Probability Notes

Antonio Jiménez-Martínez

Contents

Chapter 1

Probability Measures and Random Variables

This chapter introduces the theoretical framework that allows us to assign probabilities in a wide range of uncertain or random phenomena. In many fields, scientists attempt to formulate general laws on the basis of observation and experiment. The simplest and most used scheme of such laws has the structure:

> "if a set of conditions B is satisfied, then event A occurs."

There are many examples of such laws in chemistry, physics, biology, or social sciences, such as the law of gravity, the law of conservation of mass, or the low of demand. If event A occurs inevitably whenever the set of conditions B is satisfied, then we say that A is *certain* or *sure*, under the set of conditions B. If A can never occur whenever B is satisfied, we say that A is *impossible* (again, under the set of conditions B). As an intermediate and interesting case, randomness appears if A may or may not occur whenever B is satisfied. In this case, A is said to be a *random phenomenon* (of course, under the set of conditions B) and event A is usually referred to as a *random event*.

Random phenomena is our subject matter and, unlike certain and impossible events, the presence of randomness implies that the set of conditions B do not reflect all the necessary and sufficient conditions for event A to happen. At first glance, it might seem then almost impossible to make any general statements about random phenomena. However, repetition of random phenomena shows that many of them exhibit some statistical regularities that allows us to study them with a systematic approach. For such "regular" random phenomena, it is possible to estimate the odds of occurrence of event A by using a set of laws, which are commonly referred to as *probabilistic* or *stochastic*. Stochastic laws have the general structure:

> "if a set of conditions B is satisfied repeatedly n times, then event A occurs m times out of the n repetitions.

Given this structure, if we consider that $n \to \infty$, then "the probability that event A occurs, under the set of conditions B" can be naturally estimated as the ratio m/n.

Now, how do we assign probabilities of occurrence to random events? Historically, there have been two traditional approaches to study random phenomena, the *relative frequency method* and the *classical method*. The relative frequency method has a clear empirical motivation and it relies upon observation of occurrence of the event A under a large number of repetitions of the set of conditions B. Then, one would simply count the number of times that event A has happen and use the ratio m/n as an (asymptotic) approximation of its probability of occurrence. The classical method, whose introduction is credited to Laplace [1814], also computes the probability of the event A using the ration m/n but, in addition, it makes use of the concept of *equal likelihood*, which is taken as a primitive of the model. Under this approach, random events are regarded as the aggregation of several mutually exclusive (or disjoint), and equally likely, elementary events. Then, the probability of the event of interest is obtained as the sum of the individual probabilities of the elementary events.[1] In the 20th century, Kolmogorov [1933] proposed the *axiomatic approach*, which is consistent both with the relative frequency and the classical methods. More importantly, the *axiomatic approach* allows for a systematic and rigorous treatment of a very general set of random phenomena and it is nowadays followed by every field of science that deals with randomness. The axiomatic approach provides the foundations of all the models presented and discussed in the subsequent sections.

1.1 Measurable Spaces

1.1.1 General σ-algebras

The canonical space of random events is one that allows us to consider as many events as possible, and with very general forms. We begin with an arbitrary nonempty set Ω of *elementary events* ω. An elementary event gives us a complete and exhaustive description of a possible outcome of uncertainty.[2] Then, a random event is formally identified as a subset A of the set Ω. Therefore, one would like to take as a primitive of the model a family $\mathcal{F}$ of subsets of Ω that satisfies certain desirable properties.

Intuitively, notice that random events can be fully described by using sentences. Then, given events–sentences A and B, it makes sense to connect such sentences so as to form new sentences like "A and B," "A or B," and "not A." Thus, it would be interesting that the family of events $\mathcal{F}$

[1] In his celebrated essay (Laplace [1814]), Pierre-Simon Laplace wrote: "The theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favorable to the event whose probability is sought. The ratio of this number to that of all the cases possible is the measure of this probability, which is thus simply a fraction whose numerator is the number of favorable cases and whose denominator is the number of all the cases possible."

[2] In the social sciences, the elementary events are also known as *states of the world* or states of nature.

be closed under the set operations of intersection, union, and complement. Of course, this family $\mathcal{F}$ should also include the entire set of elementary events Ω. In addition, if one wishes to form sentences built from arbitrary, perhaps infinite, sequences of other sentences, then it is useful to add closure under (arbitrary) countable unions and intersections to our list of desiderata. With this motivation in mind, a *σ-algebra on* Ω is a non-empty family $\mathcal{F}$ of subsets of Ω such that $A \in \mathcal{F}$ implies $A^c \in \mathcal{F}$, and such that $A_1, A_2, \ldots, A_n, \cdots \in \mathcal{F}$ implies $\cup_{n=1}^{\infty} A_n \in \mathcal{F}$.[3] In what follows, we will require that any set of random events of interest be formally described by a σ-algebra. Once one associates a σ-algebra $\mathcal{F}$ to a set of elementary events Ω, we refer to the pair $(\Omega, \mathcal{F})$ as a *measurable space*. Of course, the particular σ-algebra that we select will depend on the problem at hand. A couple of consequences from the introduced notion of σ-algebra are worth mentioning. First, notice that any σ-algebra contains the empty set under our definition. Thus, it follows that its complement, $\emptyset^c = \Omega$, must also belong to any σ-algebra. The event $\emptyset$ if often referred to as the *impossible event* while the event Ω is usually known as the *sure event*. Secondly, we can make use of the convenient set-operation property known as the *de Morgan's laws*, which states $\cup_{n=1}^{\infty} A_n = \left(\cap_{n=1}^{\infty} A_n^c\right)^c$ and $\cap_{n=1}^{\infty} A_n = \left(\cup_{n=1}^{\infty} A_n^c\right)^c$, to exchangeably use either the requirement that a σ-algebra must be closed under arbitrary unions or that it must be closed under arbitrary intersections.

Using set operations, the notion of σ-algebra enables us to regard as random events some descriptions that are formed used other events–sentences. Then, we will be able to assign probabilities to these (sometimes complex) descriptions which are, in turn, derived from relatively simpler ones. For example, we will be able to assign probabilities to events–sentences such as: A^c="A does not occur," $A \cap B$="both A and B simultaneously occur," $A \cup B$="either A, or B, or both, occurs," $(A \cap B^c) \cup (B \cap A^c)$="either A or B occurs, but not both of them simultaneously," $A \cap B = \emptyset$="A and B are mutually exclusive," $A \setminus B = A \cap B^c$="$A$ occurs but B does not occur," or $(A \cup B)^c$="neither A nor B occur."

To gain intuition of the importance of the requirement that a σ-algebra be closed under arbitrary unions (or intersections), consider an example where a die is rolled arbitrarily many times. Then, it is natural to set $\Omega = \{1, \ldots, 6\} \times \{1, \ldots, 6\} \times \cdots = \{1, \ldots, 6\}^{\infty}$. Suppose that we wish to consider the event "number 2 comes up in the ith roll of the die". Then, we should certainly choose a σ-algebra on Ω that contains all sets of the form

$$A_i = \left\{(\omega_i)_{i=1}^{\infty} \in \Omega \,:\, \omega_i = 2\right\}, \quad i = 1, 2, \ldots.$$

Notice that, under this σ-algebra, the situation B ="in neither the second nor the third roll number 2 comes up" is formally another event since

$$B = \left\{(\omega_i)_{i=1}^{\infty} \in \Omega \,:\, \omega_2 \neq 2, \omega_3 \neq 2\right\} = A_2^c \cap A_3^c.$$

[3] For the special case where the $\mathcal{F}$ satisfies the requirement that it is closed *only* under finite unions, we say that $\mathcal{F}$ is an *algebra*.

Also, the situations "number 2 comes up at least once through the rolls," which is described by $\cup_{i=1}^{\infty} A_i$, and "each roll results in number 2 coming up," which is described by $\{(2, 2, \dots)\} = \cap_{i=1}^{\infty} A_i$, are formally events in this σ-algebra.

From the requirements of a σ-algebra, we observe that most sets of elementary events admit multiple σ-algebras that fit into the definition. The choice of a σ-algebra is not a trivial one and, depending on the application one is dealing with, sometimes we must choose the working σ-algebra very carefully. To see this, notice that the simplest σ-algebra of any set of elementary events Ω is $\{\emptyset, \Omega\}$. Of course, this σ-algebra rules out almost all random events of interest in most applications. On the other extreme, the largest possible σ-algebra on Ω is the *power class* 2^{Ω}, which consists of the family of all subsets of Ω. By choosing 2^{Ω}, we make sure not to lose any random even from consideration. However, the size of the family 2^{Ω} increases exponentially with the the size of the set Ω. When Ω has an infinite countable number of elementary events, or when it corresponds to a continuum, the family of events 2^{Ω} can be simply too large for many applications. As mentioned earlier, even most finite sets Ω admit multiple σ-algebras that can be ordered "in size" according to set inclusion:

$$\{\emptyset, \Omega\} \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}_n \subseteq \dots 2^{\Omega}.$$

If Ω is countably infinite, then we have an infinite sequence $\{\mathcal{F}_n\}_{n=1}^{\infty}$ of σ-algebras, ordered by set inclusion. If Ω is a continuum, then we have an ordered set of σ-algebras $\{\mathcal{F}_n \ : n \in N\}$, where N is a continuum. Then, how should we choose the most appropriate σ-algebra for a set of elementary events? A practical approach would be the following. Starting with a family $\mathcal{A}$ of subsets of Ω, which need not be σ-algebra of Ω itself, one could search for a family of subsets of Ω that contains $\mathcal{A}$, such that it is a σ-algebra on Ω, and such that it is the smallest one with respect to set inclusion. This approach is captured by the notion of *σ-algebra generated by a family of sets.* The *σ-algebra generated by the family of sets $\mathcal{A}$ of a non-empty set* Ω is the family of sets

$$\sigma(\mathcal{A}) = \bigcap_{n \in N} \left\{ \mathcal{F}_n \subseteq 2^{\Omega} : \ \text{for each } \ n \in N, \ \ \mathcal{F}_n \supseteq \mathcal{A} \ \text{ is a } \ \sigma\text{-algebra on } \ \Omega \right\}.$$

The following theorem establishes that, for a family of sets $\mathcal{A}$ of Ω, $\sigma(\mathcal{A})$ is in fact a σ-algebra on Ω and, in addition, it provides a very useful implication for the σ-algebra generated by a family of sets $\mathcal{A}$ that is included in some σ-algebra $\mathcal{F}$.

Theorem 1.1. *Given a nonempty family $\mathcal{A}$ of subsets of a nonempty set Ω, $\sigma(\mathcal{A})$ satisfies:*

(i) $\sigma(\mathcal{A})$ is a σ-algebra on Ω;

(ii) $\mathcal{A} \subseteq \sigma(\mathcal{A})$;

(iii) if $\mathcal{A} \subseteq \mathcal{F}$ and $\mathcal{F}$ is a σ-algebra on Ω, then $\sigma(\mathcal{A}) \subseteq \mathcal{F}$.

Proof of Theorem 1.1. (i) First, take $A \in \sigma(\mathcal{A})$, then $A \in \mathcal{F}_n$ for each $\mathcal{F}_n \supseteq \mathcal{A}$ and each $n \in N$. Since each $\mathcal{F}_n$ is a σ-algebra on Ω, we have that $A^c \in \mathcal{F}_n$ for each $\mathcal{F}_n \supseteq \mathcal{A}$ and each $n \in N$. Therefore,

$A^c \in \sigma(\mathcal{A})$. Second, take a sequence $\{A_n\}_{n=1}^{\infty} \subseteq \sigma(\mathcal{A})$, then $\{A_n\}_{n=1}^{\infty} \subseteq \mathcal{F}_n$ for each $\mathcal{F}_n \supseteq \mathcal{A}$ and each $n \in N$. Since each $\mathcal{F}_n$ is a σ-algebra on Ω, we have that $\cup_{n=1}^{\infty} A_n \in \mathcal{F}_n$ for each $\mathcal{F}_n \supseteq \mathcal{F}$ and each $n \in N$. Therefore, $\cup_{n=1}^{\infty} A_n \in \sigma(\mathcal{A})$. (ii) The result follows directly from the definition of $\sigma(\mathcal{A})$, taking into account the set operations of inclusion and intersection. (iii) Take a σ-algebra $\mathcal{F}$ on Ω such that $\mathcal{F} \supseteq \mathcal{A}$. Then, it must be the case that $\mathcal{F} = \mathcal{F}_n$ for some $n \in N$ so that $\sigma(\mathcal{A}) \subseteq \mathcal{F}$. ■

1.1.2 Borel σ-algebras

In many applications, it is usual to choose a very particular generated σ-algebra which is known as the *Borel σ-algebra*. To present the approach most commonly used to construct such a σ-algebra, we need first to introduce another family of subsets, which are of crucial importance in real analysis and probability. A *topology on a set* Ω is a family of subsets τ of Ω that contains the empty set and the set Ω itself, and such that it is closed under finite intersections and under arbitrary (not necessarily countable) unions. A generic element of a topology on a set is referred to as an *open set* in such a family of sets. Notice that a σ-algebra on a countable set is also a topology on that set but the converse is not true. To see this, consider the following example. Take the set $\Omega = \{a, b, c, d\}$ and its family of subsets $\gamma = \{\emptyset, \{a\}, \{a, d\}, \{b, c\}, \{a, b, c, d\}\}$. We have that γ is not a σ-algebra on Ω since, for instance, $\{a\}^c = \{b, c, d\} \notin \gamma$. Furthermore, γ is not a topology on Ω either since, for instance, $\{a\} \cup \{b, c\} = \{a, b, c\} \notin \gamma$. We can add one extra element to γ so that $\tau = \gamma \cup \{a, b, c\}$ is indeed a topology on Ω. However, τ is still not a σ-algebra on Ω. However, if we look for the σ-algebras generated, respectively, by γ and τ, we obtain

$$\sigma(\gamma) = \sigma(\tau) = \{\emptyset, \{a\}, \{a, d\}, \{b, c\}, \{a, b, c, d\}, \{a, b, c\}, \{b, c, d\}, \{d\}\}.$$

Given a set of elementary events Ω and a topology τ on Ω, the *Borel σ-algebra associated to the space* (Ω, τ) is the σ-algebra generated by the family of sets τ, $\sigma(\tau)$. The generic elements of a Borel σ-algebra are referred to as *Borel sets*. Notice that the concept of *Borel σ-algebra* depends on the chosen topology τ. When the set of elementary events Ω is a subset of some Euclidean space, it is common to choose the *Euclidean topology* as the topology of reference to generate the corresponding Borel σ-algebra. In this case, one usually starts with the notion of *neighborhood* as a primitive to propose the Euclidean topology. The *Euclidean metric*[4] in $\mathbb{R}^n$ is defined as $d(\omega, \omega') = \left[\sum_{i=1}^{n} (\omega_i - \omega_i')^2\right]^{1/2}$. Given a number $\varepsilon > 0$, the set

$$B_d(\omega, \varepsilon) = \{\omega' : d(\omega, \omega') < \varepsilon\}$$

is known as the *ε-neighborhood centered at ω*. Then, the *Euclidean topology on $\mathbb{R}^n$*, $\tau[\mathbb{R}^n]$, is the family of subsets A of $\mathbb{R}^n$ such that if $\omega \in A$, then there exists some $\varepsilon > 0$ satisfying $B_d(\omega, \varepsilon) \in A$. For a set

[4] A *metric* on a set Ω is a function $d : \Omega \times \Omega \to \mathbb{R}$ such that
(i) $d(\omega, \omega') \geq 0$ for each $\omega, \omega' \in \Omega$ and $d(\omega, \omega') = 0$ if and only if $\omega = \omega'$;
(ii) $d(\omega, \omega') = d(\omega', \omega)$ for each $\omega, \omega' \in A$;
(iii) $d(\omega, \omega'') \leq d(\omega, \omega') + d(\omega', \omega'')$ for each $\omega, \omega', \omega'' \in \Omega$.

$\Omega \subseteq \mathbb{R}^n$, the *Borel σ-algebra on* Ω, is the generated σ-algebra $\sigma(\tau[\Omega])$. We will use $\mathcal{B}_\Omega$ to denote the *Borel σ-algebra on* Ω, for any set $\Omega \subseteq \mathbb{R}^n$. The following examples illustrate that relatively simple families of sets of the real line can be alternatively used to generate the Borel σ-algebra on $\mathbb{R}$.

Example 1.1. *Consider the family of open intervals in* $\mathbb{R}$,

$$\alpha = \{(a, b) \subseteq \mathbb{R} : -\infty < a < b < +\infty\}.$$

We wish to show that $\sigma(\alpha) = \mathcal{B}_\mathbb{R}$. *First, since each open interval is an open set in* $\mathbb{R}$, *we have that* $\alpha \subseteq \sigma(\tau[\mathbb{R}])$. *Then, using Theorem 1.1 (iii), we obtain that* $\sigma(\alpha) \subseteq \sigma(\tau[\mathbb{R}])$ *because* $\sigma(\tau[\mathbb{R}])$ *is a* σ*-algebra on* $\mathbb{R}$. *Secondly, since each open set in* $\mathbb{R}$ *can be expressed as the result of the union of countably many open intervals, we know that* $\tau[\mathbb{R}] \subseteq \sigma(\alpha)$. *This is so because, as a* σ*-algebra that contains* α, $\sigma(\alpha)$ *must contain the unions of countably arbitrarily many open intervals. Then* $\sigma(\tau[\mathbb{R}]) \subseteq \sigma(\alpha)$ *follows from Theorem 1.1 (iii) since* $\sigma(\alpha)$ *is a* σ*-algebra on* $\mathbb{R}$. *Therefore,* $\sigma(\alpha) = \sigma(\tau[\mathbb{R}]) = \mathcal{B}_\mathbb{R}$.

Example 1.2. *Consider the family of all bounded right-semiclosed intervals of* $\mathbb{R}$,

$$\beta = \{(a, b] \subseteq \mathbb{R} : -\infty < a < b < +\infty\}.$$

We wish to show that $\sigma(\beta) = \mathcal{B}_\mathbb{R}$ *as well. First, note that for each* $a, b \in \mathbb{R}$ *such that* $-\infty < a < b < +\infty$, *we have*

$$(a, b] = \bigcap_{n=1}^{\infty} \left(a, b + \frac{1}{n}\right).$$

Then, $\beta \subseteq \sigma(\tau[\mathbb{R}])$ *since, as a* σ*-algebra that contains* $\tau[\mathbb{R}]$, $\sigma(\tau[\mathbb{R}])$ *must contain the intersections of countably arbitrarily many open intervals. From the fact that* $\sigma(\tau[\mathbb{R}])$ *is a* σ*-algebra on* $\mathbb{R}$, *it follows, using Theorem 1.1 (iii), that* $\sigma(\beta) \subseteq \sigma(\tau[\mathbb{R}])$. *Secondly, note that for each* $a, b \in \mathbb{R}$ *such that* $-\infty < a < b < +\infty$, *we have*

$$(a, b) = \bigcup_{n=1}^{\infty} \left(a, b - \frac{1}{n}\right].$$

Then, by an argument totally analogous to the previous one, we obtain $\tau[\mathbb{R}] \subseteq \sigma(\beta)$ *and, then,* $\sigma(\tau[\mathbb{R}]) \subseteq \sigma(\beta)$. *Therefore,* $\sigma(\beta) = \sigma(\tau[\mathbb{R}]) = \mathcal{B}_\mathbb{R}$.

Using arguments totally analogous to these in the two examples above, one can show that the Borel σ-algebra on $\mathbb{R}$ coincides also with the σ-algebras generated by the following families of sets in $\mathbb{R}$:

1. the family of all closed intervals;
2. the family of all bounded left-semiclosed intervals;
3. the family of all intervals of the form $(-\infty, a]$

4. the family of all intervals of the form $[b, +\infty)$;

5. the family of all closed sets.

Since the family of all closed sets generates $\mathcal{B}_{\mathbb{R}}$, it follows that that singletons and countable sets in $\mathbb{R}$ are members of its Borel σ-algebra.

1.2 Probability Measures

Suppose that we wish to compute the probability of occurrence of a certain event $A \subset \Omega$. Intuitively, when the elementary events ω in the set Ω are equally likely, one could count the number elementary events in both sets A and Ω, and then obtain the probability of A simply as the ratio $|A| \,/\, |\Omega|$ But, how do we count these elementary events when the sets A and Ω are not finite or are uncountable? Furthermore, which rules should we follow when the elementary events ω are not equally likely? Historically, mathematicians have been interested in proposing a notion of probability as a primitive by generalizing the intuitive notions of length, area or volume. The most useful generalization of these concept is provided by the notion of a *measure*. With a general measure as a primitive, we are able to compute probabilities in a wide range of settings. A *measure* P on a measurable space $(\Omega, \mathcal{F})$ is a set function $P : \mathcal{F} \to \mathbb{R}^*$[5] with $P(\emptyset) = 0$, $P(A) \geq 0$ for each $A \in \mathcal{F}$, and such that if $\{A_n\}_{n=1}^{\infty} \subseteq \mathcal{F}$ is a sequence of pairwise disjoint events in $\mathcal{F}$, then $P\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \sum_{n=1}^{\infty} P(A_n)$. In addition, if a measure P also satisfies $P(\Omega) = 1$, then it is referred to as a *probability measure*. A *probability space* is a triplet $(\Omega, \mathcal{F}, P)$, where Ω is an arbitrary nonempty set, $\mathcal{F}$ is a σ-algebra of subsets of Ω, and P is a probability measure on $(\Omega, \mathcal{F})$. When $\mathcal{F}$ is a Borel σ-algebra, with respect to some topology on the set Ω, and P is a probability measure on $(\Omega, \mathcal{F})$, then P is often referred to as a *Borel probability measure*.

A measure that is typically used to compute probabilities of uncountable events in $\mathbb{R}$, when one considers that the drawing of each point is equally likely, is the *Lebesgue Measure*. Given a closed interval $[a, b] \subset \mathbb{R}$, the *Lebesgue Measure on* $\mathbb{R}$ is defined as $\lambda([a, b]) = b - a$. In this case, if we start with a set of elementary events $\Omega = [a, b] \subset \mathbb{R}$ such that each $\omega \in \Omega$ is equally likely, and we consider an event $A = \cup_{i=1}^{n} [a_i, b_i] \in \mathcal{B}_{\Omega}$, where the intervals $[a_i, b_i]$ are disjoint, then $P(A) = \sum_{i=1}^{n} (b_i - a_i)/(b - a)$ gives us the associated probability measure. The random phenomena captured by this, and others closely related, probability measure as often referred to phenomena that satisfy the "uniform probability law." The Lebesgue Measure is analogously considered for n-dimension Euclidean spaces. Thus, if we consider a hyperrectangle $[a, b] = \times_{i=1}^{n} [a_i, b_i] \subset \mathbb{R}^n$, then the *Lebesgue Measure on* $\mathbb{R}^n$ is defined as $\lambda([a, b]) = \prod_{i=1}^{n} (b_i - a_i)$.

Note that the following properties can be easily derived from the definition of probability

[5]The notation $\mathbb{R}^*$ indicates the *extended real line* $\mathbb{R} \cup \{-\infty, +\infty\}$

measure. First, since $\Omega = A^c \cup A$ and A^c and A are disjoint events, we have that

$$1 = P(\Omega) = P(A^c) + P(A) \Rightarrow P(A^c) = 1 - P(A),$$

which, in turn, implies that $0 \leq P(A) \leq 1$ for any event A. Secondly, suppose that $A \subseteq B$. Then, we have $B = A \cup (A^c \cap B)$, where and A and $(A^c \cap B)$ are disjoint events, and, therefore,

$$P(B) = P(A) + P(A^c \cap B) \geq P(A).$$

Thirdly, by considering the set-operation relations $A \cup B = A \cup [B \setminus (A \cap B)]$, where A and $[B \setminus (A \cap B)]$ are disjoint, and $B = (A \cap B) \cup [B \setminus (A \cap B)]$, where $(A \cap B)$ and $[B \setminus (A \cap B)]$ are disjoint as well, we obtain that

$$\begin{aligned} P(A \cup B) = P(A) + P(B \setminus (A \cap B)) \text{ and } P(B) = P(A \cap B) + P(B \setminus (A \cap B)) \\ \Rightarrow P(A \cup B) = P(A) + P(B) - P(A \cap B), \end{aligned}$$

which, in turn, directly implies that $P(A \cup B) \leq P(A) + P(B)$. This is the case $n = 2$ of a more general expression known as *inclusion-exclusion formula*:

$$\begin{aligned} P(\cup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i) - \sum_{i<j} P(A_i \cap A_j) + \sum_{i<j<k} P(A_i \cap A_j \cap A_k) \\ + \cdots + (-1)^{n+1} P(A_1 \cap \cdots \cap A_n). \end{aligned}$$

As in the $n = 2$ case, the formula above leads directly to the inequality $P\Big(\cup_{i=1}^n A_i\Big) \leq \sum_{i=1}^n P(A_i)$, which is referred to as *Boole's inequality*. Finally, since the de Morgan's laws imply that $\cup_{n=1}^\infty A_n = \Big(\cap_{n=1}^\infty A_n^c\Big)^c$, we obtain that $P(\cup_{n=1}^\infty A_n) = 1 - P(\cap_{n=1}^\infty A_n^c)$ for any sequence $\{A_n\}_{n=1}^\infty$ of events, not necessarily be disjoint.

