Lecture Notes on Probability and Statistics

A Practical Guide

Antonio Jimenez

August 2023

The goal of this notes is to serve as a roadmap to navigate introductory courses in mathematical probability and statistics. The notes follow a practical approach. While trying to present abstract concepts rigorously, the main goal is to help articulate teaching and learning around applications. With this approach in mind, the exposition keeps the theoretical side as simple as possible, while trying to enrich discussions with a variety of motivating examples. The idea is that instructors can largely focus on working out examples and discussing their takeaways with students. The intent is to help develop sound understanding of difficult concepts through plenty of opportunities to practice. Definitions and demonstrations are presented in ways such that, most of the times, students wish to read them just for consultation purposes. In consonance, these notes should not be regarded as particularly suitable for courses more oriented toward thorough philosophical discussions on mathematical probability abstract concepts, or its foundations.

The only mathematical prerequisites are knowledge of typical first-year college calculus, including sums of infinite series, differentiation, and single and multiple integration.

Throughout its chapters, the notes introduce first the main notions briefly and then jump right away into a series of solved examples that deal closely with such notions. At the end of each chapter, a set of unsolved problems aims at complementing the presented topics. I intentionally use a concise cookbook style so that the notes can be dynamically navigated. Several parts of the notes—such as sections, solved examples, or results—are labelled using (steaming) coffee cups. Following requirements of coffee intake, the level of difficulty, or abstractness, goes from the one-coffee-cup (the easiest, or more practical) to the three-coffee-cup (the hardest, or more abstract) category. Depending on the course level, instructors might want to focus on some parts while skipping others. I personally would view one-coffee-cup content as suitable for undergraduate courses, whereas three-pepper material would perhaps suit better graduate courses.

These notes have benefited enormously from my own graduate teaching at CIDE during many years. I wish to thank many students who have helped with their patience and comments to improve these notes.

The *LaTex* class used to typeset these notes is based on the *The Legrand Orange Book LaTex* class.

Contents

I Probability

1 Probability Models 9

1.1 Introduction: Random Phenomena (☕) 9

1.2 Sample Spaces, Events, and Probability (☕) 11

1.2.1 The σ-algebra generated by a family of sets (☕☕☕) 15

1.2.2 Borel σ-algebras on Euclidean spaces (☕☕☕) 17

1.3 Measures and Probability Laws (☕☕) 19

1.4 Conditional Probability and Independence (☕) 22

1.5 Practice Exercises 31

2 Random Variables and Distributions (☕) 35

2.1 Random Variables and Random Vectors (☕) 35

2.2 Distribution Functions (☕) 37

2.3 Discrete Random Variables (☕) 41

2.4 Continuous Random Variables (☕) 44

2.5 Functions of a Random Variables (☕☕) 51

2.6 Independence of Random Variables (☕) 55

2.7 Practice Exercises 59

3 Expected Values and Moments of Distributions (☕) 65
3.1 Covariance and Correlation (☕) 72
3.2 Practice Exercises 75
4 Moment Generating Functions (☕☕) 79
4.1 Practice Exercises 84
5 Some Important Distributions (☕) 85
5.1 Some Discrete Distributions (☕) 85
5.1.1 The binomial distribution 85
5.1.2 The negative binomial distribution 87
5.1.3 The multinomial distribution 89
5.1.4 The Poisson distribution 90
5.2 Some Continuous Distributions (☕) 91
5.2.1 The uniform distribution 91
5.2.2 The Γ, χ^2, and Beta distributions 92
5.2.3 The normal distribution 96
5.2.4 The t and the F distributions 102
5.3 Practice Exercises 103
II Statistical Inference
6 Convergence of Distributions (☕) 109
6.1 Convergence in Distribution (☕) 109
6.2 Convergence in Probability and Almost Surely (☕) 111
6.3 Laws of Large Numbers (☕) 113
6.4 The Central Limit Theorem (☕) 115
6.5 Practice Exercises 116
7 Parametric Point Estimation (☕) 119
7.1 Consistent Estimation (☕) 120
7.2 Sufficient Estimation (☕) 120

7.3 Unbiased Estimation (☕) 122

7.4 Maximum Likelihood Estimation (☕) 123

7.5 Rao-Cramér Bound and Efficient Estimation (☕) 127

7.6 Practice Exercises 128

8 Hypothesis Testing 129

8.1 Neyman-Pearson Theory (☕) 129

8.2 Tests Based on the Likelihood Ratio (☕) 136

8.3 Practice Exercises 139

III Miscellanea on Probability

9 Combinatorics (☕☕) 143

9.1 Ordered Samples and Permutations 143

9.2 Combinations 145

9.3 Practice Exercises 150

10 Integration (☕☕) 153

11 Extension of Probability Measures (Continued) (☕☕☕) . . . 155

Probability

1 **Probability Models** **9**

1.1 Introduction: Random Phenomena (☕) . . . 9

1.2 Sample Spaces, Events, and Probability (☕) 11

1.3 Measures and Probability Laws (☕☕) 19

1.4 Conditional Probability and Independence (☕) 22

1.5 Practice Exercises 31

2 **Random Variables and Distributions (☕)** **35**

2.1 Random Variables and Random Vectors (☕) 35

2.2 Distribution Functions (☕) 37

2.3 Discrete Random Variables (☕) 41

2.4 Continuous Random Variables (☕) 44

2.5 Functions of a Random Variables (☕☕) . . 51

2.6 Independence of Random Variables (☕) . 55

2.7 Practice Exercises 59

3 **Expected Values and Moments of Distributions (☕)** **65**

3.1 Covariance and Correlation (☕) 72

3.2 Practice Exercises 75

4 **Moment Generating Functions (☕☕)** **79**

4.1 Practice Exercises 84

5 **Some Important Distributions (☕) 85**

5.1 Some Discrete Distributions (☕) 85

5.2 Some Continuous Distributions (☕) 91

5.3 Practice Exercises 103

1. Probability Models

1.1 Introduction: Random Phenomena (☕)

In a variety of populations, scientists try to formulate general laws about outcomes of either natural phenomena or designed experiments. A typical law has the form:

> "if a certain set B of circumstances happens, then a set of outcomes A occurs."

Some of such laws appear in chemistry, physics, or the social sciences; for instance, the law of conservation of mass, the law of gravity, or the low of demand. Suppose that we are aware that the set of circumstances contained in B have happened. Then, if the outcomes in A occur inevitably, we say that A is *certain*, or *sure*. If the outcomes in A can never occur, then we say that A is *impossible*.

Intermediate cases appear if the outcomes in A may, or may not, occur whenever the circumstances in B happen. Unlike certain and impossible sets of outcomes, the presence of randomness implies that the set of circumstances in B do not account for all the necessary and sufficient conditions for the outcomes in A to happen. In such cases, we are uncertain about the occurrence of A and thus would like to know how likely is that the outcomes in A occur. The main concern of probability theory is to assess scientifically the likelihood of occurrence of such outcomes in A. When we face this sort of intermediate cases, we say that we are in the presence of a *random phenomenon*, or *random experiment*, and A is said to be a *random event*.

In addition to assessing probabilities, by providing systematic means to model random experiments, probability theory lies the statistical foundations to draw inferences from populations.

Probability theory thus provides a set of tools that help us achieve a valuable two-way scientific goal when dealing with random phenomena or experiments. In short, it allows scientists both to (i) construct rigorously theoretical (deductive) models to investigate problems that deal with uncertainty and (ii) propose solid empirical (inductive) exercises to raise research questions and test the implications of theoretical models.

At first glance, it seems daunting to make general statements about random phenomena and the likelihood of occurrence of uncertain outcomes. However, repetition of random phenomena shows that many of such phenomena exhibit some statistical regularities that allows us to study them using a systematic approach. For random phenomena that exhibit certain regularities over repetition, it is possible to estimate the odds of occurrence of event A by using certain laws, which are referred to as *probabilistic* or *stochastic*. Stochastic laws have the general structure:

> "if the set B of circumstances happens repeatedly n times,
> then the set of outcomes 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 thought of as the fraction m/n.

Now, how do we assign probabilities of occurrence to random events? Historically, there have been two major 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 circumstances in 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 fraction m/n. However, to that logic of the fraction m/n, it adds 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

[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."

and rigorous treatment of a very general set of random phenomena. The axiomatic approach provides the foundations of modern probability theory and it is at the heart of ideas presented and discussed in these notes.

