X = i \wedge Y = j \wedge Z = k\}$. It is now shown that X and Y are conditionally independent given Z . This is achieved by verifying (1.32).

From the given data, we can compute the joint distributions of $X \wedge Z$, and of $Y \wedge Z$. This gives

$$\begin{array}{cc} X \wedge Z & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{bmatrix} \theta_{000} + \theta_{010} & \theta_{001} + \theta_{011} \\ \theta_{100} + \theta_{110} & \theta_{101} + \theta_{111} \end{bmatrix} \end{array} = \begin{array}{cc} X \wedge Z & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{bmatrix} 0.06 & 0.14 \\ 0.32 & 0.48 \end{bmatrix} \end{array}$$

Hence, if ψ denotes the joint distribution of X and Z , then

$$\psi_{\{X|Z=0\}} = \frac{1}{0.38} \begin{bmatrix} 0.06 \\ 0.32 \end{bmatrix}, \psi_{\{X|Z=1\}} = \frac{1}{0.62} \begin{bmatrix} 0.14 \\ 0.48 \end{bmatrix}.$$

Next, let η denote the joint distribution of Y and Z . Then an entirely similar computation yields that

$$\begin{array}{cc} Y \wedge Z & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{bmatrix} \theta_{000} + \theta_{100} & \theta_{001} + \theta_{101} \\ \theta_{010} + \theta_{110} & \theta_{011} + \theta_{111} \end{bmatrix} \end{array} = \begin{array}{cc} Y \wedge Z & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{bmatrix} 0.114 & 0.248 \\ 0.266 & 0.372 \end{bmatrix} \end{array}$$

Hence

$$\eta_{\{Y|Z=0\}} = \frac{1}{0.38} [0.114 \ 0.266] = [0.3 \ 0.7],$$

$$\boldsymbol{\eta}_{\{Y|Z=1\}} = \frac{1}{0.62} [0.248 \ 0.372] = [0.4 \ 0.6].$$

Next, let us compute the joint distribution of $X \wedge Y$ conditioned on Z . From either of the above computations, it is clear that the marginal distribution of Z is given by

$$\boldsymbol{\psi}_Z = \boldsymbol{\eta}_Z = [0.38 \ 0.62].$$

Therefore the joint distribution of $X \wedge Y$ conditioned on Z can be computed using (1.31). This gives

$$\begin{array}{cc} X \wedge Y | Z = 0 & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{bmatrix} 0.018/0.38 & 0.042/0.38 \\ 0.096/0.38 & 0.224/0.38 \end{bmatrix} \end{array} = \frac{1}{0.38} \begin{bmatrix} 0.06 \\ 0.32 \end{bmatrix} [0.3 \ 0.7]$$

$$= \boldsymbol{\psi}_{\{X|Z=0\}} \times \boldsymbol{\eta}_{\{Y|Z=0\}}.$$

Similarly,

$$\begin{array}{cc} X \wedge Y | Z = 1 & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{bmatrix} 0.056/0.62 & 0.084/0.62 \\ 0.192/0.62 & 0.288/0.62 \end{bmatrix} \end{array} = \frac{1}{0.62} \begin{bmatrix} 0.14 \\ 0.48 \end{bmatrix} [0.4 \ 0.6]$$

$$= \boldsymbol{\psi}_{\{X|Z=1\}} \times \boldsymbol{\eta}_{\{Y|Z=1\}}.$$

Hence X and Y are conditionally independent given Z .

Note that, if we are not “given Z ,” then X and Y are *not* independent. From earlier discussion, it follows that the joint distribution of X and Y is given by

$$\Pr\{X = a_i \wedge Y = b_j\} = \sum_{k=1}^l \Pr\{X = a_i \wedge Y = b_j | Z = c_k\} \cdot \Pr\{Z = c_k\}.$$

So if we were to write the joint distribution of X and Y in a matrix, then it would be

$$\begin{bmatrix} 0.018 & 0.042 \\ 0.096 & 0.224 \end{bmatrix} + \begin{bmatrix} 0.056 & 0.084 \\ 0.192 & 0.288 \end{bmatrix} = \begin{bmatrix} 0.074 & 0.126 \\ 0.288 & 0.512 \end{bmatrix},$$

where the columns correspond to the values of Y and the row corresponds to the values of X . Since this matrix does not have rank one, X and Y are not independent. The point is that a convex combination of rank one matrices need not be of rank one.

Once we have the notion of a conditional distribution, the notion of conditional expected value is natural. Suppose X, Y are random variables assuming values in $\mathbb{A}$ and $\mathbb{B}$ respectively, and suppose $f : \mathbb{A} \rightarrow \mathbb{R}$ is some real-valued function. Let ϕ denote the joint distribution of X and Y . Then the “unconditional” expected value of f is denoted by $E[f, P_{\phi, X}]$ or less clumsily by $E[f, \phi_X]$, and is defined as

$$E[f, \phi_X] = \sum_{i=1}^n f(a_i)(\phi_X)_i.$$

The “conditional” expected value of f is denoted by $E[f, P_{\phi, \{X|Y=b_j\}}]$ or less clumsily by $E[f, \phi_{\{X|Y=b_j\}}]$, and is defined as

$$E[f, \phi_{\{X|Y=b_j\}}] = \sum_{i=1}^n f(a_i) \phi_{\{a_i|b_j\}}.$$

We conclude this subsection by introducing another notion called the conditional expectation of a random variable. The dual usage of the adjective “conditional” is a source of endless confusion to students. The conditional *expected value* of a random variable (or a function of a random variable) is a *real number*, whereas the conditional *expectation* of a random variable is another *random variable*. Unfortunately, this dual usage is too firmly entrenched in the probability literature for the present author to deviate from it.

Suppose X, Y are random variables assuming values in finite sets $\mathbb{A} = \{a_1, \dots, a_n\}$, $\mathbb{B} = \{b_1, \dots, b_m\}$ respectively. Let $\phi \in \mathbb{S}_{nm}$ denote their joint distribution. Now suppose $h : \mathbb{A} \times \mathbb{B} \rightarrow \mathbb{R}$ is a function of both X and Y . Then we can think of $h(X, Y)$ as a real-valued random variable. Now one can ask: What is the best approximation of $h(X, Y)$ in terms of a real-valued function of X alone? In other words, we seek a function $f : \mathbb{A} \rightarrow \mathbb{R}$ such that f best approximates h . A natural error criterion is the “least-squares error,” namely,

$$J(f) = E[f - h, P_{\phi}] = \sum_{i=1}^n \sum_{j=1}^m (f_i - h_{ij})^2 \phi_{ij},$$

where we use the shorthand $f_i = f(a_i)$, $h_{ij} = h(a_i, b_j)$. The choice of f that minimizes J is easy to compute. Note that

$$\frac{\partial J}{\partial f_i} = 2 \sum_{j=1}^m (f_i - h_{ij}) \phi_{ij}.$$

Hence the optimal choice of f_i is obtained by setting this partial derivative to zero, that is,

$$f_i = \frac{\sum_{j=1}^m h_{ij} \phi_{ij}}{\sum_{j=1}^m \phi_{ij}} = \frac{\sum_{j=1}^m h_{ij} \phi_{ij}}{(\phi_X)_i} = \frac{E[h(X, Y), \phi]}{(\phi_X)_i}. \quad (1.33)$$

Hence if we define a function $f : \mathbb{A} \rightarrow \mathbb{R}$ by $f(a_i) = f_i$, then $f(X)$ is the best approximation to $h(X, Y)$ that depends on X alone. We formalize this idea through a definition.