To illustrate how the properties derived above can be used to compute probabilities in applications, suppose for instance that we toss a coin n times and wish to compute the probability of the event A ="there shows up at least one head". Here, we can take the set of elementary events as $\Omega = \{head, tail\}^n$ so that $|\Omega| = 2^n$. If we specify the event B_i ="the ith toss results in a head," then we know that $A = \cup_{i=1}^n B_i$. Note that the sets $B_1, \ldots, B_n$ are not pairwise disjoint so that we cannot obtain $P(A)$ as the sum $\sum_{i=1}^n P(B_i)$. However, using some of the properties above properties of a probability measure together, we have

$$P(A) = 1 - P(A^c) = 1 - P\Big((\cup_{i=1}^n B_i)^c\Big) = 1 - P\Big(\cap_{i=1}^n B_i^c\Big).$$

Notice that $\cap_{i=1}^n B_i^c$ consists of the event "the n tosses yield tails," i.e., $\cap_{i=1}^n B_i^c = \{(tail, \ldots, tail)\}$. Then, $P(\cap_{i=1}^n B_i^c) = 2^{-n}$ so that the probability of our event of interest can be computed as $P(A) = 1 - 2^{-n}$.

1.2.1 Extension of Probability Measures

Given a set of elementary events Ω, sometimes one starts by computing probabilities not on a σ-algebra on Ω but on an algebra $\mathcal{A}$ on such as set. A reason to follow this approach is that of considering a relatively simple family of sets since, in general, an algebra contains less sets than a σ-algebra constructed from such an algebra. Then, if we have a probability measure Q on the algebra $\mathcal{A}$ and are *only* interested in computing probabilities of occurrence of events $A \in \mathcal{A}$, such a measure Q would be enough for our purposes. However, as an algebra, $\mathcal{A}$ might not contain some relatively more complicated events, such as the unions of arbitrary countable sequences of events in $\mathcal{A}$. For these cases, we would like to know whether there is a systematic way to proceed, starting from the probability measure Q, in order to compute the probabilities of occurrence of these more complicated events.

To get a better intuition about this problem, let us consider an example where a dice is rolled an infinitely number of times so that $\Omega = \{1, \dots, 6\}^\infty$. The choice of an appropriate σ-algebra is not obvious here. To consider a suitable family of events for this case, let us fix a finite number k which indicates that we begin by focusing only on the first k draws of the dice. In particular, we wish to consider events of the form

$$A = \{((\omega_1, \dots, \omega_k), \omega_{k+1}, \dots) \in \{1, \dots, 6\}^\infty : (\omega_1, \dots, \omega_k) \in A_k\},$$

where $A_K \in 2^{\Omega_k}$, for $\Omega_k = \{1, \dots, 6\}^k$. Notice that the event A above is nothing but the event "the outcome of the first k tosses belongs to the set A_k." For instance, one can consider $k = 2$ and then ask about the probability of the event A_2 ="at least one of the first two draws results in a number larger than 4." In this case, we can resort to the modified set of elementary events $\Omega_2 = \{1, \dots, 6\}^2$ so that $|\Omega_2| = 6^2$. Since Ω_2 is finite, we can use for it the σ-algebra as 2^{Ω_2}, which is not a very complicated family. Then, if the dice is a fair one, the finiteness of Ω_2 allows us to assign probabilities to the event $A_2 = \{(5,5), (5,6), (6,5), (5,5)\}$ by using a probability measure Q on 2^Ω such that $Q(A_2) = 4/6^2 = 1/9$. For a general value of the finite number k, one would simply compute $Q(A_k) = |A_k|/6^k$. Following this reasoning, we can let k to increase so that we should be able to compute the corresponding probabilities of events when the number of draws becomes arbitrarily large. A consistent way thus to choose a family of events, for the case where the dice is rolled an infinitely number of times, would be that of selecting events of the form A_k and of considering a non-empty family $\mathcal{A}$ of such events such that be closed under complements, and *finite* unions and intersections. Such a family $\mathcal{A}$ would be an algebra on $\Omega = \{1, \dots, 6\}^\infty$ so that one could propose $\sigma(\mathcal{A})$ as a suitable σ-algebra in this case. Notice, however, that our earlier question still remains unanswered in this example. That is, we are able to assign probabilities to all the events in $\mathcal{A}$ by using the probability measure $Q(A_k) = |A_k|/6^{2^k}$ and, if needed, by using the probability rules that allows us to compute probabilities for complements, and for *finite* unions and intersections. But, how do we compute probabilities of occurrence of the events $A \in \sigma(\mathcal{A}) \setminus \mathcal{A}$?

A formal procedure to construct an extension of the measure Q to the set $\sigma(\mathcal{A}) \setminus \mathcal{A}$ was was provided by Carathéodory [1918]. Interestingly enough, this probability extension is unique, a feature that gives crucial consistency to many probability measures that are commonly proposed on complicated measurable spaces.

Theorem 1.2 (Carathéodory [1918])**.** *Let $\mathcal{A}$ be an algebra on a nonempty set Ω and let Q be a measure on $\mathcal{A}$. Then there exists a measure P on $\sigma(\mathcal{A})$ such that $P(A) = Q(A)$ for each $A \in \mathcal{A}$. Moreover, if Q is a probability measure, then P is unique.*

Thus, Theorem 1.2 allows us to construct complicated probability spaces by starting from relatively much simpler ones, such as we did in the previous example. An important implication of Theorem 1.2 is that any measure on an algebra on $\mathbb{R}$ that contains all intervals can be extended to a Borel measure on $\mathbb{R}$. This implication, in turn, provides us with a suitable foundation for the existence of the Lebesgue Measure.

1.2.2 Conditional Probability and Independence of Events

In many situations there is some information available about the outcome of the random phenomenon at the moment at which we assign probabilities. In these cases, we wish to answer questions of the form "what is the probability that event A occurs given that another event B has occurred?" Given a probability space $(\Omega, \mathcal{F}, P)$ and two events $A, B \in \mathcal{F}$ such that $P(B) > 0$, the *conditional probability of A given B* is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{1.1}$$

If $P(B) = 0$, then the conditional probability of A given B is left undefined.

Using the notion of conditional probability, we can obtain a set of *chain-rule formulas* that are useful in many applications:

$$\begin{aligned}
&P(A \cap B) = P(A)P(B|A),\\
&P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B),\\
&P(A \cap B \cap C \cap D) = P(A)P(B|A)P(C|A \cap B)P(D|A \cap B \cap C), \quad \text{and so on.}
\end{aligned}$$

Furthermore, if $\{A_n\}_{n=1}^{\infty}$ is a sequence of events that partitions the set of elementary events Ω, then the definition of conditional probability allows us to express the probability of an event B as:

$$P(B) = \sum_{n=1}^{\infty} P(A_n \cap B) = \sum_{n=1}^{\infty} P(A_n)P(B|A_n).$$

This property is often referred to as the *Law of Total Probability* .

The definition of conditional probability gives rise directly to an expression which is used to obtain conditional probabilities in many applications when we have a partition of the set of elementary events. This expression is known as *Bayes' Rule*, and it can be viewed simply as an alternative formulation of the definition of conditional probability.

Theorem 1.3 (Bayes' Rule)**.** *Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $\{A_i\}_{i=1}^{\infty}$ be a sequence of events $A_i \in \mathcal{F}$ with $P(A_i) > 0$ for each $i = 1, 2, \ldots$, and such that they partition Ω, that is, the events A_i are mutually disjoint and satisfy $\cup_{i=1}^{\infty} A_i = \Omega$. Consider an event $B \in \mathcal{F}$ such that $P(B) > 0$. Then,*

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^{\infty} P(B|A_i)P(A_i)} \quad \textit{for each given } \; k = 1, 2, \ldots.$$

Proof of Theorem 1.3. Note first that since $B = \cup_{i=1}^{\infty}(B \cap A_i)$, where the events $\{B \cap A_i\}_{i=1}^{\infty}$ are disjoint, we directly obtain the expression of the Total Probability Rule, $P(B) = \sum_{i=1}^{\infty} P(B \cap A_i)$. Then, by applying the definition of conditional probability, it follows that

$$P(B) = \sum_{i=1}^{\infty} P(B \cap A_i) = \sum_{i=1}^{\infty} P(B|A_i)P(A_i).$$

Secondly, using the definition of conditional probability again, we can write, for each given $k = 1, 2, \ldots,$

$$P(A_k|B) = \frac{P(A_k \cap B)}{P(B)} = \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^{\infty} P(B|A_i)P(A_i)},$$

as stated. ■

The following example illustrates a typical application of Bayes' Rule.

Example 1.3. *A ball is drawn from one of two urns depending on the outcome of the roll of a fair die. If the die shows* 1 *or* 2*, then the ball is drawn from Urn I, which contains* 6 *red balls and* 2 *white balls. If the die shows* 3*,* 4*,* 5*, or* 6*, then the ball is drawn from Urn II, which contains* 7 *red balls and* 3 *white balls. Suppose that we wish to know the probability that the ball came from Urn I (Urn II) given that we know that a white ball is drawn. Let us denote the event "the ball comes from Urn I (Urn II)" simply as I (II) and let us use w (r) to denote the event "the drawn ball is white (red)." Then, we can compute $P(I|w)$ and $P(II|w)$ by applying Bayes' Rule as*

$$P(I|w) = \frac{P(w|I)P(I)}{P(w|I)P(I) + P(w|II)P(II)} = \frac{(1/4)(1/3)}{(1/4)(1/3) + (3/10)(2/3)} = \frac{5}{17},$$

$$P(II|w) = \frac{P(w|II)P(II)}{P(w|I)P(I) + P(w|II)P(II)} = \frac{(3/10)(2/3)}{(1/4)(1/3) + (3/10)(2/3)} = \frac{12}{17}.$$

A very interesting feature in Probability and Statistics appears when the knowledge that an event B occurs does not change the odds that another event A occurs. In this case, we are left with

$P(A|B) = P(A)$, provided that $P(B) > 0$, and it is natural to think intuitively that A and B occur in an "independent way in probability terms." More formally, using the definition of conditional probability, we say that two events A, B are *independent events* if the probability of simultaneous occurrence of both events $P(A \cap B)$ can be decomposed multiplicatively as $P(A \cap B) = P(A)P(B)$. This is a simple criterion when one deals with a pair of events. The extension of this definition to a larger set of events $A_1, \ldots, A_n$ is nevertheless not straightforward as it requires much more than the multiplicative decomposition $P(\cap_{i=1}^{n} A_i) = P(A_1) \times \cdots \times P(A_n)$. In particular, a finite family $A_1, \ldots, A_n$ of events is *independent* if

$$P(A_{k_1} \cap \cdots \cap A_{k_j}) = P(A_{k_1}) \times \cdots \times P(A_{k_j})$$

for each relabeling of events $k_1, \ldots, k_j$, with $1 \leq k_1 < \cdots < k_j \leq n$, for each $2 \leq j \leq n$. In other words, a finite family of events is independent if each of its subfamilies is. Analogously, an infinite (perhaps uncountable) family of events is *independent* if each of its finite subfamilies is. To grasp better the subtleties behind the definition of independence for a family of events, consider a set of elementary events $\Omega = \{a, b, c, d\}$ and suppose that the probability of each $\omega \in \Omega$ is $1/4$. Consider the three events $A = \{a, b\}$, $B = \{a, c\}$, and $C = \{a, d\}$. Then, we have

$$P(A \cap B) = P(A \cap C) = P(B \cap C) = P(A \cap B \cap C) = P(\{a\}) = 1/4,$$

so that $P(A \cap B) = P(A)P(B)$, $P(A \cap C) = P(A)P(C)$, and $P(B \cap C) = P(B)P(C)$. However, $P(A \cap B \cap C) = 1/4 \neq 1/8 = P(A)P(B)P(C)$. Therefore, we obtain that events A, B, and C are pairwise independent but all three of them are not independent. Sometimes, the notion of independence of events does not have clear intuitive interpretation in terms of odds of occurrence, as it is the case in the following example.

Example 1.4. *Consider a set of elementary events* $\Omega = \left\{(x, y) \in \mathbb{R}^2 : 0 \leq x, y \leq 1\right\}$ *and consider the probability space* $(\Omega, \mathcal{B}_\Omega, \lambda)$ *where* λ *is the Lebesgue measure on* $\mathbb{R}^2$*. Suppose that we wish to know whether the events*

$$A = \left\{(x, y) \in \mathbb{R}^2 : 0 \leq x \leq 1/2,\ 0 \leq y \leq 1\right\},$$

$$B = \left\{(x, y) \in \mathbb{R}^2 : 0 \leq x \leq 1,\ 0 \leq y \leq 1/4\right\}$$

are independent or not. To answer this, we simply need to compute the area of the respective rectangles. First, notice that

$$A \cap B = \left\{(x, y) \in \mathbb{R}^2 : 0 \leq x \leq 1/2,\ 0 \leq y \leq 1/4\right\}.$$

Then, one obtains $\lambda(A) = 1/2$, $\lambda(B) = 1/4$, *and* $\lambda(A \cap B) = 1/8$, *so that* $\lambda(A \cap B) = \lambda(A)\lambda(B)$ *and* A *and* B *are independent events.*

Consider now the event

$$C = \left\{(x, y) \in \mathbb{R}^2 \,:\, 0 \leq x \leq 1/2,\ 0 \leq y \leq 1,\ y \geq x\right\}$$

We have $\lambda(C) = 1/2 - (1/2)^3 = 3/8$ *and* $\lambda(C)\lambda(B) = 3/32$. *On the other hand,* $\lambda(C \cap B) = 1/2(1/4)^2 = 1/32$ *so that C and B are not independent events.*

In other cases, the concepts of conditional probability and independence of events can be used to obtain probabilities with intuitive interpretations in terms of odds of occurrence, as the following example illustrates.

Example 1.5. *Suppose that we wish to buy two laptops from a shop that has 100 old model laptops and 1500 new model laptops in stock. An extensive market survey informs us that* 15% *of the old model, and* 5% *of the new model, laptops have some kind of defect. When an order comes in, a laptop is chosen at random from the shop stock. We decide whether to buy old or new model laptops based on the outcome of a coin toss, and, after we know the outcome of the coin toss, we order two laptops of the same model (either new or old). Suppose that we wish to know the probability that both laptops will be defective. To answer this, consider first that we choose the old model. Then, there are* $0.15 \times 100 = 15$ *defective old model laptops. Secondly, notice that the events "first is defective given that the first is old" and "second is defective and first is defective given that the second is old" are independent. Also, the simultaneous occurrence of both events gives us the event "the two laptops are defective given that the two are old." Then, probability that we choose two defective old model laptops is then*

$$\begin{aligned} P(\textit{two defective} \,|\, \textit{old}) &= P(\textit{first is defective} \,|\, \textit{first is old}) \\ &\times P(\textit{second is defective} \,|\, \textit{first is defective} \,\&\, \textit{second is old}) = \frac{15}{100} \cdot \frac{14}{99}. \end{aligned}$$

Analogously, for the case where we choose new model laptops, there are $0.05 \cdot 1500 = 75$ *defective new model laptops and, therefore, we obtain*

$$\begin{aligned} P(\textit{two defective} \,|\, \textit{new}) &= P(\textit{first is defective} \,|\, \textit{first is new}) \\ &\times P(\textit{second is defective} \,|\, \textit{first is defective} \,\&\, \textit{second is new}) = \frac{75}{1500} \cdot \frac{74}{1499}. \end{aligned}$$

Finally, using the Total Probability Law, we can compute

$$\begin{aligned} P(\textit{two defective}) &= P(\textit{two defective} \,|\, \textit{old}) \cdot P(\textit{old}) \\ &+ P(\textit{two defective} \,|\, \textit{new}) \cdot P(\textit{new}) = \frac{15}{100} \cdot \frac{14}{99} \cdot \frac{1}{2} + \frac{75}{1500} \cdot \frac{74}{1499} \cdot \frac{1}{2}. \end{aligned}$$

1.3 Random Variables and Distributions

In many applications, it is useful to have a tool that allows us to assign probabilities consistently over events on a general set of elementary events Ω by doing so over events on the real line, and

vice-versa. The notion of random variable is a key concept in probability theory as it gives us a transformation through which we can assign probabilities on subsets of arbitrary sets by means of doing so on subsets of the real line.

1.3.1 Random Variables and Random Vectors

Sometimes, we begin with relatively complicated elementary events, such as sequences or functions, and would rather be interested in working with measurable spaces whose corresponding elementary events be real numbers. This approach would then allow us to deal in a unified and consistent way with a huge variety of random phenomena. Of course, a function is the tool that can be used to move from an arbitrary initial set of elementary events to a subset of the real line. The specification of such a function would be given by the particular features of the random experiment that interests us. To illustrate this, suppose that we roll a dice ten times. In this case, the underlying elementary events would have the form $\omega = (\omega_1, \ldots, \omega_{10}) \in \Omega = \{1, \ldots, 6\}^{10}$. In some applications, however, we might be interested in the sum of the outcomes so that the required function $X : \Omega \to \mathbb{R}$ could be specified as $X(\omega) = \sum_{i=1}^{10} \omega_i$. Other applications could ask instead about the smallest outcome of the dice rolls, so that we would specify $X(\omega) = \min_{i=1,\ldots,10} \{\omega_i\}$.

However, in order to preserve the structure of the corresponding family of events, the proposed function $X : \Omega \to \mathbb{R}$ needs to satisfy a particular property, which yields the concept of *random variable*. Specifically, a *random variable* on a measurable space $(\Omega, \mathcal{F})$ is a function $X : \Omega \to \mathbb{R}$ such that for each $B \in \mathcal{B}_{\mathbb{R}}$, we have $[X \in B] \in \mathcal{F}$, where the notation

$$[X \in B] = \{\omega \in \Omega : X(\omega) \in B\}$$

is used to indicate the inverse image of X when applied to a set B. Thus, the Borel σ-algebra is usually taken as the reference σ-algebra on the real line and the crucial point of the definition of a random variable X is to guarantee that, for each Borel set $B \in \mathcal{B}_{\mathbb{R}}$, the inverse image $[X \in B]$ lies in the original σ-algebra $\mathcal{F}$.

Notice that the definition of a random variable does not depend, in principle, on any probability measure. Of course, in order to compute probabilities in applications, we need a probability measure P on the original measurable space $(\Omega, \mathcal{F})$. Then, when we make use of a random variable on a measurable space $(\Omega, \mathcal{F})$, which is in turn endowed with some probability measure, we obtain the notion of *probability distribution of the random variable*.

Given a probability space $(\Omega, \mathcal{F}, P)$ be a probability space and a random variable X a random variable on $(\Omega, \mathcal{F})$, the associated *probability distribution of the random variable* X is a probability measure ψ on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ specified by

$$\psi(B) = P[X \in B] = P(\{\omega \in \Omega \, : \, X(\omega) \in B\}).$$

The following example illustrates the notions of random variable and of its associated probability distribution.

Example 1.6. *Suppose that we roll three dice together and are interested in the sum of the numbers that show up. In principle, we could take as our primitive the probability space* $(\Omega, \mathcal{F}, P)$*, where* $\omega = (\omega_1, \omega_2, \omega_3) \in \Omega = \{1, \ldots, 6\}^3$*,* $\mathcal{F} = 2^\Omega$*, and* P *is specified by* $P(A) = |A|/6^3$ *for each* $A \in \mathcal{F}$*. Here this measurable space* $(\mathcal{F}, P)$ *is not be particularly useful since we are interested only on the sum of the numbers that show up. Then, we can make use of a function* $X : \Omega \to \mathbb{R}$ *specified as*

$$X((\omega_1, \omega_2, \omega_3)) = \omega_1 + \omega_2 + \omega_3.$$

This function X *satisfies the criterion of random variable. Now, consider the event* $B = (3, 5] \in \mathcal{B}_\mathbb{R}$*. Using the concept of probability distribution of* X*, we can compute the probability that "the sum of the numbers that show up is larger that three but no larger than five" as*

$$\psi(B) = P[X \in B] = \frac{|\{(\omega_1, \omega_2, \omega_3) \in \Omega : 3 < \omega_1 + \omega_2 + \omega_3 \leq 5\}|}{6^3} = \frac{9}{6^3}.$$

Many random phenomena exhibit several features or, in other words, take place simultaneously in more than a single dimension. To study the odds of occurrence of several features that stem from a common underlying probability space, the concept of random variable can be readily extended to that of random vector. By doing so, the relevant events would belong to a multidimensional Euclidean space. A *random vector* on a measurable space $(\Omega, \mathcal{F})$ is a function $X = (X_1, \ldots, X_n) : \Omega \to \mathbb{R}^n$ such that for each $B \in \mathcal{B}_{\mathbb{R}^n}$, we have $[X \in B] = \{\omega \in \Omega : (X_1(\omega), \ldots, X_n(\omega)) \in B\} \in \mathcal{F}$. Furthermore, a function $X = (X_1, \ldots, X_n) : \Omega \to \mathbb{R}^n$ constitutes a random vector on some $(\Omega, \mathcal{F})$ if and only if each of its components X_i is a random variable on $(\Omega, \mathcal{F})$. Thus, a random vector is simply a list of random variables. The concept of probability distribution can be also analogously extended to a random vector. The *probability distribution of the random vector* $X = (X_1, \ldots, X_n)$ associate to an underlying probability space $(\Omega, \mathcal{F}, P)$ is a probability measure ψ on $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ specified by

$$\psi(B) = P[X \in B] = P(\{\omega \in \Omega : (X_1(\omega), \ldots, X_n(\omega)) \in B\}) \quad \text{for each} \quad B \in \mathcal{B}_{\mathbb{R}^n}.$$

1.3.2 Distribution Functions

Sometimes it is useful to use an alternative formulation to compute the probabilities captured by the probability distribution of a random variable. The associated *distribution function* of the random variable X with probability distribution ψ is the function $F : \mathbb{R} \to \mathbb{R}$ specified by $F(x) = \psi((-\infty, x]) = P[X \leq x]$. Thus, a distribution function is simply a function that gives probabilities only of Borel sets of the form $(-\infty, x]$. In this sense, the distribution function F of a random variable seems, at least at first glance, more restrictive than its probability distribution ψ. Nevertheless, a few key properties of a probability measure allow us to use generally the distribution function of a random variable to compute probabilities exactly the same way as we would do using its probability distribution. Some of these properties deal with the continuity of a probability measure and are stated in the theorem below. A nice proof of this result is provided by Billingsley [1995].

Theorem 1.4 (Billingsley [1995]). *Let ψ be a probability measure on a measurable space $(\Omega, \mathcal{F})$, then:*

(i) continuity from below: if $A, A_1, \ldots, A_n, \ldots$ are events with $A_1 \subseteq A_2 \subseteq \ldots$ and $A = \cup_{n=1}^{\infty} A_n$, then $\psi(A_1) \leq \psi(A_2) \leq \ldots$ and $\lim_{n\to\infty} \psi(A_n) = \psi(A)$;

(ii) continuity from above: if $A, A_1, \ldots, A_n, \ldots$ are events with $A_1 \supseteq A_2 \supseteq \ldots$ and $A = \cup_{n=1}^{\infty} A_n$, then $\psi(A_1) \geq \psi(A_2) \geq \ldots$ and $\lim_{n\to\infty} \psi(A_n) = \psi(A)$;

(iii) if the set of elementary events Ω can be obtained as the union of some finite or countable sequence of events, then the corresponding σ-algebra $\mathcal{F}$ cannot contain an uncountable disjoint collection of events $\{A : A \in \mathcal{F}\}$ with $\psi(A) > 0$.