1.2 Sample Spaces, Events, and Probability (☕)

To deal with possible outcomes of random phenomena or experiments, it is useful first to consider some basic relations and operations for sets of elements.

Observation 1.1 — Set Relations. Take two sets A, B. Then:

1. We say that A is a subset of B, $A \subseteq B$, whenever (i) $\omega \in A$ implies $\omega \in B$;
2. We say that sets A and B are equivalent, $A = B$, whenever (i) $\omega \in A$ implies $\omega \in B$ (that is, $A \subseteq B$) and (ii) $\omega \in B$ implies $\omega \in A$ (that is, $B \subseteq A$);
3. We say that A is a strict subset of B, $A \subset B$, whenever (i) $\omega \in A$ implies $\omega \in B$ and (ii) there is some $\omega \in B$ with $\omega \notin A$, that is, $A \subseteq B$ and $A \neq B$;

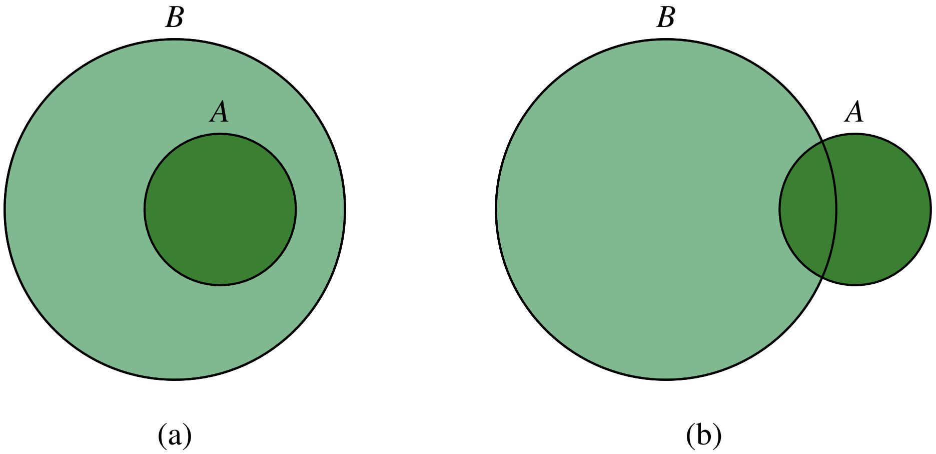

Figure 1.1: Set inclusion. (a) $A \subset B$ with $A \neq B$, (b) $A \nsubseteq B$ and $B \nsubseteq A$.

Observation 1.2 — Set Operations. Given a set Ω and two subsets $A, B \subseteq \Omega$, then:

1. The cross product of A and B, $A \times B$, is the set of all pairs such that the first entry is an element of A and the second entry is an element of B, $A \times B = \{(a,b) : a \in A, b \in B\}$;
2. The union of A and B, $A \cup B$, is the set of all elements that belong to either A, B, or both: $A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B\}$;
3. The intersection of A and B, $A \cap B$, is the set of all elements that belong to both A and B:

$A \cap B = \{\omega \in \Omega \,:\, \omega \in A \text{ and } \omega \in B\}$. Moreover, the sets A and B are disjoint if the do not have any element in common, that is, if $A \cap B = \emptyset$;

4. The subtraction of set B from set A, $A \setminus B$, is the set of all elements that belong to A and do not belong to B: $A \setminus B = \{\omega \in \Omega \,:\, \omega \in A \text{ or } \omega \notin B\}$;
5. The complement of set A—with respect to a reference set Ω—, A^c, is the set of all elements that do not belong to A—and still belong to Ω: $A^c = \{\omega \in \Omega \,:\, \omega \notin A\} = \Omega \setminus A$.

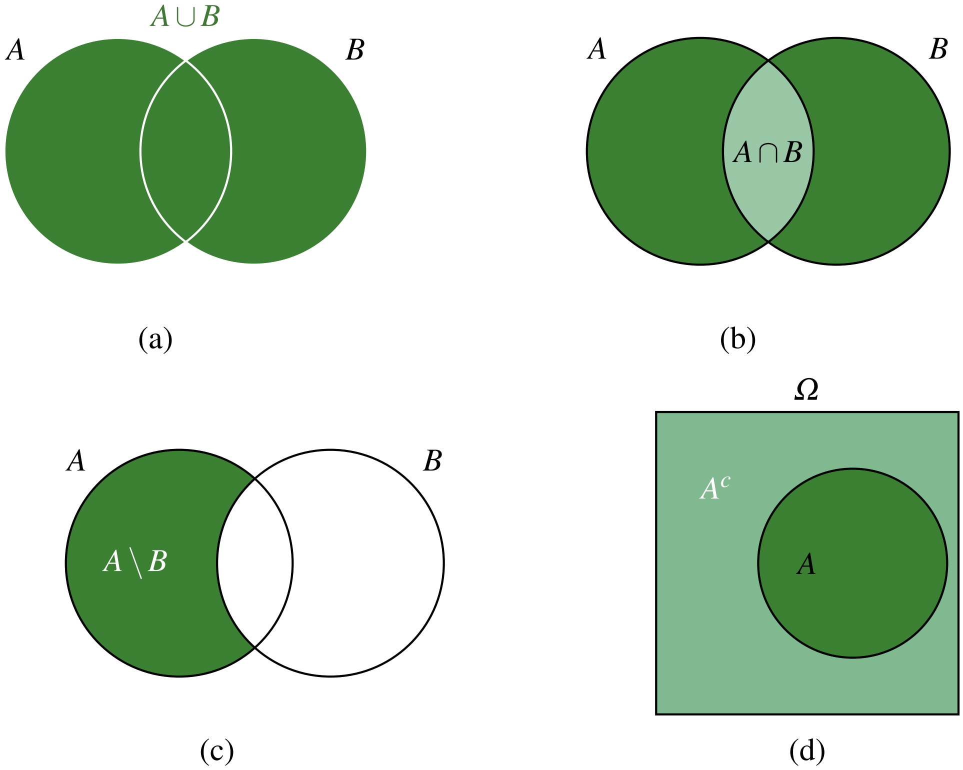

Figure 1.2: Basic set operations. (a) $A \cup B$, (b) $A \cap B$, (c) $A \setminus B$, (d) $A^c = \Omega \setminus A$.

When dealing with a random phenomenon or experiment, we would like to consider a space of random events that allows us to consider as many, and as varied, as possible sets of outcomes of uncertainty. In particular, we begin by proposing an arbitrary nonempty set Ω of *elementary events* ω. There is no fix rule to propose the set Ω and the possible outcomes of a given random experiment can be translated into elementary events in different ways. To obtain a clean model, it is advisable that any possible outcome of the experiment be included in our sample space. Also, it is convenient that each elementary event has enough detail so as to distinguish between all outcomes of interest. Therefore, it turns quite useful to take the approach of considering each elementary event $\omega \in \Omega$ as a complete and exhaustive description of a possible outcome of uncertainty.[2] The set Ω of all possible elementary events ω of a random phenomenon is known as *sample space*.

■ **Example 1.1** [☕] Consider an experiment in which a coin is flipped twice. In principle, there are several ways of proposing a sample set Ω that captures the outcomes of this experiment. First, we could consider the number of times that Heads shows up. In this case, $\Omega = \{0,1,2\}$. Secondly, we could consider the particular outcomes, independent of the order in which the outcome appears. Then, we would have $\Omega = \{\{H,H\},\{H,T\},\{T,T\}\}$. However a most complete and exhaustive description of a possible outcome would consider particular outcomes, as well as the order in which they appear. This consideration would give us the most convenient sample set to model our experiment. We would have $\Omega = \{(H,H),(HT),(T,H),(T,T)\}$. ■

Sample sets Ω can be countable, whenever we can think of a one-to-one map from the set Ω to a set of integers. Of course if Ω is finite, then it is also countable. Sample sets Ω can also be uncountable, whenever such a one-to-one map from Ω to a set of integers does not exist and we must then resort to associate Ω with a continuum real set, such as the interval $[0,1]$. At a technical level, the mathematical treatments required for countable and uncountable sample sets are different. However, the underlying intuitions are similar.