Definition 1.18 Suppose X, Y are random variables assuming values in finite sets $\mathbb{A} = \{a_1, \dots, a_n\}$, $\mathbb{B} = \{b_1, \dots, b_m\}$ respectively. Let $\phi \in \mathbb{S}_{nm}$ denote their joint distribution. Suppose $h : \mathbb{A} \times \mathbb{B} \rightarrow \mathbb{R}$. Then the **conditional expectation of h with respect to X** is the function $f : \mathbb{A} \rightarrow \mathbb{R}$ defined by (1.33), and is denoted by $h|_{\mathbb{A}}$ or $h|X$.

In the above definition, if $(\phi_X)_i = 0$ for some index i , then the corresponding value f_i can be assigned arbitrarily. This is because, if $(\phi_X)_i = 0$

for some index i , then $\phi_{ij} = 0$ for all j . As a result, we can actually just drop the corresponding element a_i from the set $\mathbb{A}$ and carry on without affecting anything.

Lemma 1.19 *Suppose $h : \mathbb{A} \times \mathbb{B} \rightarrow \mathbb{R}_+$. Then $h|X : \mathbb{A} \rightarrow \mathbb{R}_+$. Suppose $h : \mathbb{A} \times \mathbb{B} \rightarrow [\alpha, \beta]$ for some finite numbers $\alpha < \beta$. Then $h|X : \mathbb{A} \rightarrow [\alpha, \beta]$.*

Proof. The first part of the lemma says that if the original function h assumes only nonnegative values, then so does its conditional expectation $h|X$. This fact is obvious from the definition (1.33). The second part follows readily upon observing that if $h : \mathbb{A} \times \mathbb{B} \rightarrow [\alpha, \beta]$, then both $h - \alpha$ and $\beta - h$ are nonnegative-valued functions. $\square$

A very useful property of the conditional expectation is given next.

Theorem 1.20 *Suppose X, Y are random variables assuming values in finite sets $\mathbb{A} = \{a_1, \dots, a_n\}$, $\mathbb{B} = \{b_1, \dots, b_m\}$ respectively. Let $\phi \in \mathbb{S}_{nm}$ denote their joint distribution. Suppose $h : \mathbb{A} \times \mathbb{B} \rightarrow \mathbb{R}$ and that $g : \mathbb{A} \rightarrow \mathbb{R}$. Then*

$$E[g(X)h(X, Y), \phi] = E[g \cdot h|X, \phi_{\mathbb{A}}]. \quad (1.34)$$

The point of the theorem is that $g(X)h(X, Y)$ is a random variable that depends on both X and Y , whereas $g \cdot h|X$ is a random variable that depends on X alone. The theorem states that both random variables have the same expected value (with respect to their corresponding probability distributions).

Proof. This follows from just writing out the expected value as a summation. We have

$$\begin{aligned} E[g(X)h(X, Y), \phi] &= \sum_{i=1}^n \sum_{j=1}^m g_i h_{ij} \phi_{ij} \\ &= \sum_{i=1}^n g_i \sum_{j=1}^m h_{ij} \phi_{ij} \\ &= \sum_{i=1}^n g_i (h_{\mathbb{A}})_i (\phi_{\mathbb{A}})_i \\ &= E[g \cdot h|X, \phi_{\mathbb{A}}]. \end{aligned}$$

This is the desired result. $\square$

1.2.3 Bayes' Rule

The next result, known as **Bayes' rule**, is widely used.

Lemma 1.21 *Suppose X and Y are random variables assuming values in finite sets $\mathbb{A}$ and $\mathbb{B}$ of cardinality n and m respectively. Then*

$$\Pr\{X = a_i | Y = b_j\} = \frac{\Pr\{Y = b_j | X = a_i\} \cdot \Pr\{X = a_i\}}{\Pr\{Y = b_j\}}. \quad (1.35)$$

Proof. An equivalent way of writing (1.32) is:

$$\Pr\{X = a_i|Y = b_j\} \cdot \Pr\{Y = b_j\} = \Pr\{Y = b_j|X = a_i\} \cdot \Pr\{X = a_i\}.$$

But this statement is clearly true, since each side is equal to $\Pr\{X = a_i \wedge Y = b_j\}$. $\square$

Example 1.9 A typical use of Bayes' rule is when we try to invert the hypothesis and conclusion, and assess the probability of the resulting statement. To illustrate, suppose there is a diagnostic test for HIV, which is accurate 98% of the time on HIV-positive patients and 99% of the time on HIV-negative patients. In other words, the probability that the test is positive when the patient has HIV is 0.98, while the probability that the test is negative when the patient does not have HIV is 0.99. We may therefore be lulled into thinking that we have a very good test at hand. But the question that really interests us is this: What is the probability that a patient who tests positive actually has HIV?

Let us introduce two random variables: X for a patient's actual condition, and Y for the outcome of a test. Thus X assumes values in the set $X = \{H, F\}$, where H denotes that the patient has HIV, while F denotes that the patient is free from HIV. Similarly, Y assumes values in the set $Y = \{P, N\}$, where P denotes that the test is positive, while N denotes that the test is negative. The data given thus far can be summarized as follows:

$$\Pr\{Y = P|X = H\} = 0.98, \Pr\{Y = N|X = F\} = 0.99. \quad (1.36)$$

But what we really want to know is the value of

$$\Pr\{X = H|Y = P\},$$

that is, the probability that the patient really has HIV when the test is positive.

To compute this quantity, suppose the fraction of the population that has HIV is 1%. Thus the marginal probability distribution of X is

$$[\Pr\{X = H\} \quad \Pr\{X = F\}] = [0.01 \quad 0.99].$$

With this information and (1.36), we can compute the joint distribution of the variables X and Y . We get

$$\begin{bmatrix} \phi_{X=H \wedge Y=P} & \phi_{X=H \wedge Y=N} \\ \phi_{X=F \wedge Y=P} & \phi_{X=F \wedge Y=N} \end{bmatrix} = \begin{bmatrix} 0.0098 & 0.0002 \\ 0.0099 & 0.9801 \end{bmatrix}.$$

So by adding up the two columns, we get

$$[\Pr\{Y = P\} \quad \Pr\{Y = N\}] = [0.0197 \quad 0.9803].$$

Hence, by Bayes' rule, we can compute that

$$\Pr\{X = H|Y = P\} = \frac{0.0098}{0.0197} \approx 0.5.$$

So actually the diagnostic is quite unreliable, because the likelihood of a patient who tests positive *not* having HIV is just about equal to the likelihood of a patient who tests positive actually having HIV.