Some useful properties of a distribution function F can be derived directly from Theorem 1.4 above. First, by using the property of the probability measure ψ on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ that states that $A \subseteq B$ implies $\psi(A) \leq \psi(B)$, we know that F is monotone nondecreasing. Secondly, by continuity from above of the probability measure ψ on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ (Theorem 1.4 (ii)), we obtain

$$\lim_{y\to x^+} F(y) = \lim_{\varepsilon\to 0} \psi((-\infty, x + \varepsilon]) = \psi((-\infty, x]) = F(x) = P[X \leq x],$$

so that F is right-continuous. Thirdly, by continuity from below of the probability measure ψ on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ (Theorem 1.4 (i)), the left-hand limit $\lim_{y\to x^-} F(y) = \lim_{\varepsilon\to 0} \psi((-\infty, x - \varepsilon])$ exists and

$$\lim_{y\to x^-} F(y) = \psi((-\infty, x)) = P[X < x].$$

Therefore, the "jump" of F at a x is

$$P[X = x] = \psi(\{x\}) = F(x) - \lim_{y\to x^-} F(y),$$

which, combined with the result (iii) of Theorem 1.4, leads to that the distribution function F can have at most countably many points of discontinuity. Clearly, another property that follows is that $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to+\infty} F(x) = 1$. Conversely, it can be shown that any function $F : \mathbb{R} \to \mathbb{R}$ that satisfies the properties above is in fact a distribution function. Formally,

Theorem 1.5. *Let $F : \mathbb{R} \to \mathbb{R}$ be a monotone nondecreasing, right-continuous function satisfying*

$$\lim_{x\to-\infty} F(x) = 0 \quad \textit{and} \quad \lim_{x\to+\infty} F(x) = 1.$$

Then, there exists a random variable X on some probability space $(\Omega, \mathcal{F}, P)$ such that $F(x) = P[X \leq x]$.

Proof of Theorem 1.4. We consider the probability space $((0,1), \mathcal{B}_{(0,1)}, P)$, where P is the Lebesgue measure on $((0,1), \mathcal{B}_{(0,1)})$. To grasp better the logic behind the proof, suppose first that F is strictly increasing and continuous so that, in this case, $F : \mathbb{R} \to (0,1)$ is a one-to-one mapping. Let $v : (0,1) \to \mathbb{R}$ be the inverse mapping $v = F^{-1}$. Since F is a one-to-one function, then we know that

ν is a strictly increasing function. Let $X : (0,1) \to \mathbb{R}$ be the function specified as $X(\omega) = \nu(\omega)$ for $\omega \in (0,1)$. Since ν is strictly increasing, then X is a random variable on $((0,1), \mathcal{B}_{(0,1)})$. For a given $\omega \in (0,1)$, we have that $\nu(\omega) \leq x$ if and only if $\omega \leq F(x)$. Furthermore, since P is a Lebesgue measure, then we know that

$$P[X \leq x] = P\big(\{\omega \in (0,1) : \nu(\omega) \leq x\}\big) = P\big((0, F(x)]\big) = F(x) - 0 = F(x),$$

as required.

To complete the proof, consider now the case where either F has discontinuities or it is not strictly increasing. Let us define, for $\omega \in (0,1)$, $\nu(\omega) = \inf\{x \in \mathbb{R} : \omega \leq F(x)\}$. Note that, since F is nondecreasing and right-continuous, then the set $\{x \in \mathbb{R} : \omega \leq F(x)\}$ is in fact an interval with the form $[\nu(\omega), +\infty)$ for some $\omega \in (0,1)$ (i.e., it is closed on the left and stretches to $+\infty$). Therefore, we obtain again that $\nu(\omega) \leq x$ if and only if $\omega \leq F(x)$ so that, by specifying $X(\omega) = \nu(\omega)$ for $\omega \in (0,1)$, and by applying the same arguments as above, we obtain again that X is a random variable on $((0,1), \mathcal{B}_{(0,1)}, P)$ and that $P[X \leq x] = F(x)$. ∎

Hence, the previous results allow us, in many applications, to use directly the distribution function of a random variable to compute probabilities over Borel sets over than those with the form $(-\infty, a]$. For instance, some typical computations are $P[X > a] = 1 - F(a)$ or, for $b > a$, $P[a < X \leq b] = F(b) - F(a)$.

Multidimensional distribution functions can be considered when we are interested in probability calculations over several features and, therefore, must deal with random vectors. The *joint distribution function* of the random vector $X = (X_1, \ldots, X_n)$ with probability distribution ψ is the function $F : \mathbb{R}^n \to \mathbb{R}$ specified by $F(x_1, \ldots, x_n) = \psi(S_x) = P[X_1 \leq x_1, \ldots, X_1 \leq x_1]$, where $S_x = \{(y_1, \ldots, y_n) \in \mathbb{R}^n : y_i \leq x_i,\ i = 1, \ldots, n\}$ is the Euclidean region of points "southwest" with respect to x.

1.3.3 Discrete Random Variables

When a random phenomena can yield at most a countable number (perhaps infinite) of possible outcomes, the associated random variable is said to be *discrete*. Discrete random variables are associated to probability measures that assign positive probability of occurrence only to finitely or countably many points. All random experiments related to drawing a number of times (even an arbitrary number of times) elements from a finite set are described by means of discrete random variables. Coin tosses and dice rolls are typical examples of such experiments. Let us formalize this idea.

A *support* for a probability measure ψ on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ is a Borel set $S \in \mathcal{B}_{\mathbb{R}}$ satisfying $\psi(S) = 1$. Then, a random variable X and its probability distribution ψ are said to be *discrete* if ψ has a countable support $S = \{x_1, x_2, \ldots, x_n, \ldots\}$. Furthermore if ψ has a finite support, then the corresponding

random variable is said to be a *simple random variable*. Of course, a probability measure can admit an infinite number of different supports. More precisely, if S is a support of a probability measure ψ and $S \subset T \in \mathcal{B}_{\mathbb{R}}$, then T is also a support of ψ since it must be the case that $\psi(T) = 1$ and $\psi(A) = 0$ for each Borel set $A \in T \setminus S$. For simplicity, it is commonly understood that one seeks for the minimal support with respect to set inclusion. Thus, when we wish to make statements about which outcomes of a random phenomena have actually positive probability of occurrence, we are usually concerned about the support $\text{supp}(\psi) \in \mathcal{B}_{\mathbb{R}}$ of the corresponding distribution ψ such that $S \subset \text{supp}(\psi)$ implies $\psi(S) < 1$. From the result of Theorem 1.4 (ii), we know that the minimal support $\text{supp}(\psi)$ of a probability measure ψ is unique. Then, we will use $\text{supp}(X)$ to denote the minimal support of the random variable X (associated to the minimal support of the corresponding probability measure). In this case, the support of the random variable X will be simply identified with the range of X considered as a function.

When a random variable X is discrete, its corresponding probability distribution ψ is completely determined by the values $\psi(\{x_i\}) = P[X = x_i]$ for $i = 1, 2, \ldots$. For a discrete random variable X, the function $f : \text{supp}(X) \to [0, 1]$ which gives us $f(x_i) = P[X = x_i]$ is often referred to as *discrete density function* or *mass function*. Of course, as a consistency requirement with the probabilities computed by a discrete density function, it must be the case that $\sum_{x_i \in \text{supp}(X)} f(x_i) = 1$. Using the discrete density function of a random variable, we can compute values of its distribution function simply as

$$F(x) = P[X \leq x] = \sum_{x_i \leq x} P[X = x_i] = \sum_{x_i \leq x} f(x_i).$$

The following example presents an experiment that can be formalized by means of a discrete random variable and illustrates how discrete density functions can be derived.

Example 1.7. *Consider a box that contains a red balls and b black balls. We pick randomly n balls from the box. In doing so, we replace each ball back into the box after each draw. Let us use X to indicate the number of red balls finally picked along the n draws. We would like to compute the discrete density of the random variable X. To answer this, let us specify the set of balls as*

$$S = \{1, \ldots, a, a + 1, \ldots, a + b\},$$

where we follow the convention that $\{1, \ldots, a\}$ are the red balls and $\{a + 1, \ldots, a + b\}$ are the black balls. Then, since there is replacement, our set of elementary events is $\Omega = S^n$ so that $|\Omega| = (a + b)^n$. The random variable X can be then specified as

$$X(\omega) = X((\omega_1, \ldots, \omega_n)) = |\{\omega_i \in S : \omega_i \leq a, \ \ i = 1, \ldots, n\}|.$$

Since the discrete density function of X is defined as $f(x) = P[X = x]$, we need to compute the number of possible samples which have exactly a number x of its coordinates no larger than a. In other words, we must

compute the cardinality of the event

$$A = \{\omega \in \Omega : |\{\omega_i \leq a\}| = x\}.$$

Since the draws are with replacement, notice that there are a^x ways of selecting x coordinates yielding numbers no larger than a and b^{n-x} ways of selecting the remaining $n-x$ coordinates yielding numbers between $a+1$ and $a+b$. Finally, there are $\binom{n}{x}$ ways of choosing x coordinates from the n coordinates in the sample. Then, we obtain

$$f(x) = \binom{n}{x} a^x b^{n-x} (a+b)^{-n}.$$

Now, in this experiment the probability of choosing a red ball after drawing one ball from the box is $p = a/(a+b)$. This is commonly known as the probability of success in a sequence of n Bernoulli trials*. Using this probability of success, we can rewrite $f(x)$ as*

$$f(x) = \binom{n}{x} \left(\frac{a}{a+b}\right)^x \left(\frac{b}{a+b}\right)^{n-x} = \binom{n}{x} p^x (1-p)^{n-x},$$

which corresponds to the density function of a Binomial distribution *with parameter p.*

A random vector is said to be *discrete* if all its components are discrete random variables. For a discrete random vector $X = (X_1, \ldots, X_n)$, the function $f : \mathbb{R}^n \to \mathbb{R}$ defined as

$$f(x_1, \ldots, x_n) = P[X_1 = x_1, \ldots, X_n = x_n],$$

which gives us the probabilities of occurrence of each combination of outcomes, is often referred to as the *discrete joint density function* of X. Sometimes we begin with a random vector but are interested instead in computing probabilities of occurrence along a single component of vector. This idea is captured by the concept of *marginal distribution of the random vector*. For the case of a discrete random vector $X = (X_1, \ldots, X_n)$, we derive the *marginal discrete density function* of a random variable X_i simply as

$$f_i(x_i) = \sum_{\text{supp}(X_1)} \cdots \sum_{\text{supp}(X_{i-1})} \sum_{\text{supp}(X_{i+1})} \cdots \sum_{\text{supp}(X_n)} f(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$$

The idea here is that we focus only on the component X_i of interest and abstract from all the others. This procedure to obtain marginal distributions applies generally when we wish to restrict attention to a subset components included in our original random vector. Some of the concepts that we have just introduced are illustrated in the example below.

Example 1.8. *Consider a random vector (X, Y) with joint discrete density function*

$$f(x, y) = c(x^2 + y^2) \quad \textit{for} \quad x \in \{1, 2, 4\} \quad \textit{and} \quad y \in \{1, 3\},$$

where c is some real number. Note first that

$$\text{supp}(X, Y) = \{(1,1), (1,3), (2,1), (2,3), (4,1), (4,3)\}$$

and then, for $f(x, y)$ to be a density, we need

$$c(1+1) + c(1+9) + c(4+1) + c(4+9) + c(16+1) + c(16+9) = 1 \quad \Rightarrow \quad c = 1/72.$$

Using this density, we can now compute, for example, the following probabilities:

$$P[X > Y] = P(\{(2,1)\}) + P(\{(4,1)\}) + P(\{(4,3)\}) = \frac{5}{72} + \frac{17}{72} + \frac{25}{72} = \frac{47}{72},$$

$$P[Y = 3] = P(\{(1,3)\}) + P(\{(2,3)\}) + P(\{(4,3)\}) = \frac{10}{72} + \frac{13}{72} + \frac{25}{72} = \frac{48}{72}.$$

Also, we can obtain the marginal density

$$f_x(x) = \sum_{y \in \{1,3\}} f(x,y) = \begin{cases} \frac{2}{72} + \frac{10}{72} & \text{for } x = 1 \\ \frac{5}{72} + \frac{13}{72} & \text{for } x = 2 \\ \frac{17}{72} + \frac{25}{72} & \text{for } x = 4 \end{cases} = \begin{cases} \frac{12}{72} & \text{for } x = 1 \\ \frac{18}{72} & \text{for } x = 2 \\ \frac{42}{72} & \text{for } x = 4. \end{cases}$$

1.3.4 Continuous Random Variables

Unlike the cases described in the preceding subsection, other random phenomena exhibit the property that the set of its possible outcomes are uncountable. For example, many economic models, both with micro and macro emphases, assume for mathematical tractability that crucial sets are uncountable. Instances of such sets are ubiquitous in Economic Theory, such as consumption and production sets, set of prices, sets of strategies and beliefs in game theory and its applications, or sets of investment choices in finance, to name a few. The most common approach to deal with random variables in these cases is that described by *continuous random variables*. A random variable X and its probability distribution ψ are said to be *continuous* if there exists a function $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) \geq 0$ for each $x \in \mathbb{R}$ and

$$P[X \in B] = \psi(B) = \int_B f(x)dx \quad \text{for each} \quad B \in \mathcal{B}_{\mathbb{R}}. \tag{1.2}$$

The function f is referred to as *density function (with respect to Lebesgue measure)*.[6] Of course, if $B = \mathbb{R}$ in the expression (1.2) above, then f must integrate to 1.

[6] Alternatively, when a notion of integral more general than the *Riemann Integral* is required, the probability computation expressed in condition (1.2) above can be written as

$$P[X \in B] = \psi(B) = \int_B f(x)\lambda(dx),$$

where λ is the Lebesgue measure on $\mathbb{R}$.

Since the Borel σ-algebra can be generated by the family of all intervals with the form $(a, b] \subset \mathbb{R}$, it follows from Carathéodory's Theorem (Theorem 1.2) that the requirement in (1.2) above holds for each Borel set if it is satisfied for each interval $(a, b] \subset \mathbb{R}$. This enables us to, alternatively, state that a random variable X with associated distribution function F is continuous if there exists a nonnegative function $f : \mathbb{R} \to \mathbb{R}$ such that

$$P[a < X \leq b] = F(b) - F(a) = \int_a^b f(x)dx \quad \text{for each } a < b. \tag{1.3}$$

Notice that the only requirement that needs to be satisfied for X to be a continuous random variable is that it has a density function f that integrates properly as expressed in (1.3) above. In particular, the distribution function need not differentiate everywhere, neither F or f need be continuous everywhere. Nevertheless, if f is a continuous function, then it follows from the *Fundamental Theorem of Calculus* that $F'(x) = f(x)$ and that f is a density function associated to the distribution function F. The basic question that remains then is what happens when f is not continuous? First, to gain some intuition about the direction of the answer, recall that a distribution function F of any random variable can have at most countably many points of discontinuity so that it is continuous almost everywhere (with respect to Lebesgue measure). Further, a more formal the answer can be obtained by resorting to the famous *Lebesgue Differentiation Theorem* (see Lebesgue [1910]). This theorem requires that F be nondecreasing, which we already know it is the case, and states that if the condition in (1.3) is satisfied, then F can be can be differentiated almost everywhere (with respect to Lebesgue measure) and that $F'(x) = f(x)$ holds at each continuity point x of f.

Analytically, we can use continuous random variables to compute probabilities in a way very similar to the one we would follow for the cases described by discrete random variables. Notice first that the support of continuous random variable X with density function f is specified as $\text{supp}(X) = \{x \in \mathbb{R} : f(x) > 0\}$. Intuitively, the support of the random varialbe includes only Borel sets whose probability of occurrence is positive according to the corresponding density. Also, let us compare how we compute probabilities in the discrete case,

$$P[X \leq x] = F(x) = \sum_{x_i \leq x} f(x_i) \quad \text{and} \quad P[a < X \leq b] = F(b) - F(a) = \sum_{x_i > a}^{x_i \leq b} f(x_i),$$

with the case where X is a continuous random variable,

$$P[X \leq x] = F(x) = \int_{-\infty}^{x} f(y)dy \quad \text{and} \quad P[a < X \leq b] = F(b) - F(a) = \int_a^b f(x)dx.$$

A random vector is said to be *continuous* if all its components are continuous random variable. For each continuous random vector $X = (X_1, \ldots, X_n)$ there exists a density function $f : \mathbb{R}^n \to \mathbb{R}$ such that we can compute the probability of occurrence of each Borel set $B \in \mathcal{B}_{\mathbb{R}^n}$ as

$$P[(X_1, \ldots, X_n) \in B] = \int_B \cdots \int f(x_1, \ldots, x_n)dx_1 \cdots dx_n.$$

The density f is often referred to as the *joint density function* (with respect to Lebesgue measure) of X. The joint distribution function of a continuous random vector is related with its joint density function as

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_n} f(y_1, y_2, \ldots, y_n) dy_1 dy_2 \ldots dy_n.$$

For the case of a continuous random vector $X = (X_1, \ldots, X_n)$, we derive the *marginal density function* of a random variable X_i simply as

$$f_i(x_i) = \int_{\mathbb{R}^{n-1}} \cdots \int f(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) dx_1 \cdots dx_{i-1} x_{i+1} \cdots dx_n.$$

Of course, we can also relate the corresponding *marginal distribution function* $F_i(x_i) = P[X_i \leq x_i]$ to the density $f_i(x_i)$ in the usual way.

Similarly to the case of events, we sometimes wish to compute the probability that a random variable X takes certain values given that we know that some event B has occurred or that another random variable Y yields some particular values. We can do such computations by using the *conditional distribution* of the random variable X. The required definitions exhibit some technical differences depending on whether the random variables of interest are discrete or continuous. For the case where X is a discrete random variable, the *conditional discrete density* of X given that another random variable Y lies in some Borel set B, with $P[Y \in B] > 0$, is the nonnegative function

$$f_{X|B}(x) = \frac{P[\{X = x\} \cap \{Y \in B\}]}{P[Y \in B]}.$$

Since $P[Y \in B] = \sum_{x \in \text{supp}(X)} P[\{X = x\} \cap \{Y \in B\}]$, we see that $\sum_{x \in \text{supp}(X)} f_{X|B}(x) = 1$, as needed for $f_{X|B}(x)$ to be indeed a density. When Y is a discrete variable as well, it makes sense to deal with the information that the random variable has taken a particular value $Y = y$, with $P[Y = y] > 0$. In this case, the expression of the *conditional discrete density* of X becomes

$$f_{X|y}(x) = \frac{P[X = x, Y = y]}{P[Y = y]} = \frac{f(x, y)}{f_2(y)},$$

where f denotes the joint density function of the random vector (X, Y) and f_2 stands for the marginal density of Y. For the case where X is a continuous random variable, the *conditional density* of X given that another random variable Y lies in some Borel set B, with $P[Y \in B] > 0$, is the nonnegative function $f_{X|B}(x)$ that satisfies the equality:

$$\int_A f_{X|B}(x) dx = \frac{P[\{X \in A\} \cap \{Y \in B\}]}{P[Y \in B]} \qquad \text{for each } A \in \mathcal{B}_{\mathbb{R}}.$$

Since $P[Y \in B] = \int_{\text{supp}(X)} P[\{X \in A\} \cap \{Y \in B\}]$, we observe, by letting $A = \text{supp}(X)$ in the expression above, that $f_{X|B}(x)$ integrates to one over the support of X. Finally, if we consider a value y for the random variable, then we obtain the same expression as in the discrete case for the corresponding conditional density, that is, $f_{X|y}(x) = f(x, y)/f_2(y)$.

1.3.5 Functions of a Random Variables

A typical problem in many applications is that of obtaining the probability distribution of some transformation $Y = g(X)$ of a random variable X. To this end, one must verify first that Y is indeed a random variable. Then, given the requirement of the definition of a random variable, it is common to restrict attention to those cases where the transformation g is given by a one-to-one function.

The treatment of this problem is relatively simple in the discrete case, as the following example illustrates.

Example 1.9. *Consider a discrete random variable X with discrete density function $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$, for some $p \in (0,1)$, and whose support is $\{1, 2, \ldots, n\}$. Let $Y = g(X) = a + bX$ for some $a, b > 0$. We are interested in obtaining the discrete density function of Y. Let us denote such a density by h. First, notice that, by applying g to the elements in the support of X, the support of Y is $g(\{1, 2, \ldots, n\}) = \{a + b, a + 2b, \ldots, a + nb\}$. Then, we can simply compute*

$$h(y) = f((y-a)/b) = \binom{n}{(y-a)/b} p^{(y-a)/b} (1-p)^{n-(y-a)/b},$$

where $y \in \{a + b, a + 2b, \ldots, a + nb\}$.

Thus, for the case of discrete random variables, we observe that we only need to find the inverse function of the transformation g and then incorporate it directly into the corresponding discrete density function.

Obtaining the probability distribution of a transformation of continuous random variable, on the other hand, can sometimes be done using a systematic approach. Let us propose a simple rule to deal with transformations of continuous random variables. To do this in a systematic way, we need to assume that the corresponding density function is continuous and that the transformation $g : \mathbb{R} \to \mathbb{R}$ is a one-to-one continuously differentiable function function. In this case, the inverse function $T = g^{-1}$ exists and it is differentiable as well. Let us use H and h to denote, respectively, the distribution function and the density of the random variable $Y = g(X)$. Then, if the random variable X has distribution function F and continuous density f, we already know that $F'(x) = f(x)$ holds for each $x \in \mathbb{R}$. To proceed, suppose first that g is increasing. In this case, for each $y \in \mathbb{R}$, we have

$$H(y) = P[g(X) \leq y] = P[X \leq T(y)] = F(T(y)).$$

Since $T = g^{-1}$ is differentiable, we obtain

$$h(y) = \frac{d}{dy} H(y) = \frac{d}{dy} F(T(y)) = F'(T(y))T'(y) = f(T(y))T'(y).$$

Now, suppose that g is decreasing. Then, we know that

$$H(y) = P[g(X) \leq y] = P[g(X) < y] = P[X > T(y)] = 1 - F(T(y)),$$

so that

$$h(y) = \frac{d}{dy}H(y) = -F'(T(y))T'(y) = -f(T(y))T'(y).$$

Therefore, in either case the random variable $Y = g(X)$ has density

$$h(y) = f(T(y))\,|T'(y)|\,.$$

The above arguments constitute a proof to the following useful result.

Theorem 1.6. *Let $g : U \to V$ be a one-to-one continuously differentiable function, where U, V are open sets in $\mathbb{R}$. Suppose that $T = g^{-1}$ satisfies $T'(y) \neq 0$ for each $y \in V$. Then, if X is a continuous random variable with density f, supported in U, it follows that the random variable $Y = g(X)$ has density h, supported in V, given by*

$$h(y) = \begin{cases} f(T(y))\,|T'(y)| & \text{if } y \in V, \\ 0 & \text{if } y \notin V. \end{cases}$$

The following example illustrates how the result in Theorem 1.6 can be applied.

Example 1.10. *Consider a positive continuous random variable X with continuous density f and suppose that we are interested in obtaining the density function of $1/X$. Note that $T(y) = g^{-1}(y) = 1/y$, which is differentiable for each $y \geq 0$. Also, $T'(y) = -1/y^2$ so that $h(y) = f(1/y)/y^2$.*

Beyond the simple (and systematic) rule provided by Theorem 1.6, we can use the reasoning followed in its proof to obtain the density of a transformation $Y = g(X)$ in cases where g is not one-to-one, as the following example shows.