■ **Example 1.2** [☕] Suppose that we consider the experiment of flipping a coin three times. Then, a natural way of proposing the sample space would be

$$\Omega = \{(H,H,H),(T,H,H),(H,T,H),(H,H,T),(T,T,H),(T,H,T),(H,T,T),(T,T,T)\},$$

a countable finite set. Now, suppose that we consider the experiment of the quality the items produced in a certain manufacturing process with no determined lifetime. Then, we can consider that an elementary event is an infinite sequence $\omega = (\omega_1, \omega_2, \dots)$ where each $\omega_i \in \{G,D\}$, for $i = 1,2,\dots$, with $G =$"good quality item" and $D =$"defective item." In this case, $\Omega = \{G,D\} \times \{G,D\} \times \cdots = \{G,D\}^\infty$ is countable infinite. Finally, suppose that we consider the random phenomenon of the lifetime of a washing machine. Then, it would be natural to begin with $\Omega = [0,+\infty)$. In this case, Ω is uncountable. ■

A random event is formally identified as a subset A of the set Ω, $A \subseteq \Omega$. After proposing a set Ω of possibles outcomes of the random phenomenon, we would like then to add a family

[2] In the social sciences, elementary events that fit into such a description are also known as *states of the world* or states of nature.

$\mathscr{F}$ of subsets of Ω, or events, that satisfies certain desirable properties. To do so, we must keep in mind that while we will be interested in assessing probabilities of occurrence for random events in $\mathscr{F}$, subsets which do not belong to $\mathscr{F}$ will be out of what the proposed model can say. In intuitive terms, random events can be described simply by using everyday sentences. Then, given events/sentences like "*A*" and "*B*," it makes sense to connect such sentences in order to form new sentences like "*A* and *B*," "*A* or *B*," "*A* but not *B*, or not "just not *B*." In consequence, it is desirable that the family of events $\mathscr{F}$ be closed under the set operations of intersection, union, and complement. Of course, such a family $\mathscr{F}$ should also include the entire set of elementary events Ω. Finally, if we wish to form sentences using 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 these motivations in mind, let us consider the following notion.

Definition 1.1 A *σ-algebra on Ω* is a non-empty family $\mathscr{F}$ of subsets of Ω such that

1. $A \in \mathscr{F}$ implies $A^c \in \mathscr{F}$;
2. $A_1, A_2, \ldots, A_n, \cdots \in \mathscr{F}$ implies $\cup_{n=1}^{\infty} A_n \in \mathscr{F}$.

   Moreover, if $\mathscr{F}$ satisfies it is closed *only under finite unions*, we say that $\mathscr{F}$ is an *algebra*

From now on, we will want that any set of random events be formally described by a σ-algebra. Then, after associating a σ-algebra $\mathscr{F}$ to a set of elementary events Ω, we will refer to the pair $(\Omega, \mathscr{F})$ 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 a set-operation property, known as the *de Morgan's laws*, 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. In particular,

Observation 1.3 — De Morgan's Laws. Given a sample set Ω and a sequence of sets $A_1, A_2, \ldots, A_n, \cdots \subseteq \Omega$, then

$$\cup_{n=1}^{\infty} A_n = \left(\cap_{n=1}^{\infty} A_n^c\right)^c \text{ and } \cap_{n=1}^{\infty} A_n = \left(\cup_{n=1}^{\infty} A_n^c\right)^c.$$

In summary, using set operations, the idea 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 complex descriptions which can be formed using much simpler ones.

■ **Example 1.3** [☕] We can assign probabilities to events/sentences such as:

1. A^c="A does not occur;"
2. $A \cap B$="both A and B simultaneously occur,"
3. $A \cup B$="either A, or B, or both, occurs,"
4. $(A \cap B^c) \cup (B \cap A^c)$="either A or B occurs, but not both of them simultaneously,"
5. $A \cap B = \emptyset$="A and B are mutually exclusive,"
6. $A \setminus B = A \cap B^c$="$A$ occurs but B does not occur,"
7. $(A \cup B)^c$="neither A nor B occur."

■

■ **Example 1.4** [☕☕] To see why it is important that a σ-algebra be closed under arbitrary (perhaps, an infinite number of) unions (or intersections), suppose that a die is rolled arbitrarily many times. Then, we would naturally consider the sample set $\Omega = \{1,\dots,6\} \times \{1,\dots,6\} \times \cdots = \{1,\dots,6\}^\infty$. Suppose that we want to study the event "number 2 comes up in the ith roll of the die". Then, we should certainly choose a σ-algebra on Ω with the requirement that it must contain all sets of the form

$$A_i = \{(\omega_i)_{i=1}^\infty \in \Omega \,:\, \omega_i = 2\}\,, \quad \text{for each} \;\; i = 1,2,\dots. \tag{1.1}$$

Now, the event B ="in neither the second nor the third roll number 2 comes up" can be described as $B = \left\{(\omega_i)_{i=1}^\infty \in \Omega \,:\, \omega_2 \neq 2, \omega_3 \neq 2\right\} = A_2^c \cap A_3^c$. Then, such an event B is included in a σ-algebra that contains the sets of the form in Eq. (1.1).. Similarly, the events "number 2 comes up at least once through the rolls," which can be described by $\cup_{i=1}^\infty A_i$, and "each roll results in number 2 coming up," which can be described by $\{(2,2,\dots)\} = \cap_{i=1}^\infty A_i$, are events that belong to a σ-algebra that contains the sets of the form in Eq. (1.1). ■

1.2.1 The σ-algebra generated by a family of sets (☕☕☕)

Given the requirements that a σ-algebra must satisfy, we observe that most sets of elementary events admit multiple σ-algebras that fit into Definition 1.1. The choice of a σ-algebra is not always obvious. Depending on the application at hand, we might want to choose our 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 family of all subsets of Ω, known as the *power class* 2^Ω. By choosing the power class 2^Ω, we make sure not to lose any random event 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 is 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:

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

If Ω is countably infinite, then we end up with an infinite sequence $\{\mathscr{F}_n\}_{n=1}^{\infty}$ of σ-algebras, ordered by set inclusion. If Ω is a continuum, then we end up with an ordered set of σ-algebras $\{\mathscr{F}_n\ : n \in N\}$, 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 $\mathscr{A}$ of subsets of Ω—which in fact does not need to be a σ-algebra of Ω in itself—, we could search for a family of subsets of Ω that contains $\mathscr{A}$, 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.*

Definition 1.2 The *σ-algebra generated by the family of sets $\mathscr{A}$ of a non-empty set Ω* is the family of sets

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

The following theorem establishes that, for a family of sets $\mathscr{A}$ of Ω, $\sigma(\mathscr{A})$ 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 $\mathscr{A}$ that is included in some σ-algebra $\mathscr{F}$.

Theorem 1.1 [☕☕☕] Given a nonempty family $\mathscr{A}$ of subsets of a nonempty set Ω, $\sigma(\mathscr{A})$ satisfies:

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

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

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

Proof of Theorem 1.1. (i) First, take $A \in \sigma(\mathscr{A})$, then $A \in \mathscr{F}_n$ for each $\mathscr{F}_n \supseteq \mathscr{A}$ and each $n \in N$. Since each $\mathscr{F}_n$ is a σ-algebra on Ω, we have that $A^c \in \mathscr{F}_n$ for each $\mathscr{F}_n \supseteq \mathscr{A}$ and each $n \in N$. Therefore, $A^c \in \sigma(\mathscr{A})$. Second, take a sequence $\{A_n\}_{n=1}^{\infty} \subseteq \sigma(\mathscr{A})$, then $\{A_n\}_{n=1}^{\infty} \subseteq \mathscr{F}_n$ for each $\mathscr{F}_n \supseteq \mathscr{A}$ and each $n \in N$. Since each $\mathscr{F}_n$ is a σ-algebra on Ω, we have that $\cup_{n=1}^{\infty} A_n \in \mathscr{F}_n$ for each $\mathscr{F}_n \supseteq \mathscr{F}$ and each $n \in N$. Therefore, $\cup_{n=1}^{\infty} A_n \in \sigma(\mathscr{A})$. (ii) The result follows directly from the definition of $\sigma(\mathscr{A})$, taking into account the set operations of inclusion and intersection. (iii) Take a σ-algebra $\mathscr{F}$ on Ω such that $\mathscr{F} \supseteq \mathscr{A}$. Then, it must be the case that $\mathscr{F} = \mathscr{F}_n$ for some $n \in N$ so that $\sigma(\mathscr{A}) \subseteq \mathscr{F}$. ■

1.2.2 Borel σ-algebras on Euclidean spaces (☕☕☕)

In many applications where outcomes are described by real numbers, it is usual to choose a very particular generated σ-algebra, 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.