This apparent paradox is easily explained: For the sake of simplicity, assume that the test is equally accurate both for patients actually having HIV and for patients not having HIV. Let β denote the inaccuracy of the test. Thus

$$\Pr\{Y = P|X = H\} = \Pr\{Y = N|X = F\} = 1 - \beta.$$

Let α denote the fraction of the population that actually has HIV. We can carry through all of the above computations in symbolic form and obtain

$$\begin{bmatrix} \phi_{X=H \wedge Y=P} & \phi_{X=H \wedge Y=N} \\ \phi_{X=F \wedge Y=P} & \phi_{X=F \wedge Y=N} \end{bmatrix} = \begin{bmatrix} \alpha(1 - \beta) & \alpha\beta \\ (1 - \alpha)\beta & (1 - \alpha)(1 - \beta) \end{bmatrix}.$$

So

$$\Pr\{X = H|Y = P\} = \frac{\alpha(1 - \beta)}{\alpha(1 - \beta) + (1 - \alpha)\beta}.$$

If, as is reasonable, both α and β are small, we can approximate both $1 - \alpha$ and $1 - \beta$ by 1, which leads to

$$\Pr\{X = H|Y = P\} \approx \frac{\alpha}{\alpha + \beta}.$$

So, unless $\beta \ll \alpha$, we get a test that is pretty useless. On the other hand, if $\beta \ll \alpha$, then $\Pr\{X = H|Y = P\}$ is very close to one and we have an excellent diagnostic test. The point to note is that the error of the diagnostic test must be small, not in comparison with 1, but with the likelihood of occurrence of the condition that we are trying to detect.

1.2.4 MAP and Maximum Likelihood Estimates

In the previous subsections, we have discussed the issue of computing the probability distribution of one random variable, given an observation of another random variable. Now let us make the question a little more specific, and ask: What is the *most likely value* of one random variable, given an observation of another random variable? It is shown below that there are two distinct ways of formalizing this notion, and each is reasonable in its own way.

Definition 1.22 Suppose X and Y are random variables assuming values in finite sets $\mathbb{A} = \{a_1, \dots, a_n\}$ and $\mathbb{B} = \{b_1, \dots, b_m\}$ respectively. Let ϕ denote their joint distribution. Then the **maximum a posteriori (MAP) estimate** of X given an observation $Y = b_j$ is the a_{i^*} such that

$$\phi_{\{a_{i^*}|b_j\}} = \max_i \phi_{\{a_i|b_j\}}. \quad (1.37)$$

Thus the MAP estimate of X given an observation $Y = b_j$ is the most likely value of X using the conditional distribution $\phi_{\{X|Y=b_j\}}$. Since

$$\phi_{\{a_i|b_j\}} = \frac{\phi_{ij}}{(\phi_Y)_j},$$

and the denominator is independent of i , we can see that

$$i^* = \underset{i}{\operatorname{argmin}} \phi_{ij}.$$

So computing the MAP estimate is very easy. Given an observation $Y = b_j$, we simply scan down the j -th column of the joint distribution matrix, and pick the row i where the element ϕ_{ij} is the largest. (If there is a tie, we can use any sensible tie-breaking rule.)

The next definition gives an alternate way of defining the “most likely” value of X .

Definition 1.23 Suppose X and Y are random variables assuming values in finite sets $\mathbb{A} = \{a_1, \dots, a_n\}$ and $Y = \{b_1, \dots, b_m\}$ respectively. Let ϕ denote their joint distribution. Then the **maximum likelihood estimate (MLE)** of X given an observation $Y = b_j$ is defined as the index i^* such that $\Pr\{Y = b_j | X = a_i\}$ is maximized when $i = i^*$.

Thus the MLE of X given the observation $Y = b_j$ is the choice of a_i that makes the observed value the most likely one.

The choice between MAP and MLE is essentially dictated by whether we believe that X “causes” Y , or vice versa. The joint distribution ϕ is strictly neutral, and does not at all address the issue of what causes what. If we believe that Y causes X , then we should believe that, following the observation $Y = b_j$, the probability distribution of X has shifted from the marginal distribution ϕ_X to the conditional distribution $\phi_{\{X|Y=b_j\}}$. Thus MAP is the most logical way to estimate X . If on the other hand we believe that X causes Y , we should choose the MLE of X , because that estimate makes the observation most likely.

Example 1.10 To show that MAP and MLE can lead to diametrically opposite conclusions, consider the case where $n = m = 2$ and the joint distribution of X, Y is given by

$$\phi = \begin{bmatrix} 0.1 & 0.2 \\ 0.4 & 0.3 \end{bmatrix},$$

where the rows correspond to the value of X and the columns to the values of Y . Suppose we observe $Y = b_2$. Then, by examining the second column of ϕ , we see that the MAP estimate of X is a_2 , because $\phi_{22} > \phi_{12}$. On the other hand, to compute the MLE of X , we compute

$$\phi_{\{Y|X=x_1\}} = [1/3 \ 2/3], \quad \phi_{\{Y|X=x_2\}} = [4/7 \ 3/7].$$

Since $2/3 > 3/7$, b_2 is the most likely value of Y if $X = a_1$. So the MLE of X given the observation $Y = y_2$ is a_1 .

Problem 1.10 Prove (1.22).

Problem 1.11 Show that if X, Y are independent real-valued random variables, then their correlation coefficient is zero.

Problem 1.12 Suppose the joint distribution of two random variables X and Y , each of them assuming one of the five values $\{1, 2, 3, 4, 5\}$, is as shown in the table below.

$$\Phi = \begin{bmatrix} X \setminus Y & 1 & 2 & 3 & 4 & 5 \\ 1 & 0.0800 & 0.0260 & 0.0280 & 0.0320 & 0.0340 \\ 2 & 0.0280 & 0.0900 & 0.0300 & 0.0270 & 0.0250 \\ 3 & 0.0260 & 0.0200 & 0.0800 & 0.0340 & 0.0400 \\ 4 & 0.0340 & 0.0300 & 0.0290 & 0.0800 & 0.0270 \\ 5 & 0.0320 & 0.0340 & 0.0330 & 0.0270 & 0.0740 \end{bmatrix}.$$

Compute the following:

1. The five conditional probability distributions $\phi_{X|Y=n\}$, for $n = 1, \dots, 5$.
2. The five conditional probability distributions $\phi_{Y|X=n\}$, for $n = 1, \dots, 5$.
3. The five conditional expected values $E[X|Y = n]$ for $n = 1, \dots, 5$.
4. The five conditional expected values $E[Y|X = n]$ for $n = 1, \dots, 5$.
5. The MAP estimates of X given that $Y = n$ for $n = 1, \dots, 5$.
6. The MAP estimates of Y given that $X = n$ for $n = 1, \dots, 5$.
7. The correlation coefficient $C(X, Y)$.

Problem 1.13 Prove (1.25).

Problem 1.14 Show that (1.26) and (1.27) are equivalent conditions.

Problem 1.15 Prove Lemma 1.16.