Example 1.11. *Suppose that X is a continuous random variable with density*

$$f(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}, \quad \text{for } x \in \mathbb{R}.$$

This density corresponds to a distribution known as Normal *with parameters $(0, 1)$. Let $Y = X^2$ be a random variable with distribution function H and density function h. Notice that the transformation $g(X) = X^2$ is not one-to-one. However, the fact that the function f is symmetric around the origin allows us to write*

$$\begin{aligned} H(y) = P[X^2 \leq y] = P[-\sqrt{y} \leq X \leq \sqrt{y}] &= \frac{1}{\sqrt{2\pi}}\int_{-\sqrt{y}}^{\sqrt{y}} e^{-x^2/2}dx \\ &= \frac{2}{\sqrt{2\pi}}\int_{0}^{\sqrt{y}} e^{-x^2/2}dx. \end{aligned}$$

Now, we can deal with the integral above by proposing the change of variables $t = x^2$ *so that* $dx = 1/2\sqrt{t}dt$. *Recall that we need to apply this change of variables to the integration limits as well. Then, we obtain*

$$H(y) = \int_0^y \frac{1}{\sqrt{2\pi t}} e^{-t/2} dt.$$

Since it must be the case that $H(y) = \int_0^y h(t)dt$ *and* $supp(Y) = \mathbb{R}_+$, *we obtain that* $Y = X^2$ *has density*

$$h(y) = \frac{1}{\sqrt{2\pi y}} e^{-y/2} \quad \textit{for} \quad y > 0.$$

The density $h(y)$ *obtained aboe corresponds to a distribution known as* Chi-square *with parameter 1.*

Furthermore, when the transformation g is neither one-to-one nor continuous, the logic behind the result of the Theorem 1.6 can, in some cases, be applied parts of the function g which, taken separately, are one-to-one. The following result states formally such a use of the "Change of Variables Theorem."

Theorem 1.7. *Let* X *and* $Y = g(X)$ *be two continuous random variables and let* f *be the density function of* X. *Suppose that there exists a partition* $\{A_0, A_1, \ldots, A_k\}$ *of* $supp(X)$ *such that* $P[X \in A_0] = 0$ *and* f *is continuous in each* A_i $(i = 0, 1, \ldots, k)$. *Suppose that there are functions* $g_1(x), g_2(x), \ldots, g_k(x)$ *defined, respectively, on* $A_1, A_2, \ldots, A_k$ *such that*

(i) $g(x) = g_i(x)$ *for each* $x \in A_i$ $(i = 1, 2, \ldots, k)$;

(ii) each g_i *is monotone in* A_i $(i = 1, 2, \ldots, k)$;

(iii) the set $V = \{y \in \mathbb{R} : y = g_i(x) \text{ for some given } x \in A_i\}$ *is the same for each* $i = 1, 2, \ldots, k$;

(iv) $g_i^{-1}(y)$ *is continuously differentiable on* V *for each* $i = 1, 2, \ldots, k$.

Then, the density function h *of the random variable* Y *is given by*

$$h(y) = \sum_{i=1}^{k} f(g_i^{-1}(y)) \left| \frac{d}{dy} g_i^{-1}(y) \right| \quad \textit{for} \quad y \in V.$$

Changes of variables for a discrete random vectors can done quite straightforwardly, just as shown for the case of a discrete random variable. As to continuous random vectors, we can follow a systematic approach that parallels the one given by Theorem 1.6, provided that certain conditions are satisfied. More precisely, let $g : U \to V$ be a one-to-one continuously differentiable function, where $U, V \subseteq \mathbb{R}^n$ are open sets. We begin with n random variables $X_1, \ldots, X_n$, with joint density function f, and transform them into "new" random variables $Y_1, \ldots, Y_n$ using the functions

$$\begin{aligned} y_1 &= g_1(x_1, \ldots, x_n) \\ &\vdots \\ y_n &= g_n(x_1, \ldots, x_n). \end{aligned}$$

Then, we ask about the joint density function of $Y_1, \ldots, Y_n$. Let $T = g^{-1}$ denote the corresponding inverse function and suppose that its Jacobian never vanishes, that is,

$$J(y) = \left| \begin{pmatrix} \frac{\partial T_1}{\partial y_1}(y) & \cdots & \frac{\partial T_1}{\partial y_n}(y) \\ \vdots & \ddots & \vdots \\ \frac{\partial T_n}{\partial y_1}(y) & \cdots & \frac{\partial T_n}{\partial y_n}(y) \end{pmatrix} \right| \neq 0 \quad \text{for each} \quad y \in V.$$

Under these conditions we can state the following useful result, which is simply a generalization of Theorem 1.6 for the multidimensional case.

Theorem 1.8. *Let $g : U \to V$ be a one-to-one continuously differentiable function, where U, V are open sets in $\mathbb{R}^n$. Suppose that $T = g^{-1}$ satisfies $J(y) \neq 0$ for each $y \in V$. If X is a random vector with density f, supported in U, then the random vector $Y = g(X)$ has density h, supported in V, and given by*

$$h(y) = \begin{cases} f(T(y))\,|J(y)| & \text{if } y \in V, \\ 0 & \text{if } y \notin V. \end{cases}$$

The following example illustrates an application of the result in Theorem 1.8 above

Example 1.12. *Let (X_1, X_2) a continuous random vector with joint density function*

$$f(x_1, x_2) = e^{-(x_1+x_2)}, \qquad \text{for } \; x_1, x_2 \in \mathbb{R}_+.$$

Consider the transformation given by

$$y_1 = x_1 + x_2, \quad y_2 = 2x_1 - x_2.$$

Suppose that we wish to find the joint density function of the random vector (Y_1, Y_2). To do this, note first that

$$x_1 = \frac{y_1 + y_2}{3}, \quad x_2 = \frac{2y_1 - y_2}{3}.$$

Then, by applying the result in Theorem 1.8 above, we obtain

$$|J(y)| = \left| \frac{\partial x_1}{\partial y_1}\frac{\partial x_2}{\partial y_2} - \frac{\partial x_2}{\partial y_1}\frac{\partial x_1}{\partial y_2} \right| = \frac{1}{3},$$

and, consequently,

$$h(y_1, y_2) = \frac{1}{3}e^{-y_1} \quad \text{for} \quad y_1 \geq 0.$$

1.3.6 Independence of Random Variables

The reasoning behind the idea of independent random events extends to the various dimensions of a random phenomenon that is captured by a multidimensional random variable. In particular,

the study of many implications requires that we pay attention to what happens when the odds of occurrence of the events described by a random variable does not affect the probability of the events described by the other. This case can be formally characterized by several, seemingly different, conditions. One of the most traditional definitions of independence of random variables states that a set of random variables $X_1, \ldots, X_n$ are *independent* if the probability of the simultaneous occurrence of Borel sets with the from $(a_i, b_i]$, $i = 1, \ldots, n$, can be decomposed as the product of the probabilities of these events, that is, if

$$P[X_1 \in (a_1, b_1], \ldots, X_n \in (a_n, b_n]] = P[X_1 \in (a_1, b_1]] \times \cdots \times P[X_n \in (a_n, b_n]]. \tag{1.4}$$

Interestingly enough, this definition suffices to cover our earlier requirement for the case of independence of a (finite) collection of random events. Recall that, according to such a requirement we should verify that the respective multiplicative decomposition be satisfied *for each subcollection* of random variables of the original collection. Here, these conditions are automatically satisfied by the definition in (1.4) above. To see this, suppose that we wish to verify whether a subset of random variables $\{X_{j_1}, \ldots, X_{j_k}\} \subset \{X_1, \ldots, X_n\}$ meets the required multiplicative decomposition of probabilities. Then, we only need to consider the condition in the definition in (1.4) above and take $(a_{j_m}, b_{j_m}] = \mathbb{R}$ for those random variables that are not included in the subset that we consider, that is, for each $m \notin \{1, \ldots, k\}$.

As mentioned above, other alternative formulations of independence or random variable are common in Probability and its applications. The message conveyed by following Theorem is central to understand such alternative formulations.

Theorem 1.9. *Consider a set of elementary events Ω and a set of algebras $\mathcal{A}_1, \ldots \mathcal{A}_n$ on Ω. If each collection of events $A_1, \ldots, A_n$, with $A_i \in \mathcal{A}_i$ for $i = 1, \ldots, n$, is independent, then each collection of events $B_1, \ldots, B_n$, with $B_i \in \sigma(\mathcal{A}_n)$ for $i = 1, \ldots, n$ is independent too.*

In order to propose alternative conditions characterizing independence of random variables, notice first that we can take $\lim_{a_i \to -\infty}(a_i, x_i]$ in definition (1.4) above to obtain that the required condition must apply to all sets of the form $(-\infty, x_i]$, with $x_i \in \mathbb{R}$, as well. On the other hand, suppose that the condition in (1.4) above is satisfied for all sets of the form $(-\infty, x_i]$, with $x_i \in \mathbb{R}$. Then, since the Borel σ-algebra on $\mathbb{R}$ can be generated by all sets of the form $(-\infty, x_i]$, we have, by applying the result of Theorem 1.9, that such a multiplicative condition must hold for all sets of the form $(a_i, x_i]$, with $a_i \in \mathbb{R}$, as well. Therefore, we obtain that the requirement in condition (1.4) is satisfied if and only if the corresponding multiplicative decomposition can be expressed in terms of the joint distribution function of the random variables, that is, whenever

$$F(x_1, \ldots, x_n) = F_1(x_1) \times \cdots \times F_n(x_n) \quad \text{for each} \quad (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

In addition, for the case where the random variables of interest $X_1, \ldots, X_n$ are discrete, Theorem

1.9 above also allows us to state that they are independent if and only if, for each $(x_1, \ldots, x_n) \in \mathbb{R}^n$, we have

$$\begin{aligned} P[X_1 = x_1, \ldots, X_n = x_n] &= P[X_1 = x_1] \times \cdots \times P[X_n = x_n] \\ &\Leftrightarrow f(x_1, \ldots, x_n) = f_1(x_1) \times \cdots \times f_n(x_n). \end{aligned}$$

As for the case where the random variables $X_1, \ldots, X_n$ are continuous, we can simply resort to the multiplicative decomposition of the corresponding joint distribution function and to the result of the famous Fubini's Theorem on iterated integrals, to obtain a totally analogous condition in terms of the joint density function. In particular, a set of continuous random variables $X_1, \ldots, X_n$ are independent if and only if

$$f(x_1, \ldots, x_n) = f_1(x_1) \times \cdots \times f_n(x_n) \quad \text{for each} \quad (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

Beyond these quite common formulations, the result in Theorem 1.9 enables us to state a general necessary condition[7] of independence of random variables. In particular, notice that, if we know that the random variables $X_1, \ldots, X_n$ are independent, then it must be the case that

$$P[X_1 \in B_1, \ldots, X_n \in B_n] = P[X_1 \in B_1] \times \cdots \times P[X_n \in B_n]$$

for *any* collection of Borel sets $B_1, \ldots, B_n \in \mathcal{B}_{\mathbb{R}}$.

In some applications, knowing that a set random variables are independent is very helpful to obtain the distribution of transformations of random variables for which the general rule proposed earlier in Theorem 1.8 would not apply, as the following example illustrates.

Example 1.13. *Let X_1, X_2, X_3 be independent continuous random variables with common density*

$$f(x) = e^{-x} \qquad \text{for } x > 0,$$

and suppose that we are interested in obtaining the density function $h(y)$ of the random variable $Y = \min\{X_1, X_2, X_3\}$. Then, for a given number $y > 0$, we have

$$\begin{aligned} H(y) &= P[\min\{X_1, X_2, X_3\} \leq y] = 1 - P[\min\{X_1, X_2, X_3\} > y] \\ &= 1 - P[X_1 > y, X_2 > y, X_3 > y] = 1 - P[X_1 > y]P[X_2 > y]P[X_3 > y] \\ &= 1 - \left(\int_y^\infty e^{-x} dx \right)^3 = 1 - e^{-3y}. \end{aligned}$$

Consequently, $h(y) = H'(y) = 3e^{-3y}$ for $y > 0$.

[7] Of course, it is also a sufficient condition. However, for obvious tractability reasons, one does not use it a condition to check for independence in applications.

Working with independent random variables also allows us to obtain simple expressions for the distributions of sums of such random variables. Perhaps the best known of such expressions is the *convolution formula*. The following example deals with a simple application of the convolution of two continuous random variables.

Example 1.14. *Consider two continuous and independent random variables X and Y, with joint density $f(x, y)$ and marginal densities $f_1(x)$, $f_2(y)$. Suppose that we wish to obtain the density of the sum $X + Y$. To do this, let us first propose the transformations $U = X + Y$ and $V = Y$, and then let us apply the rule provided by Theorem 1.8 to the mapping $(X, Y) \mapsto (U, V) = (X + Y, Y)$. If we let $h(u, v)$ denote the joint density of the random vector (U, V), then we obtain directly that $h(u, v) = f(u - v, v)$. Now, since X and Y are independent random variables, we know that $h(u, v) = f_1(u-v)f_2(v)$. Thus, to obtain the (marginal) density of $U = X + Y$ we simply need to integrate the density $h(u, v)$ over the support of the random variable $V = Y$:*

$$h_{X+Y}(u) = \int_{supp(Y)} f_1(u - y) f_2(y) dy.$$

The formal arguments under the idea of independent random variables can be extended readily to random vectors. One simply needs to verify that any of the earlier definitions of independence holds with the appropriate changes in the formula so as to consider random vectors instead of random variables. In particular, if each X_i is a k_i-dimensional random vector, for $i = 1, \ldots, n$, then the random vectors $X_1, \ldots, X_n$ are *independent* if, for each $x_1 \in \mathbb{R}^{k_1}, \ldots, x_n \in \mathbb{R}^{k_n}$, we have

$$P[X_1 \leq x_1, \ldots, X_n \leq x_n] = P[X_1 \leq x_1] \times \cdots \times P[X_n \leq x_n].$$

1.3.7 Expected Values and Moments

The information contained in the probability distribution of a random variable can often be summarized by some characteristics of the general shape of the distribution and its location. Such characteristics are in most cases described by numbers known as the *moments of the distribution*. One of the moments that is extensively used in many applications corresponds to the *expected value* of the random variable. In intuitive terms, the expected value of a random variable gives us the weighted average, according to the probabilities of occurrence given by its distribution, of the possible values of the random variable. When X is a discrete random variable with discrete density f, its *expected value* is

$$E[X] = \sum_{x_i \in \text{supp}(X)} x_i P[X = x_i] = \sum_{x_i \in \text{supp}(X)} x_i f(x_i).$$

When X is a continuous random variable with density f, its *expected value* is

$$E[X] = \int_{x \in \text{supp}(X)} x f(x) dx,$$

provided that the function $xf(x)$ integrates properly over the support of X. We observe that the definitions for the discrete and continuous cases are very similar. For simplicity, the remaining of this Section will present its concepts only in terms of the integral notation so that the case of discrete random variables only requires that we change integrals into sums in the appropriate formulae. Also, to reduce the notational burden, many applications make use of μ_X (or simply of μ) to denote instead the expected value of the random variable X.

A few useful properties of the expected value of a random variable can be derived directly by applying some properties of the integral (or sum). Here is a list a some properties commonly used in many applications:

1. $E[\alpha] = \alpha$ for each $\alpha \in \mathbb{R}$;
2. $E[\alpha X + \beta Y] = \alpha E[X] + \beta E[Y]$ for each $\alpha, \beta \in \mathbb{R}$;
3. if X and Y are independent random variables, then $E[X \cdot Y] = E[X] \cdot E[Y]$;
4. if $X \leq Y$ almost everywhere (with respect to Lebesgue measure), then $E[X] \leq E[Y]$.

The concept of conditional distribution allows us to obtain in many applications the expected value, or the variance, of a random variable given that we have some available information about some event, or about another random variable. For instance, sometimes we would like to know the expected value of a random variable X given that the realization of another random variable Y is y. In this case, we simply need to compute

$$E[X|Y = y] = \int_{\text{supp}(X)} xf(x|y)dx.$$

The following examples illustrate how conditional expected values and variances can be obtained.

Example 1.15. *Suppose that X and Y are two continuous random variables with joint density*

$$f(x, y) = n(n-1)(y-x)^{n-2} \quad \textit{for } 0 \leq x \leq y \leq 1,$$

where $n > 2$ is some integer. We wish to compute the conditional density and conditional expected value of Y given $X = x$.

First, note that the marginal density of X is given by

$$\begin{aligned} f_1(x) &= \int_{supp(Y)} f(x, y)dy = n(n-1) \int_x^1 (y-x)^{n-2}dy \\ &= n(n-1)\left[\frac{(y-x)^{n-1}}{n-1}\right]_x^1 = n(1-x)^{n-1} \end{aligned}$$

for $0 \leq x \leq 1$. *Therefore, for* $0 \leq x \leq y < 1$, *we have that*

$$f(y|x) = \frac{(n-1)(y-x)^{n-2}}{(1-x)^{n-1}}$$

Thus, we obtain

$$\begin{aligned} E[Y|X=x] &= \int_{-\infty}^{+\infty} y f(y|x) dy \\ &= (n-1)(1-x)^{1-n} \int_x^1 y(y-x)^{n-2} dy. \end{aligned}$$

To compute the integral above, note that

$$\begin{aligned} y(y-x)^{n-2} &= [y-x+x](y-x)^{n-2} \\ &= x(y-x)^{n-2} + (y-x)(y-x)^{n-2} \\ &= x(y-x)^{n-2} + (y-x)^{n-1}. \end{aligned}$$

So, by using the algebraical identity above, we obtain

$$\begin{aligned} E[Y|X=x] &= (n-1)(1-x)^{1-n} \int_x^1 \left[x(y-x)^{n-2} + (y-x)^{n-1} \right] dy \\ &= (n-1)(1-x)^{1-n} \left[\frac{x(1-x)^{n-1}}{n-1} + \frac{(1-x)^n}{n} \right] \\ &= \frac{(n-1)(1-x)}{n} + x = \frac{n-1+x}{n}. \end{aligned}$$

Example 1.16. *Let us go back to the pair of random variables specified in Example 1.8. These were two discrete random variables* (X, Y) *with joint density function*

$$f(x,y) = \frac{1}{72}(x^2 + y^2) \quad \textit{for} \quad x \in \{1,2,4\} \quad \textit{and} \quad y \in \{1,3\},$$

so that

$$\operatorname{supp}(X,Y) = \{(1,1),(1,3),(2,1),(2,3),(4,1),(4,3)\}.$$

Also, let us consider the event $A =$ "$X \geq Y$." *We wish to obtain* $E[X]$, $\operatorname{Var}[X]$, $E[X \cdot Y]$, $E[X \,|\, A]$, *and* $\operatorname{Var}[X \,|\, A]$. *First, recall from Example 1.8 that*

$$f_X(x) = \begin{cases} \frac{12}{72} & \textit{for } x = 1 \\ \frac{18}{72} & \textit{for } x = 2 \\ \frac{42}{72} & \textit{for } x = 4. \end{cases}$$

Then, we have

$$E[X] = 1 \cdot \frac{12}{72} + 2 \cdot \frac{18}{72} + 4 \cdot \frac{42}{72} = 3$$

and

$$\mathrm{Var}[X] = (1-3)^2 \cdot \frac{12}{72} + (2-3)^2 \cdot \frac{18}{72} + (4-3)^2 \cdot \frac{42}{72} = \frac{3}{2}.$$

To compute $E[X \cdot Y]$*, note first that* $E[X \cdot Y] = E[X] \cdot E[Y]$ *does not follow necessarily.*[8] *However, we can use instead directly the joint distribution function of the vector* (X, Y)*. We obtain*

$$\begin{aligned} E[X \cdot Y] =& \frac{1}{72}\Big[1 \cdot 1 \cdot (1^2 + 1^2) + 1 \cdot 3 \cdot (1^2 + 3^2) + 2 \cdot 1 \cdot (2^2 + 1^2) + \\ & 2 \cdot 3 \cdot (2^2 + 3^2) + 4 \cdot 1 \cdot (4^2 + 1^2) + 4 \cdot 3 \cdot (4^2 + 3^2)\Big] = \frac{61}{9}. \end{aligned}$$

To calculate $E[X \mid A]$ *and* $\mathrm{Var}[X \mid A]$*, we have first to compute the conditional density*

$$f(x \mid A) = \frac{P(\{X = x\} \cap A)}{P(A)} \quad \textit{for } x \in \{1, 2, 4\}.$$

We obtain

$$\begin{aligned} P(A) &= P[X \geq Y] = P(\{(1,1)\}) + P(\{(2,1)\}) + P(\{(4,1)\}) + P(\{(4,3)\}) \\ &= \frac{2}{72} + \frac{5}{72} + \frac{17}{72} + \frac{25}{72} = \frac{49}{72} \end{aligned}$$

and, therefore,

$$f(x \mid A) = \begin{cases} \frac{(2/72)}{(49/72)} & \textit{for } x = 1 \\ \frac{(5/72)}{(49/72)} & \textit{for } x = 2 \\ \frac{(42/72)}{(49/72)} & \textit{for } x = 4 \end{cases} = \begin{cases} \frac{2}{49} & \textit{for } x = 1 \\ \frac{5}{49} & \textit{for } x = 2 \\ \frac{42}{49} & \textit{for } x = 4. \end{cases}$$

Then, we obtain

$$E[X \mid A] = 1 \cdot \frac{2}{49} + 2 \cdot \frac{5}{49} + 4 \cdot \frac{42}{49} = \frac{180}{49}$$

and

$$E[X^2 \mid A] = 1^2 \cdot \frac{2}{49} + 2^2 \cdot \frac{5}{49} + 4^2 \cdot \frac{42}{49} = \frac{694}{49},$$

so that

$$\mathrm{Var}[X \mid A] = E[X^2 \mid A] - (E[X \mid A])^2 = \frac{694}{49} - \left(\frac{180}{49}\right)^2 = \frac{1606}{2401}.$$

The expected value is just one of a (countable) family of quantities that gives us some information about the shape of the corresponding distribution. If X is a random variable with density f and $Y = g(X)$ is another random variable with density h, then the expected value of the transformation $g(X)$ is given by

$$E[g(X)] = \int_{\mathrm{supp}(X)} g(x) f(x) dx = \int_{g(\mathrm{supp}(X))} y h(y) dy.$$

[8] In general, $E[X \cdot Y] \neq E[X] \cdot E[Y]$ unless X and Y are independent.

To obtain the expression of our family of quantities of interest, let us consider $g(X) = X^r$, where r is any positive integer. The *moment of order* $r \in \{1, 2, \dots\}$ of the random variable X with density f is the number

$$m_r(X) = \int_{\text{supp}(X)} x^r f(x) dx.$$

Of course, we should be concerned about the fact that the integral above may not exist for some integers $r = 1, 2, \dots$. Let us propose a sufficient condition for the existence of moments of a distribution. Notice first that $|x|^k \leq |x|^r + 1$ for each $k = 1, \dots, r$. Therefore, we know that $\int_{\text{supp}(X)} |x|^r f(x) dx < \infty$ implies that $\int_{\text{supp}(X)} |x|^k f(x) dx < \infty$ for each $k = 1, \dots, r$. Then, since $|x|^r f(x) \geq x^r f(x)$, we obtain that

$$\int_{\text{supp}(X)} |x|^r f(x) dx < \infty$$

is a sufficient condition for the existence of all moments of order $k = 1, \dots, r$ of the random variable X.

We observe that the expected value of a random variable $\mu = E[X]$ coincides with its first moment, $m_1(X)$. Another quantity that is extensively used to study certain features of a distribution is its *variance*. The *variance* of a random variable X with density f is the number

$$\text{Var}[X] = E[(X - \mu)^2] = \int_{\text{supp}(X)} (x - \mu)^2 f(x) dx.$$

As in the case of the expected value, many applications use σ_X^2 (or simply of σ) to denote instead the variance of the random variable X. Naturally, the positive square root σ_X of the variance σ_X^2 of a random variable X also provides us with a measure of the dispersion of X. This measure σ_X is commonly known as the *standard deviation* of the random variable X. Notice that the variance of a random variable gives us a measure of its average dispersion, weighted according to its distribution, with respect to its expected value. Some straightforward algebra leads to

$$\begin{aligned} \text{Var}[X] &= \int_{\text{supp}(X)} x^2 f(x) dx - \left(\int_{\text{supp}(X)} x f(x) dx \right)^2 \\ &= E[X^2] - \mu^2 = m_2(X) - m_1^2(X). \end{aligned}$$

Therefore, moments up to order 2 can be used to study the dispersion of a distribution. Similarly, the moment of order 3 is used in many applications to measure how asymmetric a distribution is with respect to its expected value (formally, the *skewness of the distribution*) and the moment of order 4 can be used to measure the weight to the tail of the distribution (formally, its *kurtosis*). In this sense, our knowledge about the shape of a distribution improves with the number of its moments that we are able to obtain. Intuitively, if we were close to know *all* the moments of a distribution, this would be equivalent to have full information about the entire shape of the distribution. On the

other hand, even very small differences in two distributions should give us some differences in their moments. Theorem 1.12 in Subsection 1.3.9 will provide us with an interesting characterization result that can be viewed as a formal statement of such an intuition.