Definition 1.3 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.

■ **Example 1.5** [☕☕☕] 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

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

■

Definition 1.4 Given a sample set 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 commonly known as *Borel sets*.

Notice that the notion 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, we usually want to begin using the notion of *neighborhood* as a primitive to propose the Euclidean topology.

Definition 1.5 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 $\omega, \omega', \omega'' \in \Omega$.
Then, the *Euclidean metric* in $\mathbb{R}^n$ is given by $d(\omega, \omega') = +\sqrt{\sum_{i=1}^{n}(\omega_i - \omega_i')^2}$.

The Euclidean metric is an intuitive metric that simply gives us the shortest geographical distance between two points—that is, the distance according to a straight line that connects the points. Given a number $\varepsilon > 0$, the set $B_d(\omega, \varepsilon) = \{\omega' : d(\omega, \omega') < \varepsilon\}$ is known as the *ε-neighborhood (or neighborhood of size ε) centered at ω.*

Definition 1.6 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 $B_d(\omega, \varepsilon) \in A$.
Then, given a set $\Omega \subseteq \mathbb{R}^n$, the *Borel σ-algebra on Ω*, is the generated σ-algebra $\sigma(\tau_\Omega)$. We will use $\mathscr{B}_\Omega$ 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 1.6** [☕☕☕] Consider the family of open intervals in $\mathbb{R}$,

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

We wish to show that $\sigma(\alpha) = \mathscr{B}_\mathbb{R}$. 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, $\sigma(\alpha) = \sigma(\tau_\mathbb{R}) = \mathscr{B}_\mathbb{R}$. ▪

▪ **Example 1.7** [☕☕☕] 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 $\sigma(\beta) = \mathscr{B}_{\mathbb{R}}$ 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, $\sigma(\beta) = \sigma(\tau_{\mathbb{R}}) = \mathscr{B}_{\mathbb{R}}$. ■

Observation 1.4 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 $\mathscr{B}_{\mathbb{R}}$, it follows that that singletons and countable sets in $\mathbb{R}$ are members of its Borel σ-algebra.

1.3 Measures and Probability Laws (☕☕)

To compute probabilities, we need to come up with a reasonable probability law. Suppose that we wish to compute the probability of occurrence of a certain event $A \subseteq \Omega$. 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 law (as a primitive to model uncertainty) 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 working tool, we will be able to compute probabilities for a wide variety of random phenomena or experiments.[3]

Definition 1.7 A *measure* P on a measurable space $(\Omega, \mathscr{F})$ is a set function $P : \mathscr{F} \to \mathbb{R}^*$ with $P(\emptyset) = 0$, $P(A) \geq 0$ for each $A \in \mathscr{F}$, and such that if $\{A_n\}_{n=1}^{\infty} \subseteq \mathscr{F}$ is a sequence of pairwise disjoint events in $\mathscr{F}$, then $P\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \sum_{n=1}^{\infty} P(A_n)$.

Then, a *probability measure* P is a measure that also satisfies $P(\Omega) = 1$.

At this point, we can present our workhorse to model randomness, which is commonly known as *probability space*. A *probability space* is a triplet $(\Omega, \mathscr{F}, P)$, where Ω is an arbitrary nonempty sample set, $\mathscr{F}$ is a σ-algebra of event from Ω, and P is a probability measure on $(\Omega, \mathscr{F})$.

In those cases where $\mathscr{F}$ is a Borel σ-algebra, with respect to some topology on the set Ω, and P is a probability measure on $(\Omega, \mathscr{F})$, then P is referred to as a *Borel probability measure*. In addition, 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*.

Definition 1.8 Given a closed interval $[a,b] \subset \mathbb{R}$, the *Lebesgue Measure on* $\mathbb{R}$ is

$$\lambda([a,b]) = b - a.$$

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 $A = \cup_{i=1}^{n}[a_i, b_i] \in \mathscr{B}_{\Omega}$, 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

$$\lambda([a,b]) = \prod_{i=1}^{n}(b_i - a_i).$$

The following properties can be derived from the definition of probability measure.

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

Observation 1.5 1. 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.

2. If $A \subseteq B$, then, we can write. $B = A \cup (A^c \cap B)$, where and A and $(A^c \cap B)$ are disjoint events. Therefore,

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

3. By invoking the set operations and 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

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

which, in turn, directly implies that $P(A \cup B) \leq P(A) + P(B)$. Now, we can use inductively the same logic above for two sets, to obtain a general expression for n sets, known as *inclusion-exclusion formula*:

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

Furthermore, the formula above leads directly to the inequality

$$P\Big(\cup_{i=1}^n A_i\Big) \leq \sum_{i=1}^n P(A_i),$$

known as *Boole's inequality*.

Finally, since the de Morgan's laws imply that $\cup_{n=1}^\infty A_n = \left(\cap_{n=1}^\infty A_n^c\right)^c$, 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.

■ **Example 1.8** [☕☕] 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 $\Omega = \{H,T\}^n$ 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 $B_1,\dots,B_n$ 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., $\cap_{i=1}^n B_i^c = \{(T,\dots,T)\}$. Then, $P(\cap_{i=1}^n B_i^c) = 2^{-n}$ so that the probability of our event of interest can be computed as $P(A) = 1 - 2^{-n}$. ■

1.4 Conditional Probability and Independence (☕)

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 we are aware that another event B has occurred?"

Definition 1.9 Given a probability space $(\Omega,\mathscr{F},P)$ and two events $A,B \in \mathscr{F}$ such that $P(B) > 0$, the *conditional probability of A given B* is

$$P(A|B) = \frac{P(A\cap B)}{P(B)}. \tag{1.2}$$

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

Observation 1.6 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 known as the *Law of Total Probability*.

Observation 1.7 Suppose that we begin working with a probability space $(\Omega, \mathscr{F}, P)$ and then take a given event $B \in \mathscr{F}$ such that $P(B) > 0$. Then, we can consider the σ-algebra restricted to such event B, $\mathscr{F}_B = \{A \in \mathscr{F} : A \cap B \neq \emptyset\}$, which includes all the events in the σ-algebra $\mathscr{F}$ that have some element(s) in common with set B. Then, it can be shown that the set function $P(\cdot|B) : \mathscr{F}_B \to [0,1]$ given by the definition of conditional probability, $P(A \mid B) = P(A \cap B)/P(B)$, is a well defined probability measure (check it!) on the measurable space $(B, \mathscr{F}_B)$. Therefore, $(B, \mathscr{F}_B, P(\cdot|B))$ gives us another probability space. Intuitively, if we are certain that event $B \in \mathscr{F}$ has happened, then we would like to switch from the original probability space $(\Omega, \mathscr{F}_B, P)$ to the new universe $(B, \mathscr{F}, P(\cdot|B))$ to compute probabilities.

■ **Example 1.9** [☕☕] Suppose that we roll a dice twice. The dice is fair so that all faces are equally likely to come out. The two dice rolls are independent from each other. We would like to compute the probability that, given that the outcome of the second roll is a higher number than the outcome of the first roll, the sum of the two roll outcomes is 5 and their difference is 1. To tackle this question, we can naturally consider an elementary event as $\omega = (\omega_1, \omega_2) \in \Omega = \{1, \ldots, 6\} \times \{1, \ldots, 6\}$, where ω_i captures the outcome of the ith roll of the dice. Then, $|\Omega| = 6 \cdot 6 = 36$ gives us the number of possible outcomes from rolling the dice twice. Since outcomes are equally likely, we can naturally consider that for any event $A \subseteq \Omega$ of the relevant sample space, $P(A) = |A|/|\Omega|$.