Problem 1.16 Suppose, as in (1.33), that $h : \mathbb{A} \times \mathbb{B} \rightarrow \mathbb{R}$. For each $a_i \in \mathbb{A}$, define the function $h_{i\cdot} : \mathbb{B} \rightarrow \mathbb{R}$ by

$$h_{i\cdot} = h_{ij}.$$

Show that the expression (1.33) for the conditional expectation of h with respect to X can be defined as

$$(h_X)_i = E[h_{i\cdot}, \phi_{|X=a_i}].$$

1.3 RANDOM VARIABLES ASSUMING INFINITELY MANY VALUES

1.3.1 Some Preliminaries

Until now we have steadfastly restricted ourselves to random variables that assume values in a *finite* set. However, there are situations in which it is desirable to relax this assumption, and examine situations in which the range of the random variable under study is *infinite*. Within this, we make a further distinction between two situations: where the range is a *countable* set and where the range is an *uncountable* set. Recall that a set is said to be **countable** if it can be placed in one-to-one correspondence with the set of natural numbers $\mathbb{N} = \{1, 2, \dots\}$.³ For example, the set of integers and the set of rational numbers are both countable sets. Next, suppose $\mathbb{M}$ is a finite set, such as $\{H, T\}$, the set of possible outcomes of a coin toss experiment, or $\{A, C, G, T\}$, the set of nucleotides. Let $\mathbb{M}^*$ denote the set of all *finite* sequences taking values in $\mathbb{M}$. Thus $\mathbb{M}^*$ consists of all sequences $\{u_1, \dots, u_n\}$ where $u_i \in \mathbb{M}$ for all i . Then $\mathbb{M}^*$ is countable. But uncountably infinite sets are also relevant, such as with real-valued random variables. A less familiar example is that, if $\mathbb{M}$ is a finite set, then the set of *all* sequences (not just finite sequences) taking values in $\mathbb{M}$ is an uncountably infinite set.

It turns out that the method adopted thus far to define probabilities over finite sets, namely just to assign nonnegative “weights” to each element in such a way that the weights add up to one, works perfectly well on countable sets. However, the approach breaks down when the range of the random variable is an uncountably infinite set. The great Russian mathematician A. N. Kolmogorov introduced the axiomatic foundations of probability theory precisely to cope with this situation; see [81]. Though this theory is very beautiful and comprehensive, for the most part in the present book we will not be using this more advanced theory.

We first discuss the case of random variables assuming values in a countable set $\mathcal{X} = \{x_i, i \in \mathbb{N}\}$, where $\mathbb{N}$ denotes the set of natural numbers. Let $\mu_i \geq 0$ be chosen such that $\sum_{i=1}^{\infty} \mu_i = 1$. Then for every subset $A \subseteq \mathcal{X}$, we can define the corresponding probability measure P_μ in analogy with (1.2), namely

$$P_\mu(A) := \sum_{i=1}^{\infty} I_A(x_i) \mu_i.$$

We can think of $P_\mu(A)$ as the probability $\Pr\{X \in A\}$ that the random variable X belongs to the set A . With this definition, the “axiomatic” properties described just after Theorem 1.3 continue hold, namely:

1. $0 \leq P_\mu(A) \leq 1$ for all subsets $A \subseteq \mathcal{X}$.
2. $P_\mu(\emptyset) = 0$ and $P_\mu(\mathcal{X}) = 1$.

³Some authors also include 0 in $\mathbb{N}$.

3. If $\{A_i\}_{i \geq 1}$ is a countable collection of pairwise disjoint subsets of X , then

$$P_\mu \left(\bigcup_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} P_\mu(A_i).$$

If X is a *real-valued* random variable assuming values in the countable set $\mathcal{X} = \{x_i, i \in \mathbb{N}\} \subseteq \mathbb{R}$, with the associated probability measure P_μ defined above, then we define the mean, variance, and standard deviation of X just as we did earlier for finite-valued random variables, namely

$$E[X, P_\mu] := \sum_{i=1}^{\infty} x_i \mu_i,$$

$$V[X, P_\mu] = E[(X - E[X, P_\mu])^2, P_\mu],$$

$$\sigma(X, P_\mu) = \sqrt{V[X, P_\mu]}.$$

The major potential difficulty is that, because we are dealing with *infinite sums*, the above summations are not guaranteed to converge. Example 1.11 gives one such random variable. However, Schwarz's inequality (of which we got only a glimpse earlier) guarantees that, if X has finite variance, then it also has finite mean.

Example 1.11 This example is sometimes referred to as the “St. Petersburg paradox.” Suppose a gambler visits a casino where he plays a coin-tossing game. At each step, both the gambler and the casino put up equal stakes, after which the gambler predicts the outcome of the coin toss. If he calls correctly, he gets the entire stake, whereas if he calls incorrectly, the casino gets the entire stake. The game is fair, with the coin turning up heads or tails with equal probability. Moreover, each coin toss is independent of all previous coin tosses. To simplify the notation, let us suppose that the gambler always calls heads. In view of the independence and the fairness assumptions, this strategy has just as good a chance of winning as any other.

Now the following strategy is “guaranteed” to fetch a positive payoff to the gambler: At each step, he merely doubles his stake. Thus, at the “zeroth” step, he bets \$1. If he wins, he quits and goes home. If he loses, he bets \$2 at step 1. If he loses again, he bets \$4 at the next step, and so on. The game reaches n steps only if the gambler has lost all previous $n - 1$ times, meaning that his accumulated losses are $\$1 + 2 + \dots + 2^{n-1} = 2^n - 1$. At the n -th step, he bets 2^n . If he wins, his cumulative winning amount is precisely \$1, the initial bet.

One feels that there is something strange about this strategy; indeed the difficulty is that the random variable in this case is “heavy-tailed” and does not have a finite expectation. We can see that the game has a countable set of possible outcomes, namely $H, TH, T^2H, \dots, T^nH, \dots$, where T^nH denotes a sequence of n tails followed by a head. The probability of T^nH is obviously

$2^{-(n+1)}$ because the coin is fair. In this case, the accumulated losses before the n -th step are $2^n - 1$. Thus, in order to bet 2^n at the next step, the gambler must have an initial sum of $2^n - 1 + 2^n = 2^{n+1} - 1$. Therefore, the amount of money that the player must have to begin with, call it X , is an integer-valued random variable, and equals $2^{n+1} - 1$ with probability $2^{-(n+1)}$. If we try to compute the expected value of this random variable, we see that

$$\sum_{n=0}^{\infty} (2^{n+1} - 1) \cdot 2^{-(n+1)} = \sum_{n=0}^{\infty} [1 - 2^{-(n+1)}].$$

Now the second summation converges nicely to -1 . Unfortunately, the first summation blows up. Hence, unless one has an infinite amount of money to begin with, the above “strategy” will not work.