1.3.8 Covariance and Correlation

Subsection 1.3.6 presented the case where two random variables X and Y were independent. Here the odds of occurrence of the events captured by one of the variables did not affect the probability of occurrence of the events described by the other variable. Suppose now that we know that the two random variables X and Y are indeed *not* independent. In this case, it would be very interesting to have some measure about the extent to which the probabilities of occurrence of events along both dimensions are related. The notions of *covariance* and *correlation* allow us to study the degree of relation between two random variables in terms of their distributions. For two random variables X and Y with joint density $f(x, y)$, the *covariance* between them is

$$\text{Cov}[X, Y] = \int_{\text{supp}(X)} \int_{\text{supp}(Y)} (x - \mu_X)(y - \mu_Y) f(x, y) dx\, dy. \tag{1.5}$$

Given this definition, the *correlation coefficient* of the two random variables is the ratio

$$\rho(X, Y) = \frac{\text{Cov}[X, Y]}{\sigma_X \sigma_Y}.$$

If $\text{Cov}[X_1, X_2] = 0$, which in turn implies that $\rho(X, Y) = 0$ for finite standard deviations σ_X, σ_Y, then we say that the random variables X and Y are *uncorrelated*.

Let us study the relation between independence of two random variables and their correlation. Consider two random variables X and Y with density f. By applying the definition of covariance in (1.5) above, we obtain

$$\begin{aligned}
\text{Cov}[X, Y] &= \int_{\text{supp}(X)} \int_{\text{supp}(Y)} (x - \mu_X)(y - \mu_Y) f(x, y) dx\, dy \\
&= \int_{\text{supp}(X)} \int_{\text{supp}(Y)} xy f(x, y) dx\, dy + \mu_X \mu_Y \\
&\quad - \mu_X \int_{\text{supp}(Y)} y f_2(y)\, dy - \mu_Y \int_{\text{supp}(X)} x f_1(x)\, dx \\
&= E[X \cdot Y] - \mu_X \mu_Y.
\end{aligned} \tag{1.6}$$

Suppose that the random variables X and Y are independent. Then, we have $E[X \cdot Y] = E[X]E[Y]$ so that, using the equality obtained in (1.6) above, it necessarily follows that $\text{Cov}[X, Y] = 0$. Therefore, independence of two random variables implies that they are uncorrelated. However, two uncorrelated random variables need not be independent in general.

On the other hand, positive values of $\text{Cov}[X, Y]$ indicate that, according to their odds of occurrence, Y tends to increase as X does. On the other hand, negative values of $\text{Cov}[X, Y]$ indicate that Y tends to decrease when X increases. Higher values of $\text{Cov}[X, Y]$ in absolute terms reflect higher degrees of relation in the odds of occurrence of the events described by X and Y. To gain intuition about these insights consider the special case where X and Y are indicator functions, respectively, of whether any of two events A and B occurs, that is, $X = 1$ whenever A occurs and $Y = 1$ whenever B occurs. Notice that, in this case, the product $X \cdot Y$ is either one, when both events A and B occur (that is, with probability $P[X = 1, Y = 1]$), or zero, when any of the events does not occur. Then, by applying the expression derived for the covariance in (1.6) above to this discrete case, we see that

$$\text{Cov}[X, Y] = P[X = 1, Y = 1] - P[X = 1]P[Y = 1] = f(1, 1) - f_1(1)f_2(1).$$

Therefore, we obtain that $\text{Cov}[X, Y] > 0$ if and only if

$$f(1, 1) > f_1(1)f_2(1) \Leftrightarrow \frac{f(1, 1)}{f_2(1)} > f_1(1) \Leftrightarrow P[X = 1 \mid Y = 1] > P[X = 1].$$

In other words, positive covariance between X and Y in this example indicates that if event B occurs, this increases the probability of occurrence of A.

The correlation coefficient of two random variables gives us exactly the same qualitative information about relation between the variables as their covariance. By dividing the covariance over the product of the standard deviations, we obtain a normalization of the measure described by the covariance that, furthermore, is restricted to lie in the interval $[-1, 1]$. As we have already seen, if the variables are independent, then their correlation is zero. On the other hand, values of the correlation coefficient that tend to 1 indicate high positive dependence in terms of the probabilities of occurrence of the events captured by the random variables. Values that tend to -1 reflect high negative dependence. The following result is useful to see that the correlation coefficient indeed lies between -1 and 1.

Theorem 1.10 (Schwarz's Inequality)**.** *Let W and Z be two random variables, then*

$$(E[W \cdot Z])^2 \leq E[W^2]E[Z^2].$$

Proof of Theorem 1.10. Note first that if either $E[W^2] = 0$ or $E[Z^2] = 0$, then the inequality in the Theorem above holds directly. Thus, suppose that $E[Z^2] \neq 0$. Then, we obtain

$$0 \leq E\left[\left(W - \frac{E[W \cdot Z]}{E[Z^2]} \cdot Z\right)^2\right] = E\left[W^2 + \frac{(E[W \cdot Z])^2}{(E[Z^2])^2} \cdot Z^2 - 2\frac{E[W \cdot Z]}{E[Z^2]} \cdot W \cdot Z\right] =$$

$$= E[W^2] - \frac{(E[W \cdot Z])^2}{E[Z^2]} \Rightarrow (E[W \cdot Z])^2 \leq E[W^2]E[Z^2],$$

as stated. ■

Now, given two random variables X and Y, we can construct another pair of random variables (W, Z) as $W = X - \mu_X$ and $Z = Y - \mu_Y$, and then apply the result of Theorem 1.10 above to the variables W, Z. We obtain directly that

$$\begin{aligned} &\left(E[(X-\mu_X)(Y-\mu_Y)]\right)^2 \leq E\left[(X-\mu_X)^2\right]E\left[(Y-\mu_Y)^2\right] \\ \Leftrightarrow &\left(\frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{E\left[(X-\mu_X)^2\right]E\left[(Y-\mu_Y)^2\right]}}\right)^2 \leq 1 \;\Rightarrow\; |\rho(X,Y)| \leq 1. \end{aligned}$$

1.3.9 Moment Generating Functions

Subsection 1.3.7 presented some arguments suggesting that the family of all moments of a random variable provides us with very detailed information about its distribution. More formally, there exist two functions, or transformations, that generate the moments of a random variable (or random vector) and, furthermore, such that each of them completely characterizes its distribution. Therefore, dealing with such any of these transformations is equivalent to having the distribution itself. While the main caveat of these functions is that they lacks a intuitive interpretation, they are very convenient to obtain results regarding particular distributions. Since they fully characterize multidimensional distributions, these functions are particularly helpful both to study independence of random variables and to derive key results when the involved random variables are indeed independent. From a practical point of view, these functions also allow us to derive moments of a random variable without the need of computing the integral required in the definition.

The simplest of these functions, and perhaps the most used of the two, is referred to as the *moment generating function*. The *moment generating function* of a random variable X with density f is a function $\phi_X : \mathbb{R} \to \mathbb{R}$ specified as

$$\phi_X(t) = E\left[e^{tX}\right] = \int_{\text{supp}(X)} e^{tx} f(x)dx$$

for each $t \in \mathbb{R}$ for which $\phi_X(t)$ is finite. Closely related to the moment generating function, there is the other transformation that allows us to obtain the moments of a distribution. This transformation is known as the *characteristic function* of the random variable. The *characteristic function* of a random variable X with density f is a function $\varphi_X : \mathbb{R} \to \mathbb{R}$ specified as

$$\varphi_X(t) = E\left[e^{itX}\right] = \int_{\text{supp}(X)} e^{itx} f(x)dx.$$

Technically, while the moment generating function involves the function e^{tX}, known as the *Laplace transformation* of the random variable, the characteristic function uses a complex version of such a function, which is commonly known as the *Fourier transformation*. The characteristic function

has the advantage that it always exists because the transformation e^{itx} is bounded. However, for tractability reasons, most applications resort to the moment generating function of a distribution rather than to its characteristic function. Given that it always exists, the characteristic function is more often used in formal argument to obtain general results about distributions.

To see how the moment generating function can be used to easily compute the moments of the corresponding distribution, let us invoke the Tailor expansion result

$$e^{tX} = 1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \cdots = \sum_{k=0}^{\infty} \frac{t^kX^k}{k!}.$$

Now, suppose that function exists $\phi_X(t)$ throughout some interval $(-\bar{t}, \bar{t})$, for $\bar{t} > 0$. Then, we obtain

$$\phi_X(t) = \sum_{k=0}^{\infty} \frac{t^k}{k!} E[X^k] \quad \text{for } t \in (-\bar{t}, \bar{t}).$$

Therefore, the moment of order r of the random variable X can simply be calculated by taking the r-th order derivative of the function $\phi(t)$ and them by substituting $t = 0$, that is,

$$m_r(X) = \phi_X^{(r)}(t)\Big|_{t=0}. \tag{1.7}$$

Similarly, if the moment generating function of a random variable X exists for all $t \in \mathbb{R}$, then the characteristic function of such random variable can be written as

$$\varphi_X(t) = \sum_{k=0}^{\infty} \frac{(it)^k}{k!} E[X^k],$$

an expression that can be used to obtain the order r moment of X in a totally analogous way, that is, it requires that we first calculate the r-th order derivative of $\varphi_X(t)$ and then substitute $t = 0$. In particular, the analogue of the derivation in (1.7), when using instead the characteristic function, is

$$m_r(X) = i^{-r} \varphi_X^{(r)}(t)\Big|_{t=0}.$$

The definition of moment generating function can be extended readily to random vectors. If $X = (X_1, \ldots, X_n)$ is a random vector joint density f, the *moment generating function* of X is now a vector-valued function $\phi_X : \mathbb{R}^n \to \mathbb{R}$, specified as

$$\phi_X(t_1, \ldots, t_n) = \int_{\text{supp}(X)} \cdots \int e^{(t_1x_1 + \cdots + t_nx_n)} f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n$$

for each $(t_1, \ldots, t_n) \in \mathbb{R}^n$ for which $\phi_X(t_1, \ldots, t_n)$ is finite. Using the theory of Taylor expansions,

exactly as in the case of a random variable, one obtains

$$\begin{aligned}\frac{\partial^r \phi_X(0,\ldots,0)}{\partial t_i^r} &= \int_{\text{supp}(X)} \cdots \int x_i^r f(x_1,\ldots,x_n)\, dx_1 \cdots dx_n \\ &= \int_{\text{supp}(X_i)} x_i^r \int_{\mathbb{R}^{n-1}} \cdots \int f(x_1,\ldots,x_n)\, dx_1 \cdots dx_n \\ &= \int_{\text{supp}(X_i)} x_i^r f_i(x_i)\, dx_i = m_r(X_i).\end{aligned}$$

The can also work with the analog concepts for the characteristic function. If $X = (X_1,\ldots,X_n)$ is a random vector joint density f, the *characteristic function* of X is now a vector-valued function $\varphi_X : \mathbb{R}^n \to \mathbb{R}$, specified as

$$\varphi_X(t_1,\ldots,t_n) = \int_{\text{supp}(X)} \cdots \int e^{\sum_{k=1}^n i t_k x_k} f(x_1,\ldots,x_n)\, dx_1 \cdots dx_n.$$

Then, using the theory of Taylor expansions, we obtain

$$\frac{\partial^r \varphi_X(0,\ldots,0)}{\partial t_k^r} = i^{-r} m_r(X_k).$$

The moment generating function, or the characteristic function, of a random vector can be used to the study whether a set of random variables are independent or not. Suppose that the random variables $X_1,\ldots,X_n$ are independent and that each X_i has a moment generating function $\phi_i(X_i) = E[e^{t_i X_i}]$ for each $t_i \in (-\bar{t},\bar{t})$, for $\bar{t} > 0$. It then follows that $E[e^{\sum_i^n t_i X_i}] = \prod_{i=1}^n E[e^{t_i X_i}]$ so that the moment generating function of the random vector can be decomposed as the product of the moment generating functions of its components. The same argument applies for the corresponding characteristic function as well. Formally, we have

Theorem 1.11. *Let $X = (X_1,\ldots,X_n)$ be a random vector with moment generating function $\phi_X(t_1,\ldots,t_n)$ for each $t_i \in (-\bar{t},\bar{t})$, for some $\bar{t} > 0$, and with characteristic function $\varphi_X(t_1,\ldots,t_n)$. Then, the random variables $X_1,\ldots,X_n$ are independent if and only if*

$$\phi_X(t_1,\ldots,t_n) = \prod_{i=1}^n \phi_{X_i}(t_i)$$

for each $t_i \in (-\bar{t},\bar{t})$, or, equivalently, if and only if

$$\varphi_X(t_1,\ldots,t_n) = \prod_{i=1}^n \varphi_{X_i}(t_i).$$

The following example deals with the application of the moment generation function to independent random variables.

Example 1.17. *Let $X_1, X_2, \ldots, X_k$ be a set of discrete random variables with common (discrete) density function*

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad \text{for } x \in \{0, 1, 2, \ldots, n\},$$

where $n \geq 1$ is an integer and $p \in (0, 1)$. In Chapter 2, it will be shown that the moment generating function of each X_i is given by

$$\phi_{X_i}(t) = [(1-p) + pe^t]^n.$$

Suppose that the random variables $X_1, X_2, \ldots, X_k$ are independent. Then, the moment generating function of the random variable $X = \sum_i^k X_i$ can be easily obtained as

$$\begin{aligned}\phi_X(t) &= E[e^{t\sum_{i=1}^k X_i}] = E[\Pi_{i=1}^k e^{tX_i}] \\ &= \Pi_{i=1}^k E[e^{tX_i}] = \Pi_{i=1}^k [(1-p) + pe^t]^n = [(1-p) + pe^t]^{kn}.\end{aligned}$$

It follows then that the random variable X has density function

$$f(x) = \binom{kn}{x} p^x (1-p)^{kn-x}, \qquad \text{for } x \in \{0, 1, 2, \ldots, kn\}.$$

Finally, as mentioned earlier, the moment generating function, or the characteristic function, of a random variable can be used to characterize the distribution of the random variable.

Theorem 1.12 (Inversion Theorem)**.** *The moment generating function of a random variable $\phi_X(t)$ (or its characteristic function $\varphi_X(t)$) uniquely determines its probability distribution, provided that it exists for each $t \in (-\bar{t}, \bar{t})$, for some $\bar{t} > 0$.*

Billingsley [1995] (Theorem 26.2) provides a constructive proof of this result for the case where one considers the characteristic function.

The following example illustrates how the characterization result stated in Theorem 1.12 above can be exploited to find a particular probability distribution.

Example 1.18. *Let X be a continuous random variable with density function*

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad \text{for} \quad -\infty < x < \infty,$$

which corresponds to a Normal Distribution, and let us consider the transformation $Y = X^2$. Suppose that we are interested in obtaining the moment generating function of the random variable Y. We can compute

$$\phi_Y(t) = E[e^{tX^2}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-(\frac{1-2t}{2})x^2} dx.$$

To calculate the integral above, we can use the following change of variables

$$z = \left(\frac{\sqrt{1-2t}}{\sqrt{2}}\right)x, \quad dx = \frac{\sqrt{2}}{\sqrt{1-2t}}dz.$$

In this case, we obtain

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{(\frac{2t-1}{2})x^2}dx = \frac{1}{\sqrt{2\pi}}\frac{\sqrt{2}}{\sqrt{1-2t}}\int_{-\infty}^{+\infty} e^{-z^2}dz.$$

Now, we can make use of the identity of the Gaussian integral identity, *which states that* $\int_{-\infty}^{+\infty} e^{-z^2}dz = \sqrt{\pi}$, *to obtain*

$$\phi_Y(t) = (1-2t)^{-1/2} \qquad \textit{for } \ t < 1/2.$$

This particular moment generating function is known to corresponds to a continuous random variable with density function

$$h(y) = \frac{1}{\sqrt{2\pi y}}e^{-y/2}, \quad \textit{for} \quad y > 0,$$

which is the density of a distribution Chi-Square with parameter 1. This distribution was already obtained in Example 1.11.

1.4 Appendix: Combinatorics

Here I present several combinatorial formulas which are commonly used for counting the number of elements of a set. These methods are very useful to compute probabilities when the underlying set of elementary events is finite and all elementary events are equally likely. Under these conditions of equal likelihood, the probability of an event A is simply computed as $P(A) = |A| \,/\, |\Omega|$.

1.4.1 Ordered Samples and Permutations

Let us begin with a finite set $S = \{1, 2, \ldots, s\}$ of reference. We will propose our set of elementary events Ω, depending on the particular experiment of interest, taking the set S as a starting point. Suppose that we are interested in drawing a sequence of $m \leq s$ elements from the set S and that, in doing so, we care about the order of the draws. Then, the outcome of the draws can then be formally viewed as an m-tuple $\omega = (\omega_1, \omega_2, \ldots, \omega_m)$, where ω_i is the element in the ith draw. One option here is that we draw such a sequence ω by putting each drawn element back into the set before the next element is drawn. This procedure is referred to as *sampling with replacement*. Here, we have $\Omega = S^m$ so that $|\Omega| = s^m$. Another option requires that we do not return the elements into the set before the following draw. This procedure is commonly known as *sampling without replacement*. In this case, we have $\Omega = \left\{(\omega_i)_{i=1}^m : \omega_i \neq \omega_j \text{ for each } i \neq j\right\}$ so that $|\Omega| = s(s-1)(s-2)\cdots(s-m+1)$. When the sampling is without replacement and, in addition,

we care about the order of the draws, counting the elements $\omega \in \Omega$ is often referred to as counting *permutations*. Let $P_m^s = s(s-1)(s-2)\cdots(s-m+1)$ indicate the number of different possible m-tuples drawn when there is no replacement, or *permutations* of m elements out of s elements. Notice that, when the elements from S are drawn $m = s$ times without replacement, then we obtain $s! = P_s^s$ possible permutations as the outcomes of the experiment. In other words, we are computing the number of ways of obtaining an ordered subset (or tuple) of m elements from a set of $s \geq m$ elements.

In many applications, permutations are often associated with a the following type of problems. Suppose that we permute randomly $m \leq s$ elements from the set S among themselves and then ask about their final positions along some string. Here, we must identify each position i after the rearrangement with the component ω_i of ω which, in turn, corresponds to the element drawn from the set S. Also, notice that, since two distinct elements from S cannot end up in the same position, we are in fact considering random sampling without replacement. Consequently, the number of possible ways of distributing the s elements into the m final positions is given by P_m^s. As an example of this type of problems, suppose that we are interested in the event $A =$"$q \leq s$ pre-specified elements from the set S end up in m pre-specified positions along some string." Let Q be the pre-specified subset of elements of S. Given that q elements from S are required to end up in fixed positions, the number of tuples with $s-q$ coordinates that can be extracted without replacement from the set $S \setminus Q$ is $(s-q)!$. Therefore, $|A| = (s-q)!$ and the probability that q specified objects from S end up in q specified positions after permuting randomly among themselves the s distinct objects is

$$P(A) = \frac{(s-q)!}{s!} = \frac{1}{s(s-1)\cdots(s-q+1)} = \frac{1}{P_q^s}.$$

Permutations are also useful in problems where a random sample of size m is chosen from a set S of s distinct objects with replacement. In these case, we may ask about the probability of the event A="no element appears twice in the sample." Note that the cardinality of the set of elementary events in this problem is s^m. Also, the number of elementary events from the sample set where no element from S appears twice, out of the s^m possible elementary events, is nothing but the cardinality of

$$A = \left\{(\omega_i)_{i=1}^m \,:\, \omega_i \neq \omega_j \ \text{ for each } \ i \neq j\right\}.$$

But this is precisely the cardinality of the set of elementary events associated with an experiment of random sampling without replacement from that set S. Thus, the probability we are interested in can be computed as

$$P(A) = \frac{P_m^s}{s^m} = \left(1 - \frac{1}{s}\right)\left(1 - \frac{2}{s}\right)\cdots\left(1 - \frac{m-1}{s}\right). \tag{1.8}$$

A typical problem with this form is the one commonly refereed to as the "birthday problem."

Example 1.19. *Suppose that we wish to compute the probability of the event A ="no two people from a group of five friends have a common birthday." Let us ignore the leap years and make the quite unrealistic assumption that that birth rates are exactly equal likely over the year. Then, using the expression obtained in (1.8) above with $m = 5$ and $s = 365$, so we can easily compute*

$$P(A) = (1 - 1/365)(1 - 2/365)(1 - 3/365)(1 - 4/365).$$

The following example makes use of permutations as well.

Example 1.20. *Suppose that a committee of 5 members, consisting of a president, a secretary and three officials is to be selected from a club of 50 members. The officials will be ranked as official 1, 2 and 3, according to the degree of their importance within the club. The presidency and the secretary position are automatically assigned, respectively, to the oldest and the youngest members of the club. Then, the three officials are selected at random from the remaining 48 members of the club. Suppose that we wish to obtain the probability that three friends, Peter, Paul and Pierce, end up chosen, respectively, as official 1, 2 and 3. Notice that, since two pre-specified members of the set $\{1, \ldots, 50\}$ must end up in two pre-specified positions, there are P_3^{48} ways in which the three officials are selected, provided that the order of the sequence of size 3 matters. Therefore, the sought probability is $1/P_3^{48} = 1/(48 \cdot 47 \cdot 46)$.*

1.4.2 Combinations

In some problems, we are interested in computing the number of different subsets of size $m \leq s$ that can be extracted from the reference set S. In other words, we wish to compute the number of tuples that can be obtained under the restriction that the order of its coordinates does not matter. Here notice that there are P_m^s different sequences of size m that can be drawn from S without replacement. Also note that the elements of each set $M \subset S$ of m elements can be rearranged in $m!$ different ways. Then, since we wish to ignore the order in which the elements are selected, then these $m!$ reorderings of the elements of M should be considered as being the same object of interest. Therefore, there are $P_m^s/m!$ different samples of size m that can be drawn from S without replacement and regardless the order of its elements. These subsets are referred to as *combinations* of m elements from a set of s elements. Using the binomial operator, it is usual to write

$$\binom{s}{m} = \frac{P_m^s}{m!} = \frac{s!}{m!(s-m)!}.$$

To illustrate how combinations can be used in computing probabilities, let us consider again the experiment, already analyzed in Section 1.2 , where a coin is tossed n times and we wish to compute the probability of A ="at least one head shows up." A way to tackle this problem, different from the one proposed in Section 1.2, would require to consider the events A_i ="there shows up *exactly* i heads." Then, $A = \cup_{i=1}^n A_i$ with $A_i \cap A_j = \emptyset$ for each $i, j = 1, \ldots, n$ such that $i \neq j$. In this case, we

obtain $P(A) = \sum_{i=1}^{n} P(A_i)$. To compute each $P(A_i)$, notice that $\binom{n}{i}$ gives us the number of subsets of size i that can be extracted from $1, \ldots, n$, or, equivalently the cardinality of the event "i tosses result in head shows up while, at the same time, the remaining $n - i$ tosses shows up tails." This is precisely the cardinality of A_i. Therefore, we can compute

$$P(A) = \frac{\sum_{i=1}^{n} \binom{n}{i}}{2^n}.$$

The following examples deal with some of the concepts presented in this Appendix.

Example 1.21. *Suppose that some Economics department consists of* 8 *full professors,* 14 *associate professors, and* 18 *assistant professors. A committee of* 5 *is to be selected at random from the faculty of the department and suppose that we want to compute the probability that all the members of the committee are assistant professors. To answer this, notice first that in all there are* 40 *faculty members so that the committee of five can be chosen from the forty in* $\binom{40}{5}$ *possible combinations. Also, there are* 18 *assistant professors so that the committee of five can be chosen from them in* $\binom{18}{5}$ *possible ways. Therefore, we can compute the probability of our event of interest as* $\binom{18}{5} \Big/ \binom{40}{5}$.