Now, consider the events/sentences $A =$"the sum of the two roll outcomes is 5," $B =$"the difference between the two roll outcomes is exactly 1," and $C =$"the second roll outcome resulted in a higher number than the first roll." We then wish to compute the probability $P(A \cap B \mid C)$. We can first notice that

$$A = \{(1,4), (4,1), (2,3), (3,2)\}$$

and

$$B = \{(1,2), (2,1), (2,3), (3,2), (3,4), (4,3), (4,5), (5,4), (5,6), (6,5)\}.$$

In addition,

$$C = \{(\omega_1, \omega_2) : \omega_2 > \omega_1 \text{ and } \omega_1, \omega_2 \in \{1, \ldots, 6\}\}.$$

Therefore, event C captures 5 possible outcomes when $\omega_1 = 1$, 4 possible outcomes when $\omega_1 = 2$, and so on until 1 possible outcome when $\omega_1 = 5$. In consequence, $|C| = 5+4+3+2+1 = 15$. Following Observation 1.7, we can now take event C, which contains 15 elementary events, as our sample space in a new restricted universe. Then, we observe that $A \cap B = \{(2,3),(3,2)\}$ so that $A \cap B \cap C = \{(2,3)\}$. Therefore, the sought probability $P(A \cap B \mid C)$ is equal to $1/15$. ■

The following example is a classical one that allows us to practice further with the idea of events with the probability properties presented earlier in Observation 1.5, and with the conditional probability implications presented in Observation 1.6.

■ **Example 1.10** [☕☕☕] Suppose that three blindfolded women throw their necklaces on a table. Then, the necklaces are randomly mixed and each woman randomly chooses a necklace. We ask about the probability that none of the women picks her own necklace. To tackle this questions, we start by considering the events/sentences A_i =“the ith woman picks her own necklace” for each woman named as $i = 1,2,3$. Then, “none of the women picks her own necklace” is the complement event of the event A=“some of the women pick her own necklace.” We observe that $A = A_1 \cup A_2 \cup A_3$. Suppose that the necklaces owned by each woman are initially labelled as $\{1,2,3\}$, exactly as the names of the women, so that woman i owns necklace i in that set. Then, we can think of an elementary event as a triplet $(\omega_1, \omega_2, \omega_3)$, where $\omega_i \in \{1,2,3\}$ is the necklace picked by the ith woman after the random mixed of them. In this case, our sample set would be

$$\Omega = \{(\omega_1, \omega_2, \omega_3) : \omega_i \in \{1,2,3\} \text{ for } i = 1,2,3, \text{ and } \omega_i \neq \omega_j \text{ for } i \neq j\}.$$

Therefore, we observe that

$$A_1 = \{(1,2,3),(1,3,2)\},\ A_2 = \{(1,2,3),(3,2,1)\}, \text{ and } A_3 = \{(1,2,3),(2,1,3)\}.$$

In addition, we can compute $|\Omega| = 3 \cdot 2 \cdot 1 = 3! = 6$ since three possible necklaces can be chosen by the first woman, two possible remaining necklaces can be chosen by the second woman, and only one possible necklace remains available for the third woman. We can then compute $P(A_i) = 2/6 = 1/3$ for each woman $i = 1,2,3$. Intuitively, the probability that each woman choses her own necklace is $1/3$ since she is equally likely to select any of the three available necklaces.

We now notice that the probability that woman 2 picks her own necklace given that woman 1 has selected her own necklace is $P(A_2 \mid A_1) = 1/2$ since two possible necklaces remain available for the second woman, and one of such necklaces is hers. Therefore, $P(A_1 \cap A_2) = P(A_2 \mid A_1)P(A_1) = 1/2 \cdot 1/3 = 1/6$.

We can now think in a recursive fashion (as stated in Observation 1.6) to compute the probability of event $A_1 \cap A_2 \cap A_3$, which describes the outcome that each woman ends up with her own necklace. Given that women 1 and 2 have picked their own necklaces, it is very intuitive to see that the probability that the third woman selects her own necklace, $P(A_3 \mid A_1 \cap A_2)$, is one since hers is the only remaining necklace. Then,

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) = 1/3 \cdot 1/2 \cdot 1 = 1/6.$$

Now, we can resort to the inclusion-exclusion formula stated in Observation 1.5 to compute

$$\begin{aligned} P(A_1 \cup A_2 \cup A_3) &= P(A_1) + P(A_2) + P(A_3) \\ &\qquad - P(A_1 \cap A_2) - P(A_2 \cap A_3) - P(A_1 \cap A_3) + P(A_1 \cap A_2 \cap A_3) \\ &= 1/3 + 1/3 + 1/3 - 1/6 - 1/6 - 1/6 + 1/6 = 2/3. \end{aligned}$$

Finally, we obtain that the probability that none of the three women picks her own necklace is $P(A^c) = 1 - 2/3 = 1/3$. ■

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 1.2 — Bayes' Rule. [☕] Let $(\Omega, \mathscr{F}, P)$ be a probability space and let $\{A_i\}_{i=1}^{\infty}$ be a sequence of events $A_i \in \mathscr{F}$ 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 $B \in \mathscr{F}$ such that $P(B) > 0$. Then,

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

Proof of Theorem 1.2. 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. ■

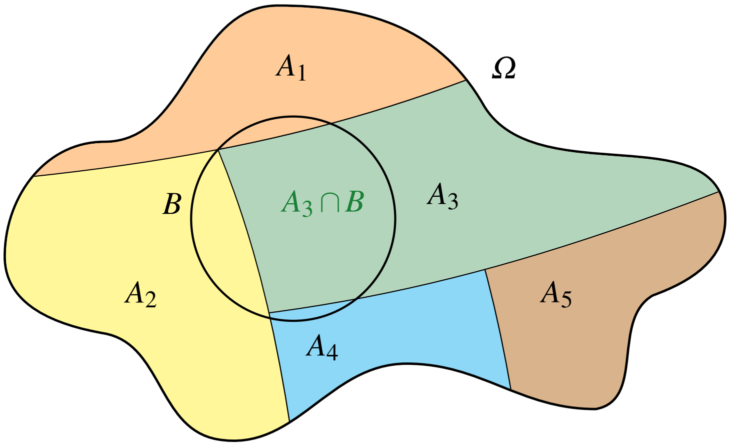

Figure 1.3: Illustration of Bayes' Rule.

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

■ **Example 1.11** 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

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

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

■

An interesting situation arises 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 $P(B) > 0$. Then, it is intuitive to view this sort of situations as A and B happening independently from each other—in probability terms. Independence is a key concept both in probability and inference theories. We can use the logic behind the idea of conditional probability, to state

Definition 1.10 Two events A, B are *independent events* if the probability of occurrence of event a does not depend on the occurrence of event B, $P(A \mid B) = P(A)$. Equivalently, A and B are independent if the probability of simultaneous occurrence of both events $P(A \cap B)$ can be decomposed multiplicatively as $P(A \cap B) = P(A)P(B)$.

The above is a simple criterion when one deals with a pair of events. However, the extension of this definition to more than two events $A_1, \dots, A_n$ is not straightforward. In particular, independence of a set of events $\{A_1, \dots, A_n\}$ requires much more than just the multiplicative decomposition $P(\cap_{i=1}^n A_i) = P(A_1) \times \cdots \times P(A_n)$. In particular,

Definition 1.11 A finite family $A_1, \dots, A_n$ 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 $k_1, \dots, k_j$, with $1 \leq k_1 < \cdots < k_j \leq n$, for each $2 \leq j \leq n$.

In short, 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.

■ **Example 1.12** [☕] To grasp 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 1.13** [☕☕] 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 $(\Omega, \mathscr{B}_\Omega, \lambda)$ 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, conditional probability and independence of events can actually have intuitive interpretations in terms of odds of occurrence.

■ **Example 1.14** [☕☕] Two people (Bob and Alice) choose at random a number in the interval $[0,2]$ under the "uniform probability law." Let us then consider the events: $A =$"the difference between the two numbers is greater than $1/3$"; $B =$"at least one of the numbers is greater than $1/3$"; $C =$"the two numbers are equal"; $D =$"only the number picked by Alice is greater than $1/3$." We wish to compute the probabilities $P(B)$, $P(C)$, and $P(A \cap D)$.

Using a graphical approach, we can compute the probability of events using Lebesgue measures. Then, we obtain

$$P(B) = 1 - P(\text{"both numbers are smaller than } 1/3\text{"}) = 1 - \frac{(1/3)(1/3)}{2 \cdot 2} = \frac{35}{36},$$

$$P(C) = \frac{\text{area of the segment } \left\{(x,y) \in [0,2]^2 \,:\, y = x\right\}}{2 \cdot 2} = \frac{0}{4} = 0,$$

$$P(A \cap D) = \frac{\text{area of the double shaded area}}{2 \cdot 2} = \frac{(5/3)(5/3)(1/2) + (4/3)(4/3)(1/2)}{4} = \frac{41}{72}.$$

■

■ **Example 1.15** [☕☕] 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

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

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

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

Finally, using the Total Probability Law, we can compute

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

■

In addition to computing conditional probabilities, we can use the following example to practice with visualization of events and probabilities in finite sample spaces.