Next we give a very cursory discussion of random variables whose range may consist of an *uncountable* subset of $\mathbb{R}$. With each real-valued random variable X we associate a so-called **cumulative distribution function (cdf)** P_X , defined as follows:

$$P_X(a) = \Pr\{X \leq a\}, \quad \forall a \in \mathbb{R}.$$

The cdf is monotonically nondecreasing, as is obvious from the definition; thus

$$a \leq b \Rightarrow P_X(a) \leq P_X(b).$$

The cdf also has a property known as “cadlag,” which is an abbreviation of the French expression “continué à droite, limité à gauche.” In English this means “continuous from the right, and limit exists from the left.” In other words, the function P_X has the property that, for each real number a ,

$$\lim_{x \rightarrow a^+} P_X(x) = P_X(a),$$

while $\lim_{x \rightarrow a^-} P_X(x)$ exists, but may or may not equal $P_X(a)$. Due to the monotonicity of the cdf, it is clear that

$$\lim_{x \rightarrow a^-} P_X(x) \leq P_X(a).$$

If the above holds with equality, then P_X is continuous at a . Otherwise it has a positive jump equal to the difference between $P_X(a)$ and the limit on the left side. For every $a \in \mathbb{R}$, it is the case that

$$\Pr\{X = a\} = P_X(a) - \lim_{x \rightarrow a^-} P_X(x).$$

So if P_X is continuous at a , then the probability that X *exactly* equals a is zero. However, if P_X has a jump at a , then the magnitude of the jump is the probability that X exactly equals a .

In general the function P_X need not be differentiable, or even continuous. However, for the purposes of the present discussion, it is sufficient to consider the case where P_X is continuously differentiable everywhere, except for a countable set of points $\{x_i\}_{i=1}^{\infty}$, where the function has a jump. Thus

$$\lim_{x \rightarrow x_i^-} P_X(x) < P_X(x_i),$$

but $P_X(\cdot)$ is continuously differentiable at all $x \neq x_i$. In such a case, the difference

$$P_X(x_i) - \lim_{x \rightarrow x_i^-} P_X(x) =: \mu_i$$

can be interpreted as the (nonzero) probability that the random variable X *exactly equals* x_i , as discussed above. For all other values of x , it is interpreted that the probability of the random variable X *exactly* equalling x is zero. Moreover, if $a < b$, then the probability of the random variable X lying in the interval $(a, b]$ is taken as $P_X(b) - P_X(a)$.

To define the expected value of the random variable X , we adapt the earlier formulation to the present situation. To simplify notation, let $P(\cdot)$ denote the cdf of X . Then we define

$$E[X, P] = \int_{-\infty}^{\infty} xP(dx),$$

where the integral is a so-called Riemann-Stieltjes integral. If P is continuously differentiable over some interval $[a, b]$, we define

$$\int_a^b f(x)P(dx) = \int_a^b f(x) \frac{dP}{dx} dx,$$

and add the term $f(x_i)\mu_i$ whenever the interval $[a, b]$ contains one of the points x_i where P_X has a jump discontinuity. As before, the existence of the expected value is not guaranteed.

1.3.2 Markov and Chebycheff Inequalities

In this subsection, we introduce two very useful inequalities known as “Markov’s inequality” and “Chebycheff’s inequality.” We use the formalism of a random variable assuming real values in order to state the result. Since the set of real numbers is uncountably infinite, the earlier approach of assigning a weight to each possible outcome does not work, and we need to adopt a different approach. What follows is a very superficial introduction to the subject, and a reader interested in a serious discussion of the subject is referred to any of the classic texts on the subject, such as [23, 19] for example.

Theorem 1.24 (Markov’s Inequality) *Suppose X is a real-valued random variable with the property that $|X|$ has finite expected value. Then, for every real number a , we have*

$$\Pr\{|X| > a\} \leq \frac{E[|X|, P]}{a}. \quad (1.38)$$

Proof. By definition, we have

$$\begin{aligned} E[|X|, P] &= \int_{|x| \leq a} |x|P(dx) + \int_{|x| > a} |x|P(dx) \\ &\geq \int_{|x| > a} |x|P(dx) \\ &\geq \int_{|x| > a} aP(dx) \\ &= a \Pr\{|X| > a\}. \end{aligned}$$

The desired inequality follows by dividing both sides by a . $\square$

Corollary 1.25 *Suppose X is a nonnegative-valued random variable with finite expected value. Then, for every real number a , we have*

$$\Pr\{X > a\} \leq \frac{E[X, P]}{a}. \quad (1.39)$$

The proof is entirely analogous to that of Theorem 1.24.

Markov's inequality in the above form is not particularly useful. However, we get a more useful version if we examine a function of X .

Corollary 1.26 *Suppose X is a real-valued random variable. Then*

$$\Pr\{X > \epsilon\} \leq \exp(-\gamma\epsilon)E[\exp(\gamma X), P], \quad \forall \gamma \geq 0 \quad \forall \epsilon > 0, \quad (1.40)$$

provided only that $\exp(\gamma X)$ has finite expected value.

Proof. Note that, whenever $\gamma \geq 0$, the function $x \mapsto \exp(\gamma x)$ is nonnegative-valued and nondecreasing. Hence, for every $\epsilon > 0$, we have

$$X > \epsilon \Leftrightarrow \exp(\gamma X) > \exp(\gamma\epsilon).$$

Now apply (1.39) with X replaced by $\exp(\gamma X)$ and a replaced by $\exp(\gamma\epsilon)$. $\square$

There is a variant of Markov's inequality for random variables that have not only finite expected value but also finite variance. As before, we define the variance of X as

$$V(X) := E[(X - E(X))^2],$$

assuming it exists of course, and the standard deviation $\sigma(X)$ as $(V(X))^{1/2}$. With this notation, we now state the next result. Recall that if a random variable has finite variance, it automatically has a finite expected value.

Theorem 1.27 (Chebycheff's Inequality) *Suppose X is a real-valued random variable with finite variance $\sigma^2(X)$ and finite expected value $E(X)$. Then for each $\epsilon > 0$, we have*

$$\Pr\{|X - E(X)| > \epsilon\} \leq \frac{\sigma^2(X)}{\epsilon^2}. \quad (1.41)$$

Proof. We reason as follows:

$$\begin{aligned}\Pr\{|X - E(X)| > \epsilon\} &= \Pr\{(X - E(X))^2 > \epsilon^2\} \\ &\leq \frac{E[(X - E(X))^2]}{\epsilon^2} \text{ by (1.39)} \\ &= \frac{\sigma^2(X)}{\epsilon^2}.\end{aligned}$$

□

Chebycheff's inequality can be used to derive a quick, but rather conservative, estimate of the rate at which empirical estimates of the mean, derived from independent samples, converge to the true value.