Example 1.22. *Consider an experiment where a die is rolled 12 times. Suppose first that we are interested in computing the probability of getting exactly 2 fives, and let A denote that event of interest. Here notice that* $\Omega = \{1, \ldots, 6\}^{12}$ *so that* $|\Omega| = 6^{12}$. *Now consider the event* $A_{(i,j)}$, *with* $i, j = 1, \ldots, 12$, $i < j$, *which describes the outcome where number 5 shows up* only *in the ith and jth rolls. Then, we have* $\left|A_{(i,j)}\right| = 5^{10}$ *regardless of the value of the particular pair* (i, j). *Also, we know that* $A_{(i,j)} \cap A_{(k,l)} = \emptyset$ *whenever* $(i, j) \neq (k, l)$ *and*

$$A = \bigcup_{(i,j) \in Q} A_{(i,j)},$$

where Q is the set specified as

$$Q = \left\{(i, j) \in \{1, \ldots, 12\}^2 \, : \, i < j\right\}.$$

Therefore, we know that

$$P(A) = |Q| \, 5^{10}/6^{10}.$$

All that we need to do then is to compute the cardinality of set Q. Note that Q is nothing but the set of different pairs of numbers that can be extracted from $\{1, \ldots, 12\}$. *Therefore, its cardinality is given by* $\binom{12}{2}$ *and we thus obtain*

$$P(A) = \binom{12}{2} \frac{5^{10}}{6^{10}}.$$

Suppose now that we wish to compute the probability that at least 1 one shows up. Let B denote that event of interest and consider the event B_k, *where* $k = 1, 2, \ldots, 12$, *which describes the outcome such that number 1 shows up exactly k times. Then, we have* $B = \cup_{k=1}^{12} B_k$ *and* $B_k \cap B_l = \emptyset$ *whenever* $k \neq l$. *Therefore,*

we know that $P(B) = \sum_{k=1}^{12} P(B_k)$. *Following the same reasoning as above, we obtain*

$$P(B) = \frac{\sum_{k=1}^{12} \binom{12}{k} 5^{12-k}}{6^{12}}.$$

Example 1.23. *Suppose that n people throw their hats into a box and, after that, each person picks one hat from the box at random. Let us consider the events: A ="each person gets his own hat back," B ="the first m people who pick up a hat get their own hats back," and C ="everyone among the first m people who pick up a hat get a hat that belongs to someone of the last m people who pick up a hat." Suppose, in addition, that each hat thrown into the box has a probability* $p \in (0,1)$ *of getting dirty (this being unrelated to what happens to other hats or to who picks them). Consider the events D ="the first m people pick up clean hats" and E ="exactly m people pick up clean hats".*

To compute the probabilities of these events, note first that

$$\Omega = \left\{(\omega_1, \ldots, \omega_n) \in \{1, \ldots, n\} \,:\, \omega_i \neq \omega_j \ \ \forall\, i \neq j\right\}$$

so that $|\Omega| = n!$. *Then,* $P(A) = P(\{\omega\})$ *for* $\omega = (1, 2, \ldots, n)$ *without loss of generality, so that* $P(A) = 1/n!$. *For B, note that the number of ways of assigning the remaining* $n - m$ *hats after the first m hats have been assigned is* $(n - m)!$ *so that* $P(B) = (n - m)!/n!$. *Also, since there are* $m!$ *ways of assigning the first m hats among the first m people and* $(n - m)!$ *ways of assigning the remaining ones, we have* $P(C) = m!(n - m!)/n! = \binom{n}{m}^{-1}$.

For event D, we have that the probability that a person picks a clean hat is $(1 - p)$ *so that, by the independence assumption,* $P(D) = (1 - p)^m$. *As for event E, note first that, for each given group* $G \subset \{1, 2, \ldots, n\}$ *of m people, we can define the event* F_G*="every person* $i \in G$ *picks a clean hat while every person* $j \notin G$ *picks a dirty hat". Then, the events* $\{F_G\}_{G \subset \{1,\ldots,n\}}$ *satisfy* $F_G \cap F_{G'} = \emptyset$ *for each* $G \neq G'$. *Notice that* $P(F_G) = (1 - p)^m p^{n-m}$ *for any* $G \subset \{1, \ldots, n\}$. *Since there are* $\binom{n}{m}$ *of such events, we finally obtain*

$$P(E) = P(\cup_{G \subset \{1,\ldots,n\}} F_G) = \sum_{G \subset \{1,\ldots,n\}} P(F_G) = \binom{n}{m} (1 - p)^m p^{n-m}.$$

Example 1.24. *Suppose that a set of n balls is distributed randomly into n boxes and that we want to compute the probability that only the first box ends up being empty. Here, an elementary event must be identified with the final position of the balls so that* ω_i *should be interpreted as the box where the ith ball ends up. Then, the sample space is* $\Omega = \{1, \ldots, n\}^n$ *so that* $|\Omega| = n^n$. *Notice that we are considering random sampling with replacement since two different balls may end up in the same box. Consider the event A ="only box* 1 *ends up being empty." Notice that this can happen if and only if exactly one of the remaining* $n - 1$ *boxes contains two balls and all the other* $n - 2$ *boxes have exactly one ball each. Consider then the event* B_i *="box* 1 *ends up empty, box i ends up with two balls, and the remaining* $n - 2$ *boxes end up with exactly one ball each." We have* $A = \cup_{i=2}^{n} B_i$ *and* $B_i \cap B_j = \emptyset$ *whenever* $i \neq j$.

To compute $P(B_i)$, *notice first that the number of subsets that can be extracted from* $\{1, \ldots, n\}$ *containing two balls is* $\binom{n}{2}$. *Then, the remaining* $(n - 2)$ *balls can be rearranged in the remaining* $(n - 2)$ *boxes in*

$(n-2)!$ *different ways. Therefore, the number of distinct ways in which one can put no ball in box* 1, *two balls into box* i, *and exactly one ball in each of the remaining boxes is* $\binom{n}{2}(n-2)!$. *We obtain*

$$P(B_i) = \frac{\binom{n}{2}(n-2)!}{n^n},$$

so that the probability of our event of interest is

$$P(A) = \sum_{i=2}^{n} P(B_i) = \frac{(n-1)\binom{n}{2}(n-2)!}{n^n} = \frac{\binom{n}{2}(n-1)!}{n^n}.$$

Some combinatorial problems are of the following type. Suppose that a box contains r red balls and b black balls, and that a random sample of size m is drawn from the box without replacement. Here we want to compute the probability that this sample contains exactly k red balls and, therefore, $m-k$ black balls. The essence of this type of problem is that the total population can be partitioned into two classes. A random sample of a certain size is taken and we ask about the probability that the sample contains a specified number of elements of the two classes. First, notice that we are interested only in the number of red and black balls in the sample and not in the order in which these balls are drawn. Thus, we are dealing with sampling without replacement and without regard to order. Then, we can take as our sample space the family of all samples of size m drawn from a set of $b+r$ without replacement and without regard to order. As argued earlier, the probability that we must assign to each of these samples is

$$\binom{r+b}{m}^{-1}.$$

We need also count the number of ways in which a sample of size m can be drawn so as to have exactly k red balls. Notice that the k red balls can be chosen from the subset of r red balls in

$$\binom{r}{k}$$

possible ways without replacement and without regard to order, and the $m-k$ black balls can be chosen from the subset of b black balls in

$$\binom{b}{m-k}$$

ways without replacement and without regard to order. Since each choice of k red balls can be paired with each choice of $m-k$ black balls, there are a total of

$$\binom{r}{k}\binom{b}{m-k}$$

possible choices. Therefore, the probability of our event of interest can be computed as

$$\binom{r}{k}\binom{b}{m-k}\Big/\binom{r+b}{m}.$$

The following example makes use of the reasoning above.

Example 1.25. *We consider a box that contains r numbered balls and draw from it a random sample of size $n < r$ without replacement. We annotate the numbers of the balls and returned them to the box. Then, we take a second random sample of size $m < r$ without replacement as well. Suppose that we wish to compute the probability that the two samples have exactly l balls in common. To answer this, notice that we can consider that the first sample makes a partition of the set of balls into two classes, these n balls which were picked and these $r - n$ that were not. This problem then requires us simply to compute the probability that the sample of size m contains exactly l balls from the first class. So, the probability of our event of interest is*

$$\binom{n}{l}\binom{r-n}{m-l}\Big/\binom{r}{m}.$$

1.5 Appendix: Integration

Here I comment very briefly on the general concept of integral for the case where it is applied to random variables. Integration is the approach used in modern mathematics to compute areas and volumes, so that it provides naturally a tool to compute measures, in particular, the Lebesgue measure.

Consider a probability space $(\Omega, \mathcal{F}, P)$ and let us begin by taking a simple random variable X on $(\mathcal{F}, P)$ so that $\text{supp}(X) = \{x_1, x_2, \ldots, x_n\}$. Let us denote by $A_i = \{\omega \in \Omega \,|\, X(\omega) = x_i\} \in \mathcal{F}$, $i = 1, \ldots, n$, the events that correspond to the realizations of the random variable. Since it is simple, the random variable X admits the following representation:

$$X(\omega) = \sum_{i=1}^{n} x_i I_{A_i}(\omega),$$

where I_{A_i} is the indicator function of the set A_i, that is, $I_{A_i}(\omega) = 1$ if $\omega \in A_i$ and $I_{A_i}(\omega) = 0$ if $\omega \notin A_i$.

Then, the *integral of the simple variable X with respect to P* is

$$\int X(\omega) dP(\omega) = \int X(\omega) P(d\omega) = \sum_{i=1}^{n} x_i P(A_i).$$

The integration problem consists of enlarging this definition so that it may be applied to more general classes of random variables. One way of defining the general notion of integral requires

that we apply it to any bounded random variable X. Then, X is said to be *P-integrable* (or *P-summable*) if

$$\sup\left\{\int Y(\omega)dP(\omega) : Y \in S_P \text{ and } Y \leq X\right\} = \inf\left\{\int Y(\omega)dP : Y \in S_P \text{ and } X \leq Y\right\},$$

where S_P denotes the set of simple random variables on the probability space $(\Omega, \mathcal{F}, P)$. If it exists, the common value above is referred to as the *integral of X with respect to P*. The integral of X with respect to P is usually denoted either as $\int X dP$, $\int X(\omega)dP(\omega)$, or $\int X(\omega)P(d\omega)$. There is a number of different approaches to construct the abstract concept of integral. One of these approaches which is closely related to the notion of measure is that of *Lebesgue integral*. On the other hand, when one deals in calculus with Euclidean spaces, the most used approach is that of the *Riemann integral*.

1.6 Exercises

1.1. Let $A_1, A_2, \ldots$ be an infinite sequence of distinct subsets of some nonempty set Ω. Show by induction that
(a) $(\cup_{n=1}^{\infty} A_n)^c = \cap_{n=1}^{\infty} A_n^c$.
(b) $(\cap_{n=1}^{\infty} A_n)^c = \cup_{n=1}^{\infty} A_n^c$.

1.2. Let $\mathcal{F}$ be a family of subsets of some nonempty set Ω.
(a) Suppose that $\Omega \in \mathcal{F}$ and that $A, B \in \mathcal{F}$ implies $A \setminus B \in \mathcal{F}$. Show that $\mathcal{F}$ is an algebra.
(b) Suppose that $\Omega \in \mathcal{F}$ and that $\mathcal{F}$ is closed under the formation of complements and finite *disjoint* unions. Show that $\mathcal{F}$ need not be an algebra.

1.3. Let $\mathcal{F}_1, \mathcal{F}_2, \ldots$ be a family of subsets of some nonempty set Ω.
(a) Suppose that $\mathcal{F}_n$ are algebras satisfying $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$. Show that $\cup_{n=1}^{\infty} \mathcal{F}_n$ is an algebra.
(b) Suppose that $\mathcal{F}_n$ are σ-algebras satisfying $\mathcal{F}_n \subseteq \mathcal{F}_{n+1}$. Show by example that $\cup_{n=1}^{\infty} \mathcal{F}_n$ need not be a σ-algebra.

1.4. Let $\Omega = \{(x, y) \in \mathbb{R}^2 : 0 < x, y \leq 1\}$, let $\mathcal{F}$ be the family of sets of Ω of the form

$$\left\{(x, y) \in \mathbb{R}^2 : x \in A,\ 0 < y \leq 1\right\},$$

where $A \in \mathcal{B}_{(0,1]}$, and let $P(\{(x, y) \in \mathbb{R}^2 : x \in A,\ 0 < y \leq 1\}) = \lambda(A)$, where λ is Lebesgue measure on $\mathbb{R}$. Show that $(\Omega, \mathcal{F}, P)$ is a probability space.

1.5. Let $(\Omega, \mathcal{F}, P)$ be a probability space and, for $A \in \mathcal{F}$, let $P_A : \mathcal{F} \to [0, 1]$ be a set function defined by $P_A(B) = P(A \cap B)$ for each $B \in \mathcal{F}$.
(a) Show that, for a given $A \in \mathcal{F}$, P_A is a measure, but not a probability measure, on $(\Omega, \mathcal{F})$.
(b) Show that, for a given $A \in \mathcal{F}$ such that $P(A) > 0$, the set function Q_A on $\mathcal{F}$ defined by $Q_A(B) = P_A(B)/P(A)$ for each $B \in \mathcal{F}$ is a probability measure on $(\Omega, \mathcal{F})$.

1.6. Let $P_1, \ldots, P_n$ be probability measures on some measurable space $(\Omega, \mathcal{F})$. Show that $Q = \sum_{i=1}^n a_i P_i$, where $a_i \in \mathbb{R}_+$ for each $i = 1, \ldots, n$ and $\sum_{i=1}^n a_i = 1$, is a probability measure on $(\Omega, \mathcal{F})$.

1.7. Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $A_1, \ldots, A_n$ be events in $\mathcal{F}$ such that $P(\cap_{i=1}^k A_i) > 0$ for each $k = 1, \ldots, n-1$.
(a) Show that

$$P(\cap_{i=1}^n A_i) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \cdots P(A_n|A_1 \cap A_2 \cap \cdots \cap A_{n-1}).$$

(b) Show that if $P(\cap_{i=1}^k A_i) = 0$ for some $k \in \{1, \ldots, n-1\}$, then $P(\cap_{i=1}^n A_i) = 0$.

1.8. Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $A_1, \ldots, A_n$ be independent events in $\mathcal{F}$. Let $B_1, \ldots, B_n$ be another sequence of events such in $\mathcal{F}$ such that, for each $i = 1, \ldots, n$, either $B_i = A_i$ or $B_i = \Omega \setminus A_i$. Show that $B_1, \ldots, B_n$ are independent events.

1.9. There are three coins in a box. One is a two-headed coin, another is a two-tailed coin, and the third is a fair coin. One of the three coins is chosen at random and flipped. It shows heads. What is the probability that it is the two-headed coin?

1.10. Two dice are rolled once and the 36 possible outcomes are equally likely. Compute the probability that the sum of the numbers on the two faces is even.

1.11. A box has 10 numbered balls. A ball is picked at random and then a second ball is picked at random from the remaining 9 boxes. Compute the probability that the numbers on the two selected balls differ by two or more.

1.12. A box has 10 balls, 6 of which are black and 4 of which are white. Three balls are removed at random from the box, but their colors are not noted.
(a) Compute the probability that a fourth ball removed from the box is white.
(b) Suppose now that it is known that at least one of the three removed balls is black. Compute the probability that all three of the removed balls are black.

1.13. A box has 5 numbered balls. Two balls are drawn independently from the box with replacement. It is known that the number on the second ball is at least as large as the number on the first ball. Compute the probability that the number on the first ball is 2.

1.14. Let $(\Omega, \mathcal{F}, P)$ be a probability space and let A, B and C be three events in $\mathcal{F}$ such that $P(A \cap B \cap C) > 0$. Show that $P(C|A \cap B) = P(C|B)$ implies $P(A|B \cap C) = P(A|B)$.

1.15. Two points are randomly chosen from the interval $[0, 1]$. Compute the probability that the length of each of the three segments formed in this way be above $1/4$.

1.16. Given the digits 1, 2, 3, 4, and 5, how many four-digit numbers can be formed if
(a) there is no repetition;
(b) there can be repetition;
(c) the number must be even and there is no repetition;
(d) if the digits 2 and 3 must appear in that order in the number and there is no repetition.

1.17. A bridge deck has 52 cards dividend into 4 suits of 13 cards each: hearts, spades, diamonds, and clubs. Compute the probability that, when drawing 5 cards form a bridge deck (a poker hand),
(a) all of them are diamonds;
(b) one card is a diamond, one a spade, and the other three are clubs;
(c) exactly two of them are hearts if it is known that four of them are either hearts or diamonds;
(d) none of them is a queen;
(e) exactly two of them are kings;
(f) exactly three of them are of the same suit.

1.18. In a hand of 13 cards drawn from a bridge deck, compute the probability of getting exactly 5 clubs, 3 diamonds, 4 hearts, and 1 spade.

1.19. A man has 8 keys one of which fits the lock. He tries the keys one at a time, at each attempt choosing at random from the keys that were not tried earlier. Find the probability that the 6th key tried is the correct one.

1.20. A set of n balls is distributed at random into n boxes. Compute the probabilities of the following events:
(a) exactly one box is empty;
(b) only one box is empty if it is known that box 1 is empty;
(c) box 1 is empty if it is known that only one box is empty.

1.21. Suppose that n balls are distributed at random into r boxes. Compute the probability that the box 1 contains exactly k balls, where $0 \leq k \leq n$.

1.22. A group of 3 balls are drawn simultaneously from a box that contains 10 numbered balls. Compute the probability that balls 1 and 4 are among the three picked balls.

1.23. A random sample of size n is drawn from a set of s elements. Compute the probability that none of k pre-specified elements is in the sample if the method used is:

(a) sampling without replacement;
(b) sampling with replacement.

1.24. A set of n objects are permuted among themselves. Show that the probability that k pre-specified objects occupy k pre-specified positions is $(n-k)!/n!$.

1.25. Two boxes contains n numbered balls each. A random sample of $k \leq n$ is drawn without replacement from each box. Compute the probability that the samples contain exactly l balls having the same numbers in common.

1.26. Show that, for two positive integers s and n such that $s \geq n$, we have

$$\left(1 - \frac{n-1}{s}\right)^{n-1} \leq \frac{(s)_n}{s^n} \leq \left(1 - \frac{1}{s}\right)^{n-1},$$

where $(s)_n = s(s-1)\cdots(s-n+1)$.

1.27. A die is rolled 12 times. Compute the probability of getting at most 3 fours.

1.28. Let X be a random variable on some probability space $(\Omega, \mathcal{F}, P)$ and let $g : \mathbb{R} \to \mathbb{R}$ be a one-to-one function. Show that $Y = g(X)$ is a random variable on $(\Omega, \mathcal{F}, P)$.

1.29. Let $F_1, \ldots, F_n$ be distribution functions on some probability space $(\Omega, \mathcal{F}, P)$. Show that $G = \sum_{i=1}^n a_i F_i$, where $a_i \in \mathbb{R}_+$ for each $i = 1, \ldots, n$ and $\sum_{i=1}^n a_i = 1$, is a distribution function on $(\Omega, \mathcal{F}, P)$.

1.30. Let X be an absolutely continuous random variable on some probability space $(\Omega, \mathcal{F}, P)$ with density $f(x) = 1/2e^{-|x|}$ for $x \in \mathbb{R}$. Compute $P[X \geq 0]$, $P[|X| \leq 2]$, and $P[1 \leq |X| \leq 2]$.

1.31. Any point in the interval $[0, 1)$ can be represented by its decimal expansion $.x_1x_2\ldots$. Suppose that a point is chosen at random from the interval $[0, 1)$. Let X be the first digit in the decimal expansion representing the point. Compute the density of X considered as a random variable on some probability space.

1.32. A box contains 6 red balls and 4 black balls. A random size of n balls is drawn from the box. Let X be the number of red balls picked. Compute the density of X, considered as a random variable on some probability space, if the sampling is without replacement.

1.33. Let n be a positive integer and let h be a real-valued function defined by

$$h(x) = \begin{cases} c2^x & \text{if } x = 1, 2, \ldots n, \\ 0 & \text{otherwise.} \end{cases}$$

Find the value of c such that h is a discrete density function on some probability space.

1.34. Let X be a discrete random variable on some probability space with support

$$\{-3, -1, 0, 1, 2, 3, 5, 8\}$$

and discrete density function f specified by $f(-3) = .2$, $f(-1) = .15$, $f(0) = .2$, $f(1) = .1$, $f(2) = .1$, $f(3) = .15$, $f(5) = .05$, and $f(8) = .05$. Compute the following probabilities:
(a) X is negative;
(b) X is even;
(c) X takes a value between 1 and 5 inclusive;
(d) $P[X = -3|X \leq 0]$;
(e) $P[X \geq 3|X > 0]$.

1.35. A box contains 12 numbered balls. Two balls are drawn with replacement from the box. Let X be the larger of the two numbers on the balls. Compute the density of X considered as a random variable on some probability space.

1.36. Let X be a random variable on some probability space $(\Omega, \mathcal{F}, P)$ such that $P[|X - 1| = 2] = 0$. Express $P[|X - 1| \geq 2]$ in terms of the distribution function F of X.

1.37. Show that the distribution function F of a random variable is continuous from the right and that

$$\lim_{x \to -\infty} F(x) = 0 \quad \text{and} \quad \lim_{x \to +\infty} F(x) = 1.$$

1.38. A point is chosen at random from the interior of a sphere of radius r. Each point in the sphere is equally likely of being chosen. Let X be the square of the Euclidean distance of the chosen point from the center of the sphere. Find the distribution function of X considered as a random variable on some probability space.

1.39. The distribution function F of some random variable X on some probability space is defined by

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ 1 - e^{-\lambda x} & \text{if } x > 0, \end{cases}$$

where $\lambda > 0$. Find a number m such that $F(m) = 1/2$.

1.40. Let X be a random variable (on some probability space) with distribution function

$$F(x) = \begin{cases} 0 & \text{if } x < 0, \\ x/3 & \text{if } 0 \le x < 1, \\ x/2 & \text{if } 1 \le x < 2, \\ 1 & \text{if } x \ge 2. \end{cases}$$

Compute the following probabilities:
(a) $P[1/2 \le X \le 3/2]$;
(b) $P[1/2 \le X \le 1]$;
(c) $P[1/2 \le X < 1]$;
(d) $P[1 \le X \le 3/2]$;
(e) $P[1 < X < 2]$.

1.41. The distribution function F of some random variable X (on some probability space) is defined by

$$F(x) = \frac{1}{2} + \frac{x}{2(|x|+1)}, \quad x \in \mathbb{R}.$$

Find a density function f for F. At what points x will $F'(x) = f(x)$?

1.42. Let X be an absolutely continuous random variable with density f. Find a formula for the density of $Y = |X|$.

1.43. Let X be a positive absolutely continuous random variable with density f. Find a formula for the density of $Y = 1/(X+1)$.

1.44. Let T be a positive absolutely continuous random variable on some probability space $(\Omega, \mathcal{F}, P)$. Let T denote the failure date of some system. Let F be the distribution function of T, and assume that $F(t) < 1$ for each $t > 0$. Then, we can write $F(t) = 1 - e^{-G(t)}$ for some one-to-one function $G : \mathbb{R}_{++} \to \mathbb{R}_{++}$. Assume also that $G'(t) = g(t)$ exists for each $t > 0$.
(a) Show that T has density f satisfying

$$\frac{f(t)}{1 - F(t)} = g(t), \quad t > 0.$$

(b) Show that for $s, t > 0$,

$$P[T > t + s | T > t] = e^{-\int_t^{t+s} g(m)dm}.$$

1.45. Compute the density functions of the following transformations $Y = g(X)$:
(a) $f(x) = \frac{1}{2}e^{-|x|}$, $x \in \mathbb{R}$, with $g(X) = |X|^3$;

(b) $f(x) = \frac{3}{8}(x+1)^2$, $x \in (-1,1)$, with $g(X) = 1 - X^2$;
(c) $f(x) = \frac{3}{8}(x+1)^2$, $x \in (-1,1)$, with $g(X) = 1 - X^2$ for $X \leq 0$ and $g(X) = 1 - X$ for $X > 0$.