■ **Example 1.16** [☕☕] Suppose that we roll a dice twice. The dice is fair so that all faces are equally likely to come out. The two dice rolls are independent from each other. We can naturally consider an elementary event as $\omega = (\omega_1, \omega_2) \in \Omega = \{1,\dots,6\} \times \{1,\dots,6\}$, where ω_i captures the outcome of the ith roll of the dice. Then, $|\Omega| = 6 \cdot 6 = 36$ gives us the number of possible outcomes from rolling the dice twice. Since outcomes are equally likely, we can naturally consider that for any event $A \subseteq \Omega$ of the relevant sample space, $P(A) = |A|/|\Omega|$.

Consider first the following events/sentences: $A =$"the sum of the two roll outcomes is 12," $B =$"at least one roll resulted in 6," and $C =$"at least one roll resulted in 1."

1. We ask whether or not events A and B are independent. Then, we notice that $A = \{(6,6)\}$ since the only way of obtaining a sum equal to 12 is that both rolls come out exactly with number 6 as outcome. Therefore, $P(A) = 1/36$. On the other hand, we observe that

$$B = A \cup \{(6,\omega_2) : \omega_2 \in \{2,\dots,6\}\} \cup \{(\omega_1,6) : \omega_1 \in \{2,\dots,6\}\},$$

   so that $|B| = 1+5+5 = 11$. Therefore, $P(B) = 11/36$. In addition, we observe that $A \subset B$ so that $B \cap A = A$. Therefore, $P(B \mid A) = P(B \cap A)/P(A) = P(A)/P(A) = 1 \neq P(B)$ so that events A and B are not independent.
2. We ask whether or not events A and C are independent. First, exactly as in the previous case, we observe that $|C| = 11$ and $P(C) = 11/36$ since there are exactly 11 ways of having at least outcome 1 turning out. However, we observe that $A \nsubseteq C$ since the sum attained by having at least one 1 showing up cannot be higher than number 7. Therefore, $C \cap A = \emptyset$ so that $P(C \mid A) = P(C \cap A)/P(A) = P(\emptyset)/P(A) = 0 \neq P(C)$. Thus, events A and C are not independent either.

Now, consider the events/sentences $D =$"the sum of the two roll outcomes is 9," $E =$"the difference between the two roll outcomes is exactly 1," and $F =$"the second roll outcome resulted in a higher number than the first roll."

1. We ask whether or not events E and F are independent. Then, we can notice that

$$E = \{(1,2),(2,1),(2,3),(3,2),(3,4),(4,3),(4,5),(5,4),(5,6),(6,5)\}.$$

   Therefore, $|E| = 10$ and $P(E) = 10/36$. On the other hand, we observe that

$$F = \{(\omega_1,\omega_2) : \omega_2 > \omega_1 \text{ and } \omega_1,\omega_2 \in \{1,\dots,6\}\}.$$

   Therefore, event F captures 5 possible outcomes when $\omega_1 = 1$, 4 possible outcomes when $\omega_1 = 2$, and so on until 1 possible outcome when $\omega_1 = 5$. In consequence,

$|F| = 5+4+3+2+1 = 15$ and $P(F) = 15/36$. In addition, we observe that $E \cap F = \{(1,2),(2,3),(3,4),(4,6),(5,6)\}$ so that $|E \cap F| = 5$ and $P(E \cap F) = 5/36$. Then, we can compute $P(F \mid E) = P(E \cap F)/P(E) = 1/2 \neq P(F) = 15/36$. Thus, E and F are not independent events.

2. Now, we ask whether or not events E and F are independent given event D. One way to tackle this question is to resort directly to the definition of independence of two events. Using the definition, the answer will be affirmative if we can verify that

$$P(E \cap F \mid D) = P(E \mid D)P(F \mid D).$$

In this case, we can notice that

$$D = \{(3,6),(6,3),(4,5),(5,4)\}.$$

Therefore, $E \cap D = \{(4,5),(5,4)\}$ so that $P(E \cap D) = 2/36$ and, since $P(D) = 4/36 \neq 0$, we have $P(E \mid D) = P(E \cap D)/P(D) = 1/2$. In addition, it can be noted that $F \cap D = \{(3,6),(4,5)\}$ so that $P(F \cap D) = 2/36$ and, since $P(D) = 4/36 \neq 0$, we have $P(F \mid D) = P(F \cap D)/P(D) = 1/2$. Finally, we observe that $E \cap F \cap D = \{(4,5)\}$ and, since $P(D) = 4/36 \neq 0$, we have $P(E \cap P \mid D) = 1/4$. By putting together all these computations, we obtain that

$$P(E \cap F \mid D) = 1/4 = P(E \mid D)P(F \mid D) = 1/2 \cdot 1/2.$$

Therefore, E and F are independent events given event D.

■

1.5 Practice Exercises

Exercise 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) $(\cap_{n=1}^{\infty} A_n)^c = \cup_{n=1}^{\infty} A_n^c$.

Exercise 1.2 Let $\mathscr{F}$ be a family of subsets of some nonempty set Ω.

(a) Suppose that $\Omega \in \mathscr{F}$ and that $A, B \in \mathscr{F}$ implies $A \setminus B \in \mathscr{F}$. Show that $\mathscr{F}$ is an algebra.

(b) Suppose that $\Omega \in \mathscr{F}$ and that $\mathscr{F}$ is closed under the formation of complements and finite *disjoint* unions. Show that $\mathscr{F}$ need not be an algebra.

Exercise 1.3 Let $\mathscr{F}_1, \mathscr{F}_2, \ldots$ be a family of subsets of some nonempty set Ω.

(a) Suppose that $\mathscr{F}_n$ are algebras satisfying $\mathscr{F}_n \subseteq \mathscr{F}_{n+1}$. Show that $\cup_{n=1}^{\infty} \mathscr{F}_n$ is an algebra.

(b) Suppose that $\mathscr{F}_n$ are σ-algebras satisfying $\mathscr{F}_n \subseteq \mathscr{F}_{n+1}$. Show by example that $\cup_{n=1}^{\infty} \mathscr{F}_n$ need not be a σ-algebra.

Exercise 1.4 Let $\Omega = \{(x,y) \in \mathbb{R}^2 : 0 < x, y \leq 1\}$, let $\mathscr{F}$ 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 $A \in \mathscr{B}_{(0,1]}$, 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 $(\Omega, \mathscr{F}, P)$ is a probability space.

Exercise 1.5 Let $(\Omega, \mathscr{F}, P)$ be a probability space and, for $A \in \mathscr{F}$, let $P_A : \mathscr{F} \to [0,1]$ be a set function defined by $P_A(B) = P(A \cap B)$ for each $B \in \mathscr{F}$.

(a) Show that, for a given $A \in \mathscr{F}$, P_A is a measure, but not a probability measure, on $(\Omega, \mathscr{F})$.

(b) Show that, for a given $A \in \mathscr{F}$ such that $P(A) > 0$, the set function Q_A on $\mathscr{F}$ defined by $Q_A(B) = P_A(B)/P(A)$ for each $B \in \mathscr{F}$ is a probability measure on $(\Omega, \mathscr{F})$.

Exercise 1.6 Let $P_1, \ldots, P_n$ be probability measures on some measurable space $(\Omega, \mathscr{F})$. 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 $(\Omega, \mathscr{F})$.

Exercise 1.7 Let $(\Omega, \mathscr{F}, P)$ be a probability space and let $A_1, \ldots, A_n$ be events in $\mathscr{F}$ 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 $P(\cap_{i=1}^{n} A_i) = 0$.

Exercise 1.8 Let $(\Omega, \mathscr{F}, P)$ be a probability space and let A, B and C be three events in $\mathscr{F}$ such that $P(A \cap B \cap C) > 0$. Show that $P(C|A \cap B) = P(C|B)$ implies $P(A|B \cap C) = P(A|B)$.

Exercise 1.9 Let $(\Omega, \mathscr{F}, P)$ be a probability space and let $A_1, \ldots, A_n$ be independent events in $\mathscr{F}$. Let $B_1, \ldots, B_n$ be another sequence of events such in $\mathscr{F}$ 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.

2. Random Variables and Distributions (☕)

In many probability applications, it is useful to have a tool that allows us to assign probabilities consistently over events in a general sample set Ω by computing probabilities over the real line instead. The notion of random variable is a key concept in probability theory that gives us a transformation through which we can move from computing probabilities over arbitrary sample sets to doing so over the real line.