Theorem 1.28 *Suppose X is a real-valued random variable with finite variance $\sigma^2(X)$ and finite expected value $E(X)$. Let $X_1, \dots, X_l$ be independent copies of X , and define the l -fold average*

$$A_l := \frac{1}{l} \sum_{i=1}^l X_i. \quad (1.42)$$

Then

$$\Pr\{|A_l - E(X)| > \epsilon\} \leq \frac{\sigma^2(X)}{l\epsilon^2}. \quad (1.43)$$

Proof. Note that, by the linearity of the expected value, it follows that

$$E(A_l) = E(X), \quad \forall l.$$

Also, the independence of the X_i leads to the observation that if we define S_l as $S_l = \sum_{i=1}^l X_i$, then $V(S_l) = l\sigma^2(X)$. In turn this implies that $V(A_l) = \sigma^2(X)/l$. Now apply Theorem 1.27, and specifically (1.41), to A_l . This readily leads to the desired inequality (1.43). □

Chebycheff's inequality can be used to estimate the probability that an empirical mean differs from its true mean. Specifically, suppose X is a real-valued random variable with finite variance $\sigma^2(X)$, and it is desired to estimate its expected value $E(X)$. One way to do this is to generate independent samples $x_1, \dots, x_l$ of X , and define

$$\hat{E}(\mathbf{x}_1^l) := \frac{1}{l} \sum_{i=1}^l x_i, \quad (1.44)$$

where $\mathbf{x}_1^l := (x_1, \dots, x_l) \in \mathbb{R}^n$. Then $\hat{E}(\mathbf{x}_1^l)$ is called the **empirical mean** of X based on the samples $x_1, \dots, x_l =: \mathbf{x}_1^l$. At this point the reader might wonder what the difference is between the average A_l and the empirical mean $\hat{E}(\mathbf{x}_1^l)$. Think of A_l as a random variable and $\hat{E}(\mathbf{x}_1^l)$ as a realization of the random variable A_l . Both notations prove useful in different situations, so both are introduced. Now Chebycheff's inequality tells us that

$$\Pr\{|\hat{E}(\mathbf{x}_1^l) - E(X)| > \epsilon\} \leq \frac{\sigma^2(X)}{l\epsilon^2}. \quad (1.45)$$

Therefore it follows that

$$\lim_{l \rightarrow \infty} \Pr\{|\hat{E}(\mathbf{x}_1^l) - E(X)| > \epsilon\} = 0, \quad \forall \epsilon > 0.$$

Note that we can also express this as

$$\lim_{l \rightarrow \infty} \Pr\{|A_l - E(X)| > \epsilon\} = 0, \quad \forall \epsilon > 0.$$

This says that the sequence of random variables $\{A_l\}$ “converges in probability” to the true mean $E(X)$ as $l \rightarrow \infty$.

1.3.3 Hoeffding’s Inequality

In this subsection we present a very powerful result known as Hoeffding’s inequality, which is applicable only to *bounded* real-valued random variables. The advantage of Hoeffding’s inequality over Chebycheff’s inequality is that the estimates converge to zero much more quickly.

Theorem 1.29 (Hoeffding’s Inequality) *Suppose $Y_1, \dots, Y_l$ are independent random variables, where Y_i assumes values in the bounded interval $[a_i, b_i]$. Then for every $\epsilon > 0$, we have*

$$\Pr\left\{\sum_{i=1}^l [Y_i - E(Y_i)] > \epsilon\right\} \leq \exp\left[-2\epsilon^2 / \sum_{i=1}^l (b_i - a_i)^2\right], \quad (1.46)$$

where $E(Y_i)$ denotes the expected value of Y_i .

The proof of Hoeffding’s inequality uses the following auxiliary lemma.

Lemma 1.30 *Suppose X is a zero-mean random variable assuming values in the interval $[a, b]$. Then for any $s > 0$, we have*

$$E[\exp(sX)] \leq \exp(s^2(b-a)^2/8).$$

Proof. (Of Lemma 1.30): We are forced to invoke Theorem 2.6, which has not yet been proven. Since the exponential is a convex function, the value of e^{sx} is bounded by the corresponding convex combination of its extreme values; that is,

$$\exp(sx) \leq \frac{x-a}{b-a} e^{sb} + \frac{b-x}{b-a} e^{sa}, \quad \forall x \in [a, b].$$

Now take the expected values of both sides, and use the fact that $E(X) = 0$. This gives

$$\begin{aligned} E[\exp(sX)] &\leq \frac{b}{b-a} e^{sa} - \frac{a}{b-a} e^{sb} \\ &= (1-p + pe^{s(b-a)})e^{-ps(b-a)} \\ &=: \exp(\phi(u)), \end{aligned}$$

where

$$p := -a/(b-a), u := s(b-a), \phi(u) := -pu + \ln(1-p + pe^u).$$

Clearly $\phi(u) = 0$. Moreover, a routine calculation shows that

$$\phi'(u) = -p + \frac{p}{p + (1-p)e^{-u}},$$

whence $\phi'(u) = 0$ as well. Moreover,

$$\phi''(u) = \frac{p(1-p)e^{-u}}{[p + (1-p)e^{-u}]^2} \leq 0.25 \quad \forall u > 0.$$

Hence by Taylor's theorem, there exists a $\theta \in [0, u]$ such that

$$\phi(u) = \frac{\phi''(\theta)u^2}{2} \leq \frac{u^2}{8} = \frac{s^2(b-a)^2}{8}.$$

This completes the proof. $\square$

Proof. (Of Theorem 1.29): For any nonnegative random variable, we have from Corollary 1.26 that

$$\Pr\{X > \epsilon\} \leq e^{-s\epsilon} E[\exp(sX)].$$

Now apply this inequality to the random variable

$$Z_l := \sum_{i=1}^l (Y_i - E(Y_i)),$$

which has zero mean by the linearity of the expected value. Thus

$$\begin{aligned} \Pr\{Z_l > \epsilon\} &\leq e^{-s\epsilon} E \left[\exp \left(s \sum_{i=1}^l (Y_i - E(Y_i)) \right) \right] \\ &= e^{-s\epsilon} \prod_{i=1}^l E[\exp(s(Y_i - E(Y_i)))] \text{ by independence} \\ &\leq e^{-s\epsilon} \prod_{i=1}^l \exp[s^2(b_i - a_i)^2/8] \text{ by Lemma 1.30} \\ &= \exp \left[-s\epsilon + s^2 \sum_{i=1}^l \frac{(b_i - a_i)^2}{8} \right] \\ &= \exp \left[\frac{-2\epsilon^2}{\sum_{i=1}^l (b_i - a_i)^2} \right], \end{aligned} \tag{1.47}$$

where the last step follows by choosing

$$s = \frac{4\epsilon}{\sum_{i=1}^l (b_i - a_i)^2}.$$

This completes the proof. $\square$

A useful (and widely used) “corollary” of Hoeffding’s inequality is obtained when we take repeated and independent measurements of *the same* random variable. Because of its importance, we state the “corollary” as a theorem.