1.46. Let (x, y) be a point randomly chosen from the square $[0,1]^2$ and let X be the random variable which assigns the number $x + y$ to the poing (x, y). Compute the distribution function of X.

1.47. Let be a (X_1, X_2) random vector with joint density

$$f(x_1, x_2) = \frac{1}{2\pi} e^{-\frac{1}{2}(x_1^2 + x_2^2)}, \qquad x_1, x_2 > 0.$$

Consider the transformation g to polar coordinates so that $T = g^{-1}$ is given by

$$(x_1, x_2) = T(y_1, y_2) = (y_1 \cos y_2, y_1 \sin y_2),$$

and $g(\mathbb{R}_{++}) = \left\{(y_1, y_2) \in \mathbb{R}^2 \,:\, y_1 > 0,\, 0 < y_2 < 2\pi\right\}$. Let h denote the joint density of (Y_1, Y_2), and let h_1 and h_2 be the marginal densities of Y_1 and Y_2, respectively. Show that
(a) $h(y_1, y_2) = (2\pi)^{-1} y_1 e^{-y_1^2/2}$,
(b) $h_1(y_1) = y_1 e^{-y_1^2/2}$,
(c) $h_2(y_2) = (2\pi)^{-1}$.

1.48. Let X and Y be two absolutely continuous random variables whose respective densities, given two numbers $\sigma, \tau > 0$,

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}$$

and

$$l(y) = \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{y^2}{2\tau^2}}$$

are supported in $\mathbb{R}$. Show that if X and Y are independent, then $S = X + Y$ has density

$$m(s) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2 + \tau^2}} e^{-\frac{s^2}{2(\sigma^2+\tau^2)}},$$

supported in $\mathbb{R}$.

1.49. Suppose that X and Y are independent absolutely continuous random variables. Derive formulas for the joint density for $(X + Y, X)$, the density of $X + Y$, and the density of $Y - X$.

1.50. Let X and Y be absolutely continuous random variables with joint distribution function F and joint density f. Find the joint distribution function and the joint density of the random variables $W = X^2$ and $Z = Y^2$. Show that if X and Y are independent, then W and Z are independent too.

1.51. Let X and Y be two independent absolutely continuous random variables (on some probability space $(\Omega, \mathcal{F}, P)$) having the same density each, $f(x) = g(y) = 1$ for $x, y \in (0, 1]$. Find

(a) $P[|X - Y| \leq .5]$;

(b) $P\left[\left|\frac{X}{Y} - 1\right| \leq .5\right]$;

(c) $P[Y \geq X | Y \geq 1/3]$.

1.52. Let X and Y be absolutely continuous random variables with joint density

$$f(x, y) = \begin{cases} \rho^2 e^{-\rho y} & \text{if } 0 \leq x \leq y, \\ 0 & \text{otherwise,} \end{cases}$$

where $\rho > 0$. Find the marginal density of X and Y. Find the joint distribution function of X and Y.

1.53. Let $f(x, y) = ce^{-(x^2 - xy + 4y^2)/2}$ for $x, y \in \mathbb{R}$ How should c be chosen to make f a joint density for two random variables X and Y? Find the marginal densities of f.

1.54. Let X, Y and Z be absolutely continuous random variables with joint density

$$f(x, y, z) = \begin{cases} c & \text{if } x^2 + y^2 + z^2 \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

How should c be chosen to make f indeed a joint density of X, Y and Z. Find the marginal density of X. Are X, Y and Z independent?

1.55. Let X be an absolutely continuous random variable with density $f(x) = 1/2$ for $x \in (-1, 1]$. Let $Y = X^2$. Show that X and Y are uncorrelated but not independent.

1.56. Let (X_1, X_2) be a random vector. Using the concept of moment generating function, show that

$$\text{Cov}[X_1, X_2] = \frac{\partial^2 \Phi_{(X_1, X_2)}(0, 0)}{\partial t_1 \partial t_2} - \frac{\partial \Phi_{(X_1, X_2)}(0, 0)}{\partial t_1} \cdot \frac{\partial \Phi_{(X_1, X_2)}(0, 0)}{\partial t_2}.$$

1.57. Let (X, Y) be an absolutely continuous random vector with joint density

$$f(x, y) = \frac{1}{2\pi \sigma_x \sigma_y \sqrt{1 - \rho^2}} \exp\{Q\},$$

where

$$Q = -\frac{1}{2(1 - \rho^2)} \left[\frac{(x - \mu_x)^2}{\sigma_x^2} + \frac{(y - \mu_y)^2}{\sigma_y^2} - 2\rho \frac{(x - \mu_x)(y - \mu_y)}{\sigma_x \sigma_y} \right].$$

Show that

$$f(x|y) = \frac{1}{\sqrt{2\pi}\sqrt{(1-\rho^2)\sigma_x^2}} \exp\left\{-\frac{1}{2(1-\rho^2)\sigma_x^2}\left[(x-\mu_x) - \rho\frac{\sigma_x}{\sigma_y}(y-\mu_y)\right]^2\right\}.$$

1.58. Let X be a random variable on some probability space $(\Omega, \mathcal{F}, P)$ which takes only the values $0, 1, 2, \ldots$. Show that $E[X] = \sum_{n=1}^{\infty} P[X \geq n]$.

1.59. Let X be an absolutely continuous random with $\text{supp}(X) = [0, b]$, where $b > 0$, with distribution function F, and with density function f. Show that

$$E[X] = \int_0^b [1 - F(x)]\, dx.$$

1.60. Let X and Y be random variables with joint density

$$f(x, y) = \begin{cases} c & \text{if } x^2 + y^2 \leq 1 \\ 0 & \text{if } x^2 + y^2 > 1. \end{cases}$$

Find the conditional density of X given $Y = y$ and compute the conditional expected value $E[X|Y = y]$.

1.61. Let $X_1, \ldots, X_n$ be independent random variables having a common density with mean μ and variance σ^2. Set $\overline{X}_n = (X_1 + \cdots + X_n)/n$.
(a) By writing $X_k - \overline{X}_n = (X_k - \mu) - (\overline{X}_n - \mu)$, show that

$$\sum_{k=1}^{n}(X_k - \overline{X}_n)^2 = \sum_{k=1}^{n}(X_k - \mu)^2 - n(\overline{X}_n - \mu)^2.$$

(b) From (a) obtain

$$E\Big[\sum_{k=1}^{n}(X_k - \overline{X}_n)^2\Big] = (n-1)\sigma^2.$$

1.62. Let X and Y be two random variables (on some probability space $(\Omega, \mathcal{F}, P)$) such that $P[|X - Y| \leq a] = 1$ for some constant $a \in \mathbb{R}$. Show that $|E[X] - E[Y]| \leq a$.

1.63. Show that $\text{Var}[aX] = a^2\text{Var}[X]$ for any random variable X and constant $a \in \mathbb{R}$.

1.64. Let X and Y be absolutely continuous random variables with joint density

$$f(x, y) = \begin{cases} \rho^2 e^{-\rho y} & \text{if } 0 \leq x \leq y, \\ 0 & \text{otherwise,} \end{cases}$$

where $\rho > 0$. Find the conditional density $f(y|x)$.

1.65. Let X and Y be absolutely continuous random variables with joint density

$$f(x, y) = ce^{-(x^2 - xy + y^2)/2},$$

for each $x, y \in \mathbb{R}$. Find the conditional expected value of Y given $X = x$.
Hint: Use the Gaussian integral identity: $\int_{-\infty}^{+\infty} e^{-z^2} dz = \sqrt{\pi}$.

1.66. Let X and Y be two absolutely continuous random variables with joint density

$$f(x, y) = \begin{cases} n(n-1)(y-x)^{n-2} & \text{if } 0 \leq x \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Find the conditional expected value of X given $Y = y$.

Chapter 2

A Few Distributions of Interest in Economics

2.1 Discrete Distributions

We begin with the binomial distribution.

6.1.1. *The binomial distribution*

A **Bernoulli trial** is a random experiment with two possible mutually exclusive outcomes. Without loss of generality we can call these outcomes "success" and "failure" (e.g., defective or non-defective, female or male). Denote by $p \in (0,1)$ the probability of success. A sequence of independent Bernoulli trials, in the sense that the outcome of any trial does not affect the outcome of any other trial, are called **binomial or Bernoulli trials**.

Let X be the random variable associated with the number of successes in the n trials. The number of ways of selecting x successes out of n trials is $\binom{n}{x}$. Since trials are independent and the probability of each of these ways is $p^x(1-p)^{n-x}$, the discrete density function of X is given by

$$f(x) = P[X = x] = \binom{n}{x} p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, 2, \ldots, n.$$

Recall that this density function was obtained earlier in example 18. The probability distribution of X is called **binomial distribution** and we write $X \sim b(n,p)$. Using the fact that, for a positive integer n, $(a+b)^n = \sum_{x=0}^{n} \binom{n}{x} b^x a^{n-x}$, we can obtain

$$\Phi_X(t) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x} = [(1-p) + pe^t]^n.$$

Then,

$$\Phi'_X(t) = n[(1-p) + pe^t]^{n-1} pe^t$$

and

$$\Phi''_X(t) = n(n-1)[(1-p)+pe^t]^{n-2}p^2e^{2t} + n[(1-p)+pe^t]^{n-1}pe^t.$$

It follows that

$$E[X] = m_1(X) = \Phi'(0) = n[1-p+p]^{n-1}p = np$$

and

$$\begin{aligned}\text{Var}[X] &= m_2(X) - m_1^2(X) = \Phi''_X(0) - \big(E[X]\big)^2 \\ &= n(n-1)[1-p+p]^{n-2}p^2 + np - n^2p^2 \\ &= n^2p^2 - np^2 + np - n^2p^2 \\ &= np(1-p).\end{aligned}$$

Consider now the special case of a Bernoulli distribution that one obtains when $n = 1$. Then, X is the random variable associated with the outcome of a single Bernoulli trial so that $X(\text{success}) = 1$ and $X(\text{failure}) = 0$. The probability distribution of X is called **Bernoulli distribution**. We write $X \sim b(1,p)$ and the discrete density function of X is

$$f(x) = P[X = x] = p^x(1-p)^{1-x} \quad \text{for } x = 0, 1.$$

One can easily compute

$$E[X] = (0)(1-p) + (1)(p) = p;$$

$$\text{Var}[X] = (0-p)^2(1-p) + (1-p)^2(p) = p(1-p);$$

$$\Phi_X(t) = e^{t(0)}(1-p) + e^{t(1)}(p) = 1 + p(e^t - 1).$$

Notice that the binomial distribution can be also considered as the distribution of the sum of n independent, identically distributed $X_i \sim b(1,p)$ random variables. For a sequence of n Bernoulli trials, let X_i be the random variable associated with the outcome of the ith trial so that $X_i(\text{success}) = 1$ and $X_i(\text{failure}) = 0$. Clearly, the number of successes is given by $X = X_1 + \cdots + X_n$. Following this approach, we have

$$E[X] = \sum_{i=1}^{n} E[X_i] = np$$

and

$$\text{Var}[X] = \text{Var}\Big[\sum_{i=1}^{n} X_i\Big] = np(1-p).$$

Theorem 2.1. *Let $X_i \sim b(n_i, p)$, $i = 1, \ldots, k$, be independent random variables. Then,*

$$Y_k = \sum_{i=1}^{k} X_i \sim b(\sum_{i=1}^{k} n_i, p).$$

Corollary 2.1. *Let $X_i \sim b(n,p)$, $i = 1,\dots,k$, be independent random variables. Then,*

$$Y_k = \sum_{i=1}^{k} X_i \sim b(kn,p).$$

This result has already been demonstrated in Example 1.7.

6.1.2. *The negative binomial distribution*

Consider now a sequence (maybe infinite) of Bernoulli trials and let X be the random variable associated to the number of failures in the sequence before the rth success, where $r \geq 1$. Then, $X + r$ is the number of trials necessary to produce exactly r successes. This will happen if and only if the $(X + r)$th trial results in a success and among the previous $(X + r - 1)$ trials there are exactly X failures or, equivalently, $r - 1$ successes. We remark that we need to take into account the probability that the $(X + r)$th trial results in a success. It follows by the independence of trials that

$$f(x) = P[X = x] = \binom{x + r - 1}{x} p^r (1 - p)^x = \binom{x + r - 1}{r - 1} p^r (1 - p)^x \quad \text{for } x = 0, 1, 2, \ldots$$

We say that the random variable X has **negative binomial distribution** and write $X \sim NB(r, p)$. For the special case given by $r = 1$, we say that X has **geometric distribution** and write $X \sim G(p)$. For the negative binomial distribution, we have

$$\Phi_X(t) = p^r [1 - (1 - p)e^t]^{-r};$$

$$E[X] = r(1 - p)/p;$$

$$\text{Var}[X] = r(1 - p)/p^2.$$

Example 2.1. *Suppose that a mathematician carries two matchboxes, box 1 and box 2, containing k matches each. Each time he needs a match, he is equally likely to take it from either box. Suppose that at a certain moment he reaches into box 1 and discovers that it is empty. Then, what is the probability that there remains exactly $r \leq k$ matches in box 2? We can identify "a match is taken from box 1" as failure and "a match is taken from box 2" as success. Thus, we have a sequence of Bernoulli trials with $p = 1/2$. Note that, right before the moment at which box 1 is empty and box 2 has r matches, there has been $k + k - r$ trials, k of which have been failures and $k - r$ of which have been successes. Then if X is the random variable which identifies the number of failure before the $(k - r)$-th successes, we know that $X \sim NB(k - r, 1/2)$. Therefore*

$$P[X = k] = \binom{2k - r}{k} (1/2)^{k-r} (1/2)^k = \binom{2k - r}{k} (1/2)^{2k-r} \quad \text{for } r = 0, 1, \ldots, k.$$

Note that, in contrast with the density specification given above for the negative binomial distribution, we are now interested exactly in the $(k - r)$-th success. Hence, the probability now obtained is slightly different to what one would obtain applying directly the formula above.

Example 2.2. *Consider two independent geometric random variables $X, Y \sim G(p)$ and suppose that we wish to compute the probability $P[X = m \mid X + Y = n]$ for $m \in \{1, 2, \ldots, n - 1\}$. First note that, by the definition of conditional probability, we have*

$$P[X = m \mid X + Y = n] = \frac{P(\{X = m\} \cap \{X + Y = n\})}{P(\{X + Y = n\})}.$$

Now, the event $\{X = m\} \cap \{X + Y = m\}$ *in the numerator is equivalent to* $\{X = m\} \cap \{Y = n - m\}$ *and, since* X *and* Y *are independent, we have*

$$P(\{X = m\} \cap \{Y = n - m\}) = P[X = m]P[Y = n - m].$$

As for the denominator, note that we can use the total probability law, together with the independence of X *and* Y*, to obtain:*

$$\begin{aligned} P(\{X + Y = n\}) &= \sum_{m'=1}^{n-1} P[X + Y = n \mid X = m']P[X = m'] \\ &= \sum_{m'=1}^{n-1} P[Y = n - m' \mid X = m']P[X = m'] \\ &= \sum_{m'=1}^{n-1} P[Y = n - m']P[X = m']. \end{aligned}$$

Then, since X *and* Y *have (identical) geometric distributions, we have:*

$$\begin{aligned} P[X = m \mid X + Y = n] &= \frac{P[X = m]P[Y = n - m]}{\sum_{m'=1}^{n-1} P[Y = n - m']P[X = m']} \\ &= \frac{p(1-p)^m p(1-p)^{n-m}}{\sum_{m'=1}^{n-1} p(1-p)^{n-m'}p(1-p)^{m'}} \\ &= \frac{(1-p)^n}{\sum_{m'=1}^{n-1}(1-p)^n} = \frac{1}{n-1}. \end{aligned}$$

6.1.3. *The multinomial distribution*

The binomial distribution is generalized in a natural way to the **multinomial distribution** as follows. Suppose that a random experiment is repeated n independent times. Each repetition of the experiment results in on of k mutually exclusive and exhaustive events $A_1, A_2, \ldots, A_k$. Let p_i be the probability that the outcome (of any repetition) is an element of A_i and assume that each p_i remains constant throughout the n repetitions. Let X_i be the random variable associated with the number of outcomes which are elements of A_i. Also, let $x_1, x_2, \ldots, x_{k-1}$ be nonnegative numbers such that $x_1 + x_2 + \cdots + x_{k-1} \leq n$. Then, the probability that exactly x_i outcomes terminate in A_i, $i = 1, 2, \ldots, k-1$, and, therefore, $x_k = n - (x_1 + x_2 + \cdots + x_{k-1})$ outcomes terminate in A_k is

$$P[X_1 = x_1, \ldots, X_k = x_k] = \frac{n!}{x_1!x_2!\cdots x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}.$$

This is the joint discrete density of a **multinomial distribution**.

6.1.4. *The Poisson distribution*

Recall that, for each $r \in \mathbb{R}$, we have

$$e^r = 1 + r + \frac{r^2}{2!} + \frac{r^3}{3!} + \cdots = \sum_{x=0}^{\infty} \frac{r^x}{x!}.$$

Then, given $r > 0$, consider the function $f : \mathbb{R} \to \mathbb{R}_+$ defined by

$$f(x) = \frac{r^x e^{-r}}{x!} \quad \text{for } x = 0, 1, 2, \ldots$$

One can check that

$$\sum_{x=0}^{\infty} f(x) = e^{-r} \sum_{x=0}^{\infty} \frac{r^x}{x!} = e^{-r} e^r = 1.$$

Hence, f satisfies the conditions required for being a discrete density function. The distribution associated to the density function above is known as the **Poisson distribution** and, for a random variable X that follows such distribution, we write $X \sim P(r)$. Empirical evidence indicates that the Poisson distribution can be used to analyze a wide class of applications. In those applications one deals with a process that generates a number of changes (accidents, claims, etc.) in a fixed interval (of time or space). If a process can be modeled by a Poisson distribution, then it is called a **Poisson process**. Examples of random variables distributed according to the Poisson distributions are: (1) X indicates the number of defective goods manufactured by a productive process in a certain period of time, (2) X indicates the number of car accidents in a unit of time, and so on. For $X \sim P(r)$, we have

$$E[X] = \text{Var}[X] = r$$

and

$$\begin{aligned} \Phi_X(t) &= \sum_{x=0}^{\infty} e^{tx} \frac{r^x e^{-r}}{x!} = e^{-r} \sum_{x=0}^{\infty} \frac{(re^t)^x}{x!} \\ &= e^{-r} e^{re^t} = e^{r(e^t - 1)}. \end{aligned}$$

Theorem 2.2. *Let $X_i \sim P(r_i)$, $i = 1, \ldots, k$, be independent random variables. Then,*

$$S_k = \sum_{i=1}^{k} X_i \sim P(r_1 + \cdots + r_k).$$

The following results relate the Poisson with the binomial distribution.

Theorem 2.3. *Let $X \sim P(r_x)$ and $Y \sim P(r_y)$ be independent random variables. Then the conditional distribution of X given $X + Y$ is binomial. In particular, $(X|X + Y = n) \sim b(n, \frac{r_x}{r_x+r_y})$ (that is, for a sequence of n Bernoulli trials). Conversely, let X and Y are independent nonnegative integer-valued random variables with strictly positive densities. If $(X|X + Y = n) \sim b(n, p)$, then $X \sim P(\theta p/(1 - p))$ and $Y \sim P(\theta)$ for an arbitrary $\theta > 0$.*

Theorem 2.4. *If $X \sim P(r)$ and $(Y|X = x) \sim b(x, p)$, then $Y \sim P(rp)$.*

2.2 Continuous Distributions

In this section we introduce some of the most frequently used absolutely continuous distributions and describe their properties.

6.2.1. *The uniform distribution*

A random variable X is said to have **uniform distribution** on the interval $[a, b]$ if its density function is given by

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b; \\ 0 & \text{otherwise.} \end{cases}$$

We write $X \sim U[a, b]$. Intuitively, the uniform distribution is related to random phenomena where the possible outcomes have the same probability of occurrence. One can easily obtain that

$$F(x) = \begin{cases} 0 & \text{if } x \le a; \\ \frac{x-a}{b-a} & \text{if } a < x \le b; \\ 1 & \text{if } x > b. \end{cases}$$

$$E[X] = \frac{a+b}{2}, \quad \text{Var}[X] = \frac{(b-a)^2}{12}, \quad \text{and} \quad \Phi_X(t) = \frac{e^{tb} - e^{ta}}{t(b-a)}.$$

Example 2.3. *Let X be a random variable with density*

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \textit{if } x > 0; \\ 0 & \textit{otherwise,} \end{cases}$$

where $\lambda > 0$. One can easily obtain

$$F(x) = \begin{cases} 0 & \textit{if } x \le 0; \\ 1 - e^{-\lambda x} & \textit{if } x > 0. \end{cases}$$

Consider the transformation $Y = F(X) = 1 - e^{-\lambda X}$. *We note then:* $x = T(y) = -\ln(1-y)/\lambda$ *and* $T'(y) = 1/\lambda(1-y)$ *so that the density of* Y *is given by*

$$\begin{aligned} h(y) &= f(T(y))\,|T'(y)| \\ &= \lambda e^{-\lambda\left(-\ln(1-y)/\lambda\right)} \frac{1}{\lambda(1-y)} = 1 \end{aligned}$$

for $0 \leq y < 1$.

So, is it a mere coincidence that in the example above $F(X)$ is uniformly distributed on the interval $[0,1]$? The following theorem answers this question and provides us with a striking result about the uniformity of the distribution of any distribution function.

Theorem 2.5. *Let* X *be a random variable with a continuous distribution function* F. *Then* $F(X)$ *is uniformly distributed on* $[0,1]$. *Conversely, let* F *be any distribution function and let* $X \sim U[0,1]$. *Then, there exists a function* $g : [0,1] \to \mathbb{R}$ *such that* $g(X)$ *has* F *as its distribution function, that is,* $P[g(X) \leq x] = F(x)$ *for each* $x \in \mathbb{R}$.

6.2.2. *The* Γ, χ^2, *and Beta distributions*

It is a well known result that the integral

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$$

yields a finite positive number for $\alpha > 0$. Integration by parts gives us

$$\Gamma(\alpha) = (\alpha-1)\int_0^\infty y^{\alpha-2} e^{-y} dy = (\alpha-1)\Gamma(\alpha-1).$$

Thus, if α is a positive integer, then

$$\Gamma(\alpha) = (\alpha-1)(\alpha-2)\cdots(2)(1)\Gamma(1) = (\alpha-1)!.$$

Let us now consider another parameter $\beta > 0$ and introduce a new variable by writing $y = x/\beta$. Then, we have

$$\Gamma(\alpha) = \int_0^\infty \left(\frac{x}{\beta}\right)^{\alpha-1} e^{-\frac{x}{\beta}} \left(\frac{1}{\beta}\right) dx$$

Therefore, we obtain

$$1 = \int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} dx.$$

Hence, since $\Gamma(\alpha), \alpha, \beta > 0$, we see that

$$f(x) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} & \text{if } x > 0; \\ 0 & \text{otherwise} \end{cases}$$

is a density function of an absolutely continuous random variable. A random variable X with the density above is said to have the **gamma distribution** and we write $X \sim \Gamma(\alpha, \beta)$. The special case when $\alpha = 1$ yields the **exponential distribution** with parameter β. In that case, we write $X \sim exp(\beta) \equiv \Gamma(1, \beta)$ and the corresponding density function is, therefore,

$$f(x) = \begin{cases} \frac{1}{\beta} e^{-x/\beta} & \text{if } x > 0. \\ 0 & \text{otherwise} \end{cases}$$

The gamma distribution is often used to model waiting times.