2.1 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 instead. This approach allows 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 arbitrary events to events in the real line. The specification of such a function would be given by the particular features of the random experiment of interest. To illustrate this, suppose that we roll a dice ten times. In this case, the underlying elementary events would have the form $\omega = (\omega_1, \dots, \omega_{10}) \in \Omega = \{1, \dots, 6\}^{10}$. 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 $X(\omega) = \min_{i=1,\dots,10}\{\omega_i\}$.

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*.

Definition 2.1 A *random variable* on a measurable space $(\Omega, \mathscr{F})$ is a function $X : \Omega \to \mathbb{R}$ such that for each $B \in \mathscr{B}_{\mathbb{R}}$, we have

$$X \in B = \{\omega \in \Omega : X(\omega) \in B\} \in \mathscr{F},$$

where we use $X \in B$ as short-hand notation for the inverse image of the set B through the function X.

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 $B \in \mathscr{B}_{\mathbb{R}}$, the inverse image $X \in B = \{\omega \in \Omega : X(\omega) \in B\}$ actually lies in the original σ-algebra $\mathscr{F}$. However, once we have understood the logic behind the notion of random variable, we can simply use a random variable X to describe a set of possible outcomes of a random experiment—under the condition that such outcomes are expressed as real numbers.

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

Definition 2.2 Given a probability space $(\Omega, \mathscr{F}, P)$ and a random variable X a random variable on $(\Omega, \mathscr{F})$, the associated *probability distribution of the random variable* X is a probability measure ψ on $(\mathbb{R}, \mathscr{B}_{\mathbb{R}})$ 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 2.1** [☕☕] 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 $(\Omega, \mathscr{F}, P)$, where $\omega = (\omega_1, \omega_2, \omega_3) \in \Omega = \{1, \ldots, 6\}^3$, $\mathscr{F} = 2^{\Omega}$, and P is specified by $P(A) = |A|/6^3$ for each $A \in \mathscr{F}$. However, the measurable space $(\mathscr{F}, P)$ 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 $X((\omega_1, \omega_2, \omega_3)) = \omega_1 + \omega_2 + \omega_3$. Since $\mathscr{F} = 2^{\Omega}$, this function X is a random variable because for each subset $B \subseteq [3,18]$, we can guarantee that

$$[X \in B] = \{(\omega_1, \omega_2, \omega_3) \in \{1, \dots, 6\}^3 \,:\, \omega_1 + \omega_2 + \omega_3 \in B\} \in 2^{\Omega}.$$

Now, consider the event $B = (3,5] \in \mathscr{B}_{\mathbb{R}}$. 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 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. In such cases, the relevant events are subsets of a multidimensional Euclidean space.

Definition 2.3 A *random vector* on a measurable space $(\Omega, \mathscr{F})$ is a function $X = (X_1, \dots, X_n) : \Omega \to \mathbb{R}^n$ such that for each $B \in \mathscr{B}_{\mathbb{R}^n}$, we have

$$X \in B = \{\omega \in \Omega \,:\, (X_1(\omega), \dots, X_n(\omega)) \in B\} \in \mathscr{F}.$$

Furthermore, it can be proved that a function $X = (X_1, \dots, X_n) : \Omega \to \mathbb{R}^n$ is a random vector on some $(\Omega, \mathscr{F})$ if and only if each of its components X_i is a random variable on $(\Omega, \mathscr{F})$. In other words, a random vector is simply a list of random variables. The idea of probability distribution works analogously for a random vector.

Definition 2.4 The *probability distribution of the random vector* $X = (X_1, \dots, X_n)$ associated to an underlying probability space $(\Omega, \mathscr{F}, P)$ is a probability measure ψ on $(\mathbb{R}^n, \mathscr{B}_{\mathbb{R}^n})$ specified by

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

2.2 Distribution Functions (☕)

Sometimes it is useful to use an alternative formulation to compute the probabilities captured by the probability distribution of a random variable.

Definition 2.5 The *distribution function* of the random variable X given by the probability distribution ψ is the function $F : \mathbb{R} \to \mathbb{R}$ specified by $F(x) = \psi((-\infty, x]) = P(X \leq x)$.

In words, a distribution function is simply a function that gives probabilities only of Borel sets which have the particular 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 proof of this result is beyond the scope of this notes.

Theorem 2.1 — Billingsley (1995). [☕☕☕] Let ψ be a probability measure on a measurable space $(\Omega, \mathscr{F})$, 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 $\mathscr{F}$ cannot contain an uncountable disjoint collection of events $\{A : A \in \mathscr{F}\}$ with $\psi(A) > 0$.

Observation 2.1 Some useful properties of a distribution function F can be derived directly from Theorem 2.1 above. First, from the result that $A \subseteq B$ implies $\psi(A) \leq \psi(B)$, we learn that F is monotone nondecreasing. Secondly, by continuity from above of the probability measure ψ on $(\mathbb{R}, \mathscr{B}_{\mathbb{R}})$ (Theorem 2.1 (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 $(\mathbb{R}, \mathscr{B}_{\mathbb{R}})$ (Theorem 2.1 (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 2.1, leads to that the distribution function F can have at most countably many points of discontinuity. Finally, another property that follows from Theorem 2.1 is that $\lim_{x\to-\infty}F(x)=0$ and $\lim_{x\to+\infty}F(x)=1$.

Furthermore, the converse implication follows. More precisely, any function $F:\mathbb{R}\to\mathbb{R}$ that satisfies the properties described in 2.1 is in fact a distribution function.

Theorem 2.2 [☕☕] Let $F:\mathbb{R}\to\mathbb{R}$ be a monotone nondecreasing, right-continuous function satisfying

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

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

Proof of Theorem 2.1. We consider the probability space $((0,1),\mathscr{B}_{(0,1)},P)$, where P is the Lebesgue measure on $((0,1),\mathscr{B}_{(0,1)})$. To grasp the logic of 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 $\nu:(0,1)\to\mathbb{R}$ be the inverse mapping $\nu=F^{-1}$. 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 $((0,1),\mathscr{B}_{(0,1)})$. 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 $((0,1),\mathscr{B}_{(0,1)},P)$ and that $P(X \leq x) = F(x)$. ■

For many applications, the previous results allow us 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 $b > a$, $P(a < X \leq b) = F(b) - F(a) = -0.2$.

▪ **Example 2.2** [☕] Suppose that we roll an unfair five side dice. The random variable X describes the outcome of the roll. The underlying probability distribution is described in Figure 2.4, which plots the distribution function $F(x) = P(X \leq x)$ for each possible outcome $x = 1,\dots,5$ of the roll. We then observe that the probability that the outcome of the roll be higher than number 2 is $P(X > 2) = 1 - F(2) = 1 - 0.4 = 0.6$. In addition the probability that the outcome of the roll be be either number or number 3 is $P(X \in \{2,3\}) = P(1 < X \leq 3) = F(3) - F(2) = 4/15 - 0.2 = 1/15$.

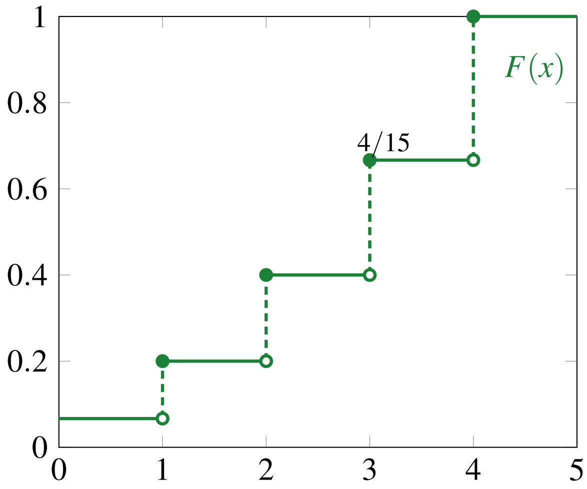

Figure 2.1: Illustration of distribution function.

■

We will need to resort to multidimensional distribution functions when we are interested in probability calculations over several features, that is, in situations where we are dealing with random vectors.