Theorem 1.31 (Hoeffding’s Inequality for Empirical Means) Suppose X is a random variable assuming values in a bounded interval $[a, b]$, and that $x_1, \dots, x_l$ are independent samples of X . Define the empirical mean $\hat{E}(\mathbf{x}_1^l)$ as in (1.44). Then for each $\epsilon > 0$, we have

$$\Pr\{\hat{E}(\mathbf{x}_1^l) - E(X) > \epsilon\} \leq \exp[-2l\epsilon^2/(b-a)^2], \quad (1.48)$$

$$\Pr\{\hat{E}(\mathbf{x}_1^l) - E(X) < -\epsilon\} \leq \exp[-2l\epsilon^2/(b-a)^2], \quad (1.49)$$

$$\Pr\{|\hat{E}(\mathbf{x}_1^l) - E(X)| > \epsilon\} \leq 2 \exp[-2l\epsilon^2/(b-a)^2]. \quad (1.50)$$

Proof. To prove (1.48), apply (1.46) with ϵ replaced by $l\epsilon$, and $a_i = a, b_i = b$ for all i . To prove (1.49), apply (1.48) with Y replaced by $-Y$. Finally (1.50) is a direct consequence of (1.48) and (1.49). $\square$

Comparing (1.45) and (1.50) brings out clearly the superiority of Hoeffding’s inequality over Chebycheff’s inequality. In (1.45) the right side is $O(1/l)$ and thus decays very slowly with respect to the number of samples. In contrast, the right side of (1.50) decays as $\exp(-2l\epsilon^2)$.

Hoeffding’s inequality was proved in 1963; see [65]. Since then various researchers had attempted to improve the bound, but could not succeed in doing so. And it is no wonder. In 1990, Massart [95] proved that Hoeffding’s inequality is, in a very precise sense, the “best possible” inequality.

Note that Hoeffding’s inequality is applicable to real-valued random variables. So how can it be applied to random variables that assume values in some discrete set that has no obvious interpretation as a subset of the real numbers (e.g., the set of nucleotides)? The next result gives the answer.

Theorem 1.32 Suppose X is a random variable that assumes values in a finite set $\mathbb{A} = \{a_1, \dots, a_n\}$, and let $\phi \in \mathbb{S}_n$ denote the “true but unknown” probability distribution of X . Suppose $x_1, \dots, x_l$ are independent samples of X . For each $j \in \{1, \dots, n\}$, define the empirical frequency

$$\hat{\phi}_{l,j} := \frac{1}{l} \sum_{i=1}^l I_{\{x_i = a_j\}},$$

and the vector $\hat{\phi}_l \in \mathbb{S}_n$ as

$$\hat{\phi}_l := [\hat{\phi}_{l,j}, j = 1, \dots, n].$$

Then

$$\Pr\{\|\hat{\phi}_l - \phi\|_\infty > \epsilon\} \leq 2n \exp(-2l\epsilon^2). \quad (1.51)$$

Equivalently,

$$\Pr\{|\hat{\phi}_{l,j} - \phi_j| \leq \epsilon \forall j\} \geq 1 - 2n \exp(-2l\epsilon^2). \quad (1.52)$$

Proof. For each index $j \in \{1, \dots, n\}$, we can define an associated binary-valued random variable Y_j as follows: $Y_j = 1$ if $X = a_j$, and $Y_j = 0$

otherwise. With this association, it is clear that the expected value of Y_j is precisely $\Pr\{X = a_j\} =: \phi_j$. Now define associated subsets $S_{l,j}(\epsilon)$ as follows:

$$S_{l,j}(\epsilon) := \{\mathbf{x} \in \mathbb{A}^l : |\hat{\phi}_{l,j} - \phi_j| > \epsilon\}.$$

Then it is clear that

$$\{\mathbf{x} \in \mathbb{A}^l : \|\hat{\phi}_l - \phi\|_\infty > \epsilon\} = \bigcup_{j=1}^n S_{l,j}(\epsilon).$$

Therefore

$$\Pr\{\|\hat{\phi}_l - \phi\|_\infty > \epsilon\} \leq \sum_{j=1}^n P_\phi^l(S_{l,j}(\epsilon)).$$

However, by (1.50), each of the summands on the right side is bounded by $2 \exp(-2l\epsilon^2)$. $\square$

Since n , the cardinality of the set $\mathbb{A}$, appears explicitly on the right side of the above equation, this approach is not useful for infinite sets, and alternate approaches need to be devised. However, for random variables X assuming values in a finite set of cardinality n , (1.51) states that, after l independent trials, we can state with confidence $1 - 2n \exp(-2l\epsilon^2)$ that *every one of the n estimates $\hat{\phi}_{j,l}$ is within ϵ of its true value ϕ_j* .

If X assumes just two values, say H and T to denote “heads” and “tails,” then it is possible to get rid of the factor 2 on the right side of (1.52). This is because, if the empirical estimate $\hat{p}_H$ is within ϵ of the true value p_H , then so is $\hat{p}_T$. However, this argument does not work when there are more than two possible outcomes.

1.3.4 Monte Carlo Simulation

In this subsection we give a brief introduction to Monte Carlo simulation, which is one of the most widely used randomized methods. There is no universal agreement on what constitutes the (or even a) Monte Carlo method, but the method given below is certainly often referred to as such. Also, while the original applications of the method were to problems in atomic physics, the application given here is to the computation (or more precisely, the estimation) of complex integrals.

Suppose $\mathbf{X}$ is a random variable assuming values in $\mathbb{R}^d$, $d \geq 1$. For such a random variable, the corresponding d -dimensional probability distribution function is defined in a manner analogous to the case of real-valued random variables. For each d -tuple $(a_1, \dots, a_d) =: \mathbf{a}$, we define

$$P_{\mathbf{X}}(\mathbf{a}) = \Pr\{X_1 \leq a_1 \wedge \dots \wedge X_d \leq a_d\}.$$

The reader is reminded that the wedge symbol $\wedge$ denotes “and.” To make things simple, it is assumed that the distribution function is jointly differentiable with respect to all of its arguments, with the derivative denoted by

$p : \mathbb{R}^d \rightarrow \mathbb{R}_+$, which is called the **probability density function**. Therefore, for all $\mathbf{a} \in \mathbb{R}^d$, we have

$$\Pr\{X_1 \leq a_1 \wedge \dots \wedge X_d \leq a_d\} = \int_{-\infty}^{a_1} \dots \int_{-\infty}^{a_d} p(x_1, \dots, x_d) dx_d \dots dx_1.$$

Now suppose $f : \mathbb{R}^d \rightarrow [a, b]$ is a continuous function. The expected value of $f(\mathbf{X})$ can be defined as

$$E[f(\mathbf{X})] = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, \dots, x_d) p(x_1, \dots, x_d) dx_d \dots dx_1,$$

where the existence of the integral is guaranteed by the fact that f assumes values in a bounded interval. Now, depending on the form of the function f and/or the density p , it might be difficult or impossible to compute the above integral exactly. So one possibility is to use “gridding,” especially when $\mathbf{X}$ assumes values only in some hypercube in $\mathbb{R}^d$. Specifically, suppose

$$p(x_1, \dots, x_d) = 0 \text{ if } (x_1, \dots, x_d) \notin \prod_{i=1}^d [\alpha_i, \beta_i].$$

Thus the density is zero outside the hypercube $\prod_{i=1}^d [\alpha_i, \beta_i]$. The gridding technique consists of choosing some integer k , dividing each interval $[\alpha_i, \beta_i]$ into k equal intervals by choosing $k + 1$ grid points, such as

$$c_i^l = \alpha_i + l(\beta_i - \alpha_i)/k, l = 0, \dots, k.$$

Then the set of points

$$\mathcal{G} = (c_1^{l_1}, \dots, c_d^{l_d}), 0 \leq l_i \leq k$$

forms the grid. It is of course possible to choose different values of k for different indices i . Then one can evaluate f at each of these grid points, and then average them.