The distribution function associated to a gamma distribution is

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0; \\ \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^x y^{\alpha-1} e^{-y/\beta} dy & \text{if } x > 0. \end{cases}$$

The corresponding moment generating function is obtained as follows. First,

$$\begin{aligned} \Phi_X(t) &= \int_0^\infty e^{tx} \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} dx \\ &= \int_0^\infty \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x(1-\beta t)/\beta} dx. \end{aligned}$$

Second, by setting $y = x(1 - \beta t)/\beta$ or, equivalently,

$$x = \frac{\beta y}{1 - \beta t} \quad \text{and} \quad dx = \frac{\beta}{1 - \beta t} dy,$$

we obtain

$$\begin{aligned} \Phi_Y(t) &= \int_0^\infty \frac{\beta/(1-\beta t)}{\Gamma(\alpha)\beta^\alpha} \left(\frac{\beta y}{1 - \beta t} \right)^{\alpha-1} e^{-y} dy \\ &= \frac{1}{(1-\beta t)^\alpha} \cdot \frac{1}{\Gamma(\alpha)} \int_0^\infty y^{\alpha-1} e^{-y} dy = \frac{1}{(1-\beta t)^\alpha} \quad \text{for } t < 1/\beta. \end{aligned}$$

Therefore, for the gamma distribution, we obtain

$$E[X] = \Phi'_X(0) = \alpha\beta \quad \text{and} \quad \text{Var}[X] = \Phi''_X(0) - (E[X])^2 = \alpha(\alpha+1)\beta^2 - \alpha^2\beta^2 = \alpha\beta^2.$$

We turn now to consider the special case of the gamma distribution when $\alpha = r/2$, for some positive integer r, and $\beta = 2$. This gives the distribution of an absolutely continuous random variable X with density

$$f(x) = \begin{cases} \frac{1}{\Gamma(r/2)2^{r/2}} x^{r/2-1} e^{-x/2} & \text{if } x > 0; \\ 0 & \text{otherwise.} \end{cases}$$

This distribution is called the **chi-square distribution** and we write $X \sim \chi^2(r)$ where, for no obvious reason, r is called the number of degrees of freedom of the distribution. The moment generating function of the chi-square distribution is

$$\Phi_X(t) = \frac{1}{(1-2t)^{r/2}} \quad \text{for } t < 1/2,$$

and its expected value and variance are, respectively, $E[X] = r$ and $\text{Var}[X] = 2r$.

Theorem 2.6. *Let $X_i \sim \Gamma(\alpha_i, \beta)$, $i = 1, \ldots, k$, be independent random variables. Then, $Y_k = \sum_{i=1}^k X_i \sim \Gamma(\sum_{i=1}^k \alpha_i, \beta)$.*

Theorem 2.7. *Let $X \sim U[0,1]$. Then, $Y = -2 \ln X \sim \chi^2(2)$.*

Theorem 2.8. *Let $X \sim \Gamma(\alpha_x, \beta)$ and $Y \sim \Gamma(\alpha_y, \beta)$ be two independent random variables. Then, $X+Y$ and X/Y are independent random variables and $X+Y$ and $X/(X+Y)$ are also independent random variables.*

Theorem 2.9. *Let $\{X_n\}_{n=1}^\infty$ be a sequence of independent random variables such that $X_n \sim exp(\beta)$ for each $n = 1, 2, \ldots$. Let $Y_n = \sum_{i=1}^n X_i$ for $n = 1, 2, \ldots$ and let Z be the random variable corresponding to the number of $Y_n \in [0, t]$ for $t > 0$. Then $Z \sim P(t/\beta)$.*

We close this subsection by introducing another important distribution related with the gamma distribution. Let U, V be two independent random variables such that $U \sim \Gamma(\alpha, 1)$ and $V \sim \Gamma(\beta, 1)$ The joint density function of (U, V) is then

$$h(u, v) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)} u^{\alpha-1} v^{\beta-1} e^{-u-v}, \quad \text{for } \ 0 < u, v < \infty.$$

Consider the change of variables given by $X = U/(U+V)$ and $Y = U+V$. Using the "change of variables formula," one obtains

$$f(x, y) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} y^{\alpha+\beta-1} e^{-y}, \quad \text{for } \ 0 < x < 1 \ \text{ and } \ 0 < y < \infty.$$

The marginal distribution of X is then

$$\begin{aligned} f_1(x) &= \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)} \int_0^\infty y^{\alpha+\beta-1} e^{-y} dy \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \quad \text{for } \ 0 < x < 1. \end{aligned}$$

The density function above is that of the **beta distribution** with parameters α and β, and we write $X \sim B(\alpha, \beta)$. Now, it follows from Theorem 2.8 above that X and Y are independent random variables. Therefore, since $f(x, y) = f_1(x) f_2(y)$, it must be the case that

$$f_2(y) = \frac{1}{\Gamma(\alpha+\beta)} y^{\alpha+\beta-1} e^{-y}, \quad \text{for } \ 0 < y < \infty.$$

The function $f_2(u)$ above corresponds to the density function of a gamma distribution such that $Y \sim \Gamma(\alpha + \beta, 1)$.

It can be checked that the expected value and the variance of X, which has a beta distribution, are given by

$$E[X] = \frac{\alpha}{\alpha + \beta} \quad \text{and} \quad \text{Var}[X] = \frac{\alpha\beta}{(\alpha + \beta + 1)(\alpha + \beta)^2}.$$

There is no closed expression for the moment generating function of a beta distribution.

The intuition given above regarding the relation between the gamma and the beta distributions can be extended by the following result.

Theorem 2.10. *Let $U \sim \Gamma(\alpha, \gamma)$ and $V \sim \Gamma(\beta, \gamma)$ be two independent random variables. Then $X = U/(U + V) \sim B(\alpha, \beta)$.*

6.2.3. *The normal distribution*

We introduce now one of the most important distributions in the study of probability and mathematical statistics, the normal distribution. The Central Limit Theorem shows that normal distributions provide a key family of distributions for applications and for statistical inference.

Definition 2.1. *A random variable X is said to have the* **normal distribution** *if its density function is given by*

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

The parameters μ and σ^2 correspond, respectively, to the mean and variance of the distribution. We write $X \sim N(\mu, \sigma^2)$. The **standard normal distribution** *is the normal distribution obtained when $\mu = 0$ and $\sigma^2 = 1$.*

Suppose that $X \sim N(0, 1)$ and consider the transformation $Y = a + bX$ for $b > 0$. Using the "change of variable formula," we can derive the expression for the density function of Y as

$$h(y) = f\left(\frac{y - a}{b}\right)\frac{1}{b} = \frac{1}{b} \cdot \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{y - a}{b}\right)^2\right\},$$

so that $Y \sim N(a, b^2)$. For $a = \mu$ and $b^2 = \sigma^2$ one can obtain the converse implication by applying the "change of variable formula" too. Therefore, the following claim holds.

Theorem 2.11. *A random variable X has a $N(\mu, \sigma^2)$ distribution if and only if the random variable $(X - \mu)/\sigma$ has a $N(0, 1)$ distribution.*

Using the result above, we can obtain the moment generating function of a random variable $X \sim N(\mu, \sigma^2)$ by using the fact that $X = \sigma Z + \mu$ for some random variable $Z \sim N(0,1)$. This is done as follows. First, note that

$$\begin{aligned}\Phi_X(t) &= E[e^{tX}] = E[e^{t\sigma Z + t\mu}] = e^{t\mu} E[e^{t\sigma Z}] \\ &= e^{\mu t} \int_{-\infty}^{+\infty} e^{t\sigma z} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz.\end{aligned}$$

Second, we compute the integral above as

$$\begin{aligned}\int_{-\infty}^{+\infty} e^{t\sigma z} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz &= e^{\sigma^2 t^2/2} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-(z-\sigma t)^2/2} dz \\ &= e^{\sigma^2 t^2/2} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-s^2/2} ds \\ &= e^{\sigma^2 t^2/2},\end{aligned}$$

using the change of variable $s = z - \sigma t$ and the fact that $\int_{-\infty}^{+\infty} 1/\sqrt{2\pi} e^{-s^2/2} ds = 1$. Therefore, we finally obtain

$$\Phi_X(t) = e^{\mu t} e^{\sigma^2 t^2/2} = e^{\mu t + \sigma^2 t^2/2}.$$

Even though many applications can be analyzed using normal distributions, normal density functions usually contain a factor of the type $\exp\left\{-s^2\right\}$. Therefore, since antiderivatives cannot be obtained in closed form, numerical integration techniques must be used. Given the relation between a normal distribution and the standard normal distribution, we make use of numerical integration computations as follows. Consider a random variable $X \sim N(\mu, \sigma^2)$, denote by F its distribution function and by $H(z) = \int_{-\infty}^{z} 1/\sqrt{2\pi} e^{-s^2/2} ds$ the distribution function of the random variable $Z = (X - \mu)/\sigma \sim N(0,1)$. Now, suppose that we wish to compute $F(x) = P[X \leq x]$. Then, we use the fact that

$$P[X \leq x] = P\left[Z \leq \frac{x - \mu}{\sigma}\right] = H\left(\frac{x - \mu}{\sigma}\right).$$

Therefore, all that we need are numerical computations for $H(z)$.

We close this section with a few important results concerning normal distributions.

Theorem 2.12. *Let X be a standard normal random variable. Then,*

$$P[X > x] \approx \frac{1}{\sqrt{2\pi} x} e^{-x^2/2} \quad as \quad x \to \infty.$$

Theorem 2.13. *If X and Y are independent normally distributed random variables, then $X + Y$ and $X - Y$ are independent.*

Theorem 2.14. *Let $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$, be independent random variables. Then, for $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$, we have*

$$\sum_{i=1}^{n} \alpha_i X_i \sim N\left(\sum_{i=1}^{n} \alpha_i \mu_i, \sum_{i=1}^{n} \alpha_i^2 \sigma_i^2\right).$$

Theorem 2.15. *If $X \sim N(\mu, \sigma^2)$, then $(X - \mu)^2/\sigma^2 \sim \chi^2(1)$.*

The result above has already been demonstrated in Examples 23 and 29.

6.2.4. *The multivariate normal distribution*

Here we consider the generalization of the normal distribution to random vectors.

Definition 2.2. *A random vector $X = (X_1, \ldots, X_n)$ is said to have the **n-variate normal distribution** if its density function is given by*

$$f(x) = f(x_1, \ldots, x_n) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu)'\Sigma^{-1}(x - \mu)\right\},$$

where $\Sigma \in \mathbb{R}^n \times \mathbb{R}^n$ is a symmetric, positive semi-definite matrix and $\mu = (\mu_1, \ldots, \mu_n) \in \mathbb{R}^n$. We write $X = (X_1, \ldots, X_n) \sim N(\mu, \Sigma)$. The vector is μ is called the **mean vector** *and the matrix Σ is called the* **dispersion matrix** *or* **variance-covariance matrix** *of the multivariate distribution.*[1]

The special case $n = 2$ yields the **bivariate normal distribution**. Consider a random vector $(X, Y) \sim N(\mu, \Sigma)$, where

$$\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}.$$

Here σ_{xy} denotes the covariance between X and Y. Thus, if ρ is the correlation coefficient between X and Y, then we have $\sigma_{xy} = \rho\sigma_x\sigma_y$, where the symbol σ_k stands for the **standard deviation**, $\sigma_k = +(\sigma_k^2)^{1/2}$, of the corresponding random variable $k = x, y$. After noting these notational rearrangements, matrix Σ above can be easily inverted to obtain

$$\Sigma^{-1} = \frac{1}{\sigma_x^2\sigma_y^2(1 - \rho^2)} \begin{pmatrix} \sigma_y^2 & -\rho\sigma_x\sigma_y \\ -\rho\sigma_x\sigma_y & \sigma_x^2 \end{pmatrix}.$$

Therefore, the joint density function of (X, Y) is

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1 - \rho^2}} \exp\{Q\},$$

where

$$Q = -\frac{1}{2(1 - \rho^2)}\left[\left(\frac{x - \mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x - \mu_x}{\sigma_x}\right)\left(\frac{y - \mu_y}{\sigma_y}\right) + \left(\frac{y - \mu_y}{\sigma_y}\right)^2\right].$$

[1] Some authors refer to Σ^{-1}, instead of Σ, as the variance-covariance matrix.

The following result is crucial to analyze the relation between a multivariate normal distribution and its marginal distributions.

Theorem 2.16. *Let $X \sim N(\mu, \Sigma)$ such that X, μ, and Σ can be partitioned as*

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Then, $X_s \sim N(\mu_s, \Sigma_{ss})$, $s = 1, 2$. Moreover, X_1 and X_2 are independent random vectors if and only if $\Sigma_{12} = \Sigma_{21} = \underline{0}$.

The result in the theorem above tells us that any marginal distribution of a multivariate normal distribution is also normal and, further, its mean and variance-covariance matrix are those associated with that partial vector. It also asserts that, for the normal case, independence of the random variables follows from their no correlation.

Let us consider the bivariate case to fix ideas. It follows from the theorem above that if $(X, Y) \sim N(\mu, \Sigma)$, with

$$\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix},$$

then $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$. Suppose now that X and Y are uncorrelated. Then, $\rho = 0$ and we can use the expression above for $f(x, y)$ to conclude that $f(x, y) = f_x(x) f_y(y)$, where

$$f_k(k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left\{ \left(\frac{k - \mu_k}{\sigma_k} \right)^2 \right\} \quad \text{for} \quad k = x, y.$$

Hence, if (X, Y) is bivariate normally distributed with the parameters given above, and X and Y are uncorrelated, then $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$. This follows simply from the fact that (X, Y) is bivariate normally distributed as stated in the theorem above. Furthermore, X and Y are independent!

However, it is possible for two random variables X and Y to be distributed jointly in a way such that each one alone is marginally normally distributed, and they are uncorrelated, but they are not independent. This can happen only if these two random variables are not distributed jointly as bivariate normal.

Example 2.4. *Suppose that X has a normal distribution with mean 0 and variance 1. Let W be a random variable which takes the values either 1 or -1, each with probability $1/2$, and assume W is independent of X. Now, let $Y = WX$. Then, it can be checked that*

(i) X and Y are uncorrelated,

(ii) X and Y have the same normal distribution, and

(iii) X and Y are not independent.
To see that X and Y are uncorrelated, notice that

$$\begin{aligned}\text{Cov}[X,Y] &= E[XY] - E[X]E[Y] = E[XY] \\ &= E[XY|W=1]P[W=1] + E[XY|W=-1]P[W=-1] \\ &= E[X^2](1/2) + E[-X^2](1/2) = 1(1/2) - 1(1/2) = 0.\end{aligned}$$

To see that X and Y have the same normal distribution notice that

$$\begin{aligned}F_Y(x) &= P[Y \le x] = P[Y \le x|W=1]P[W=1] + P[Y \le x|W=-1]P[W=-1] \\ &= P[X \le x](1/2) + P[-X \le x](1/2) \\ &= P[X \le x](1/2) + P[X \ge -x](1/2) = P[X \le x] = F_X(x).\end{aligned}$$

Finally, to see that X and Y are not independent, simply note that $|Y| = |X|$.

We have already seen how to obtain the marginal distributions from a multivariate normal distribution. We have learned that the marginal distributions are also normal. We now ask whether putting together two normal distributions yields a bivariate normal distribution. The answer to this question depends crucially on whether the two random variables are independent or not.

Theorem 2.17. *Let $X \sim N(\mu_x, \sigma_x^2)$ and $Y \sim N(\mu_y, \sigma_y^2)$ be two independent random variables. Then $(X, Y) \sim N(\mu, \sigma)$, where*

$$\mu = \begin{pmatrix}\mu_x \\ \mu_y\end{pmatrix} \quad \textit{and} \quad \Sigma = \begin{pmatrix}\sigma_x^2 & 0 \\ 0 & \sigma_y^2\end{pmatrix}.$$

However, in general the fact that two random variables X and Y both have a normal distribution does not imply that the pair (X, Y) has a joint normal distribution. A simple example is one in which X has a normal distribution with expected value 0 and variance 1, and $Y = X$ if $|X| > c$ and $Y = -X$ if $|X| < c$, where c is approximately equal to 1.54. In this example the two random variables X and Y are uncorrelated but not independent.

The following result tells us about the distribution of a linear transformation of a normal random vector.

Theorem 2.18. *Let $X \sim N(\mu, \Sigma)$, and let $A \in \mathbb{R}^m \times \mathbb{R}^n$ and $b \in \mathbb{R}^m$. Then,*

$$Y = [A \cdot X + b] \sim N(A \cdot \mu + b, A \cdot \Sigma \cdot A').$$

The following result clarifies the relation between a multivariate normal distribution and its conditional distributions.

Theorem 2.19. *Let $X \sim N(\mu, \Sigma)$ such that X, μ, and Σ can be partitioned as*

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Assume that Σ is positive definite. Then, the conditional distribution of $X_1|X_2 = x_2$ is

$$N\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right).$$

For the bivariate normal case, one can use the expression given above for the joint density of (X, Y) to obtain, after dividing such expression by the marginal density of X,

$$[Y|X = x] \sim N\left(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x), \sigma_y^2(1 - \rho^2)\right),$$

as stated in the result in the theorem above. We conclude by emphasizing that the conditional expected value of Y given $X = x$ is linear in x:

$$E[Y|X = x] = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x).$$

6.2.5. *The t and the F distributions*

Definition 2.3. *A random variable X is said to have the* ***t*** **distribution** *if its density function is given by*

$$f(x) = \frac{\Gamma((\alpha+1)/2)}{(\alpha\pi)^{1/2}\Gamma(\alpha/2)}\left(1 + \frac{x^2}{\alpha}\right)^{-(\alpha+1)/2} \quad \text{for each } x \in \mathbb{R}.$$

We write $X \sim t(\alpha)$ and α is called the **degree of freedom** *of the distribution.*

The t distribution is important in statistics because of the following results.

Theorem 2.20. *Let $X \sim N(0, 1)$ and $Y \sim \chi^2(n)$ be independent random variables. Then*

$$T = \frac{X}{\sqrt{Y/n}} \sim t(n).$$

Theorem 2.21. *Let $X_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$, be independent random variables and let $\overline{X}_n$ and S_n^2 be the random variables defined as*

$$\overline{X}_n = \sum_{i=1}^{n} X_i/n \quad \text{and} \quad S_n^2 = \sum_{i=1}^{n}(X_i - \overline{X}_n)^2/(n-1).$$

Then:
(i) $\overline{X}_n \sim N(\mu, \sigma^2/n)$;
(ii) $\overline{X}_n$ *and* S_n^2 *are independent;*
(iii) $(n-1)S_n^2/\sigma^2 \sim \chi^2(n-1)$;
(iv) $(\overline{X}_n - \mu)/(S_n/\sqrt{n}) \sim t(n-1)$.

Definition 2.4. *A random variable X is said to have the* $\boldsymbol{F}$ **distribution** *if its density function is given by*

$$f(x) = \frac{\Gamma((\alpha+\beta)/2)\alpha^{\alpha/2}\beta^{\beta/2}}{\Gamma(\alpha/2)\Gamma(\beta/2)} \cdot \frac{x^{(\alpha/2)-1}}{(\beta+\alpha x)^{(\alpha+\beta)/2}} \quad \text{for } x > 0,$$

and $f(x) = 0$ for $x \leq 0$. We write $X \sim F(\alpha, \beta)$, and α and β are called the **degrees of freedom** *of the distribution.*

The F distribution is important in statistical work because of the following result.

Theorem 2.22. *Let $X \sim \chi^2(\alpha)$ and $Y \sim \chi^2(\beta)$ be independent random variables. Then*

$$Z = \frac{X/\alpha}{Y/\beta} \sim F(\alpha, \beta).$$

2.3 Exercises

2.1. Let X be a random variable with moment generating function

$$\Phi_X(t) = \left(\frac{3}{4} + \frac{1}{4}e^t\right)^6.$$

Obtain the density function of X.

2.2. Let X be the random variable associated to the number of successes throughout n independent repetitions of a random experiment with probability p of success. Show that X satisfies the following form of the *Weak Law of Large Numbers*:

$$\lim_{n\to\infty} P\left[\left|\frac{X}{n} - p\right| < \varepsilon\right] = 1 \quad \text{for each given} \quad \varepsilon > 0.$$

2.3. Let X be a random variable with geometric distribution. Show that

$$P[X > k + j | X > k] = P[X > j].$$

2.4. Let X be a random variable with moment generating function

$$\Phi_X(t) = e^{5(e^t-1)}.$$

Compute $P[X \leq 4]$.

2.5. Let $X \sim P(1)$. Compute, if it exists, the expected value $E[X!]$.

2.6. Prove Theorem 2.6.

2.7. Let X_1, X_2, and X_3 be independent and identically distributed random variables, each with density function $f(x) = e^{-x}$ for $x > 0$. Find the density function of $Y = \min\{X_1, X_2, X_3\}$.

2.8. Let $X \sim U[0,1]$. Find the density function of $Y = -\ln X$.

2.9. Prove Theorem 2.13.

2.10. Let (X_1, X_2, X_3) have a multivariate normal distribution with mean vector $\underline{0}$ and variance-covariance matrix

$$\Sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}.$$

Find $P[X_1 > X_2 + X_3 + 2]$.

2.11. Let $X \sim N(0,1)$ and let n be a positive natural number. Using the result

$$\int_0^{+\infty} s^{2n+1} e^{-s^2/2} ds = 2^n n!,$$

show that

$$E\left[\, |X|^{2n+1} \,\right] = 2^n n! \sqrt{\frac{2}{\pi}}.$$

2.12. Let $X, Y \sim N(0, \sigma^2)$ be two independent random variables and let $U = +\sqrt{X^2 + y^2}$ and $W = X/Y$. Compute the marginal densities of U and W. Are they independent random variables?

2.13. Let $X_i \sim N(0,1)$, $i = 1, \ldots, 4$, be independent random variables. Show that $Y = X_1 X_2 + X_3 X_4$ has the density function $f(y) = (1/2)\exp\{-|y|\}$ for each $y \in \mathbb{R}$.

2.14. Let X and Y be two random variables distributed standard normally. Denote by f and F the density function and the distribution function of X, respectively. Likewise, denote by g and G the density function and the distribution function of Y. Let (X, Y) have joint density function

$$h(x,y) = f(x)g(y)[1 + \alpha(2F(x) - 1)(2G(y) - 1)],$$

where α is a constant such that $|\alpha| \leq 1$. Show that $X + Y$ is not normally distributed except in the trivial case $\alpha = 0$, i.e., when X and Y are independent.

2.15. Give a closed expression for $E[X^r]$, $r = 1, 2, \ldots$, where $X \sim F(\alpha, \beta)$.

2.16. Let $X \sim \chi^2(n)$ and $Y \sim \chi^2(m)$ be independent random variables. Find the density of $Z = X/(X + Y)$.

2.17. Let $(X, Y) \sim N(\mu, \Sigma)$. Determine the distribution of the random vector $(X + Y, X - Y)$. Show that $X + Y$ and $X - Y$ are independent if $\text{Var}[X] = \text{Var}[Y]$.

2.18. Let $X \sim N(2, 4)$. Compute $P[1 < X < 6]$ using only the function $\gamma(y) = 1/\sqrt{2\pi} \int_0^y e^{-s^2/2} ds$.

2.19. Let (X, Y) have joint density function:

$$f(x, y) = \frac{1}{6\pi\sqrt{7}} \exp\left\{-\frac{8}{7}\left(\frac{x^2}{16} - \frac{31x}{32} + \frac{xy}{8} + \frac{y^2}{9} - \frac{4y}{3} + \frac{71}{16}\right)\right\} \quad \text{for } x, y \in \mathbb{R}.$$

(a) Find the means and variances of X and Y. Find $\text{Cov}[X, Y]$ too.
(b) Find the conditional density of $Y|X = x$, $E[Y|X = x]$, and $\text{Var}[Y|X = x]$.
(c) Find $P[4 \leq Y \leq 6|X = 4]$.

2.20. Let $X \sim t(\alpha)$. Show that $X^2 \sim F(1, \alpha)$. Let $f_\alpha(x)$ denote the density function of X. Show that

$$\lim_{\alpha \to \infty} f_\alpha(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

for each $x \in \mathbb{R}$.

Bibliography

P. Billingsley. *Probability and Measure*. John Wiley and Sons, 3rd edition, 1995.

C. Carathéodory. *Vorlesungen über reelle Funktionen*. Leipzig: Teubner, 1918.

A. N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin, 1933.

P. S. Laplace. *Essai Philosophique sur les Probabilités*. Courcier Imprimeur, Paris, 1814.

H. Lebesgue. Sur l'intégration des fonctions discontinues. *Annales scientifiques de l'École Normale Supérieure*, 27(3):361'45', 1910.

Index