The main shortcoming of the gridding method is colorfully referred to as the “curse of dimensionality.” Specifically, the grid $\mathcal{G}$ consists of $(k + 1)^d$ points. Clearly this number grows exponentially with respect to the dimension d . The method described next overcomes the curse of dimensionality by making the number of sample points independent of d .

Generate l independent samples $\mathbf{x}_1, \dots, \mathbf{x}_l$ of $\mathbf{X}$, and define the empirical mean as

$$\hat{E}_l(f) := \frac{1}{l} \sum_{j=1}^l f(\mathbf{x}_j).$$

Then it readily follows from Hoeffding’s inequality that

$$\Pr\{|\hat{E}(f) - E[f(\mathbf{X})]|\} > \epsilon\} \leq 2 \exp[-2l\epsilon^2/(b - a)^2].$$

Thus, in order to estimate the expected value to within a specified accuracy ϵ with confidence $1 - \delta$, it is enough to choose l samples provided

$$2 \exp[-2l\epsilon^2/(b - a)^2] \leq \delta, \text{ or } l \geq \frac{\log(2/\delta)}{2\epsilon^2}.$$

The key point is that the dimension d does not appear anywhere in the required number of samples!

The success of the Monte Carlo method depends on the ability to generate independent samples of the random variable $\mathbf{X}$ that follow the specified distribution function $P_{\mathbf{X}}$. There are many standard methods available for generating pseudorandom sequences, and the reader can find them through a routine literature search.

Just to look ahead, Markov Chain Monte Carlo (MCMC) refers to the situation where the samples $\mathbf{x}_j$ are not independent, but are instead the state sequence of a Markov chain. This is discussed in Section 4.3.

1.3.5 Introduction to Cramér’s Theorem

In this subsection we give a brief introduction to a famous theorem by Cramér, which was one of the first results in “large deviation theory” for real-valued random variables. A fully rigorous treatment of the theory for real-valued random variables is beyond the scope of this book, and the interested reader is referred to an advanced text such as [41]. The simpler case of random variables assuming values in a finite set, which lends itself to analysis using combinatorial arguments, is studied in Chapter 5.

Suppose X is a real-valued random variable with the property that $e^{\lambda X}$ has finite expected value for every $\lambda \geq 0$. This means that if the probability distribution function $P_X(\cdot)$ is continuously differentiable everywhere, then the derivative $p(x) := dP_X(x)/dx$, known as the **probability density function**, has to decay faster than any exponential. The Gaussian random variable with mean μ and standard deviation σ , which has the density

$$p_{\text{Gauss}}(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp[-(x - \mu)^2/2\sigma^2],$$

satisfies this assumption; but other random variables might not.

For such a random variable, the function

$$\text{mgf}(\lambda; X) := E[e^{\lambda X}]$$

is called the **moment generating function** of X . Note that, once X is fixed, $\text{mgf}(\cdot; X) : \mathbb{R}_+ \rightarrow \mathbb{R}$. The rationale for the name is that, if we ignore the niceties associated with differentiating inside an integral, interchanging summation and integration, etc., then formally we have that

$$\left. \frac{d^n \text{mgf}(\lambda; X)}{d\lambda^n} \right|_{\lambda=0} = E \left[\left. \frac{d^n e^{\lambda X}}{d\lambda^n} \right|_{\lambda=0} \right] = E[X^n]$$

is the so-called n -th moment of X . So formally at least, we can write

$$\text{mgf}(\lambda; X) = \sum_{n=0}^{\infty} \frac{\lambda^n E[X^n]}{n!}.$$

The **logarithmic moment generating function (lmgf)** is defined as

$$\Lambda(\lambda; X) := \log \text{mgf}(\lambda; X) = \log E[e^{\lambda X}].$$

The problem studied in this subsection is the same as that addressed by the various inequalities presented thus far, namely: Suppose $X_1, \dots, X_l$ are independent copies of X , and define the l -fold average A_l as in (1.42). Given a real number a , can we estimate $\Pr\{A_l > a\}$? Note that earlier we were estimating $\Pr\{A_l - E(X) > \epsilon\}$, but the difference is purely cosmetic—all we have to do is to replace a by $E(X) + \epsilon$ to get bounds similar to the earlier ones.

Since our aim is only to give an introduction to Cramér’s theorem, and not to prove it rigorously, we deviate from the usual “theorem-proof” format wherein the end objective is stated up-front. Let us compute the mgf and lmfg of A_l . We have

$$\exp(\lambda A_l) = \exp\left[(\lambda/l) \sum_{i=1}^l X_i\right] = \prod_{i=1}^l \exp[(\lambda/l) X_i].$$

Therefore

$$E[\exp(\lambda A_l)] = \prod_{i=1}^l E[\exp((\lambda/l) X_i)] = (E[\exp((\lambda/l) X)])^l,$$

where the first step follows from the independence of the X_i ’s, and the second step follows from the fact that all X_i ’s have the same distribution. Therefore

$$\begin{aligned} \text{mgf}(\lambda; A_l) &= [\text{mgf}(\lambda/l; X)]^l, \\ \Lambda(\lambda; A_l) &= l\Lambda(\lambda/l; X). \end{aligned} \tag{1.53}$$

Next, let us apply Markov’s inequality as stated in (1.40) to A_l . This leads to

$$\Pr\{A_l > a\} \leq e^{-\gamma a} E[e^{\gamma A_l}] = e^{-\gamma a} \text{mgf}(\gamma; A_l).$$

If we set $\gamma = \lambda l$ and make use of (1.53), we get

$$\begin{aligned} \Pr\{A_l > a\} &\leq e^{-\lambda l a} [\text{mgf}(\lambda; X)]^l, \\ \log \Pr\{A_l > a\} &\leq -\lambda l a + l\Lambda(\lambda; X), \\ \frac{1}{l} \Pr\{A_l > a\} &\leq -[\lambda a - \Lambda(\lambda; X)]. \end{aligned} \tag{1.54}$$

Note that the above inequality holds for *every* $\lambda \geq 0$. So we can “optimize” this inequality by tuning λ for each a . Define

$$F(a) := \sup_{\lambda \geq 0} [\lambda a - \Lambda(\lambda; X)].$$

Then (1.54) implies that

$$\frac{1}{l} \Pr\{A_l > a\} \leq -F(a). \tag{1.55}$$

Now a famous theorem of Cramér states that as $l \rightarrow \infty$ the bound in (1.55) is asymptotically tight. In other words, under suitable conditions,

$$\lim_{l \rightarrow \infty} \frac{1}{l} \Pr\{A_l > a\} = -F(a).$$

To put it another way, as $l \rightarrow \infty$,

$$\Pr\{A_l > a\} \sim \text{const.} e^{-lF(a)}.$$

This gives the rate at which the “tail probability” $\Pr\{A_l > a\}$ decays as a function of a . Again, the reader is directed to advanced texts such as [41] for full details, including a formal statement and proof.