Definition 2.6 The *joint distribution function* of the random vector $X = (X_1,\dots,X_n)$ with probability distribution ψ is the function $F : \mathbb{R}^n \to \mathbb{R}$ specified by $F(x_1,\dots,x_n) = \psi(S_x) = P(X_1 \leq x_1,\dots,X_1 \leq x_1)$, where $S_x = \{(y_1,\dots,y_n) \in \mathbb{R}^n \, : \, y_i \leq x_i, \; i = 1,\dots,n\}$ is the Euclidean region of points “southwest” with respect to x.

2.3 Discrete Random Variables (☕)

When a random phenomenon can generate 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. We now present the formal ingredients.

Definition 2.7 A *support* for a probability measure ψ on $(\mathbb{R}, \mathscr{B}_{\mathbb{R}})$ is a Borel set $S \in \mathscr{B}_{\mathbb{R}}$ satisfying $\psi(S) = 1$.

More intuitively, a support of a random variable X is a simply set that includes all realizations that can generate the random phenomena captured by X. Of course, a probability measure can admit an infinite number of different supports an we need further qualification here. More precisely, if S is a support of a probability measure ψ and $S \subset T \in \mathscr{B}_{\mathbb{R}}$, 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. In other word, we are mainly interested in the set that includes exactly all realizations that can generate the random phenomena, disregarding realizations that have zero probability of occurrence. Then, when we wish to consider outcomes of a random phenomena that have actually positive probability of occurrence, we focus on the the particular support $S \in \mathscr{B}_{\mathbb{R}}$ of the corresponding distribution ψ such that $S' \subset S$ implies $\psi(S) < 1$. From the result of Theorem 2.1 (ii), we know that the minimal support $\mathrm{supp}(\psi)$ of a probability measure ψ is unique. Then, we will use $\overline{X}$ 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 the random variable X when we think of it as a function $X : \Omega \to \mathbb{R}$. Thus, we can restrict attention to the function $X : \Omega \to \overline{X}$ as realizations $x \notin \overline{X}$ will have zero probability of occurrence.

Definition 2.8 A random variable X and its probability distribution ψ are said to be *discrete* if ψ has a countable support $S = \{x_1, x_2, \dots, x_n, \dots\}$. Furthermore if ψ has a finite support, then the corresponding random variable is said to be a *simple random variable*.

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 $i = 1, 2, \dots$. For a discrete random variable

X, the function $f : \overline{X} \to [0,1]$ 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 $\sum_{x_i \in \overline{X}} f(x_i) = 1$. 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 2.3** [☕☕] This example uses some ideas of combinatorics, as described in chapter 9. 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

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

where we follow the convention that $\{1, \dots, a\}$ are the red balls and $\{a+1, \dots, a+b\}$ 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

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

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

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

We will see later that the density function above corresponds to 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 $X = (X_1, \dots, X_n)$, the function $f : \mathbb{R}^n \to \mathbb{R}$ defined as

$$f(x_1, \dots, x_n) = P(X_1 = x_1, \dots, X_n = x_n),$$

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 the vector. This idea is captured by the concept of *marginal distribution of the random vector*. For the case of a discrete random vector $X = (X_1, \dots, X_n)$, we derive the *marginal discrete density function* of a random variable X_i simply as

$$f_i(x_i) = \sum_{\overline{X}_1} \cdots \sum_{\overline{X}_{i-1}} \sum_{\overline{X}_{i+1}} \cdots \sum_{\overline{X}_n} f(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$$

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 2.4** [☕☕] Consider a random vector (X,Y) with joint discrete density function

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

where c is some real number. Note first that

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

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

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

■

2.4 Continuous Random Variables (☕)

Unlike the random experiments suitably captured by discrete random variables, other random phenomena exhibit the property that the sets of their possible outcomes are uncountable. For example, many economic models, assume for mathematical tractability that consumption and production sets, or set of prices are uncountable. The most common way to deal with random variables in these cases is to consider *continuous random variables*.

Definition 2.9 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

$$P(X\in B)=\psi(B)=\int_B f(x)dx \;\text{ for each }\; B\in\mathscr{B}_{\mathbb{R}}. \tag{2.1}$$

The function f is referred to as *density function (with respect to Lebesgue measure)*.[1] Of course, if $B=\mathbb{R}$ in the expression (Eq. (2.1)) above, then f must integrate to 1.

[1] Alternatively, when a notion of integral more general than the *Riemann Integral* is required, the probability computation expressed in condition (Eq. (2.1)) above can be written as

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

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

Observation 2.2 Since the Borel σ-algebra $\mathscr{B}_{\mathbb{R}}$ can be generated by the family of all intervals with the form $(a,b] \subset \mathbb{R}$, it follows from the celebrated Carathéodory's Theorem (Theorem 11.1) that the requirement in (Eq. (2.1)) 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

$$P(a < X \leq b) = F(b) - F(a) = \int_a^b f(x)dx \qquad \text{for each } \ a < b. \tag{2.2}$$

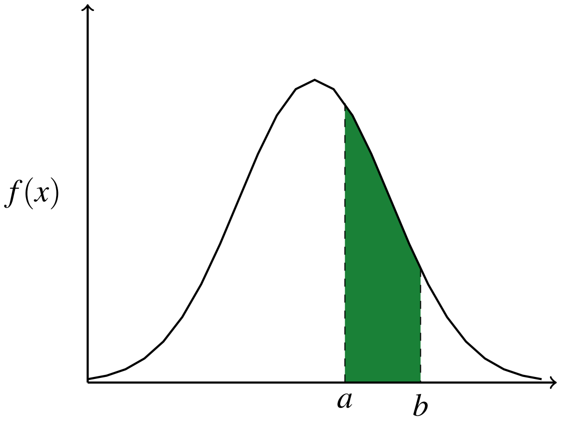

Figure 2.2: Illustration of probability computation for a continuous random variable following Eq. (2.2).

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 (Eq. (2.2)) above. In particular, the distribution function is not required to differentiate everywhere, neither F or f are required to 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 (Eq. (2.2)) 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. The support of continuous random variable X with density function f is specified as $\overline{X} = \{x \in \mathbb{R} : f(x) > 0\}$. Intuitively, the support of the random variable includes only Borel sets whose probability of occurrence is positive according to the corresponding density.

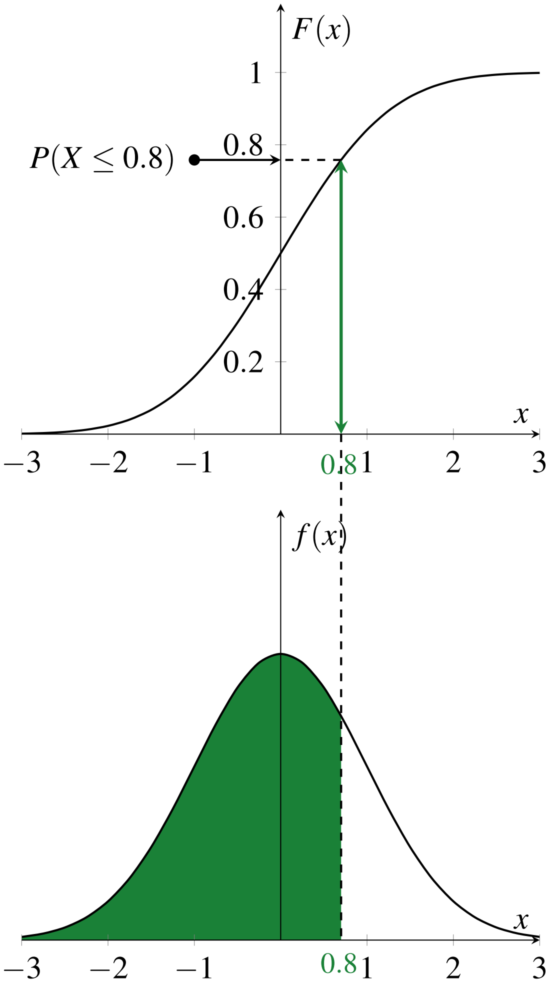

Figure 2.3: Density and distribution function of a continuous random variable. $P(X \leq 0.8) = F(0.8) = \int_{-\infty}^{0.8} f(x)dx$.

Observation 2.3 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. Leaving aside the technical differences we can appreciate the similarities between discrete and continuous random variables by comparing 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 $B \in \mathscr{B}_{\mathbb{R}^n}$ as

$$P((X_1, \ldots, X_n) \in B) = \underset{B}{\int \cdots \int} f(x_1, \ldots, x_n) dx_1 \cdots dx_n.$$

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

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

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.

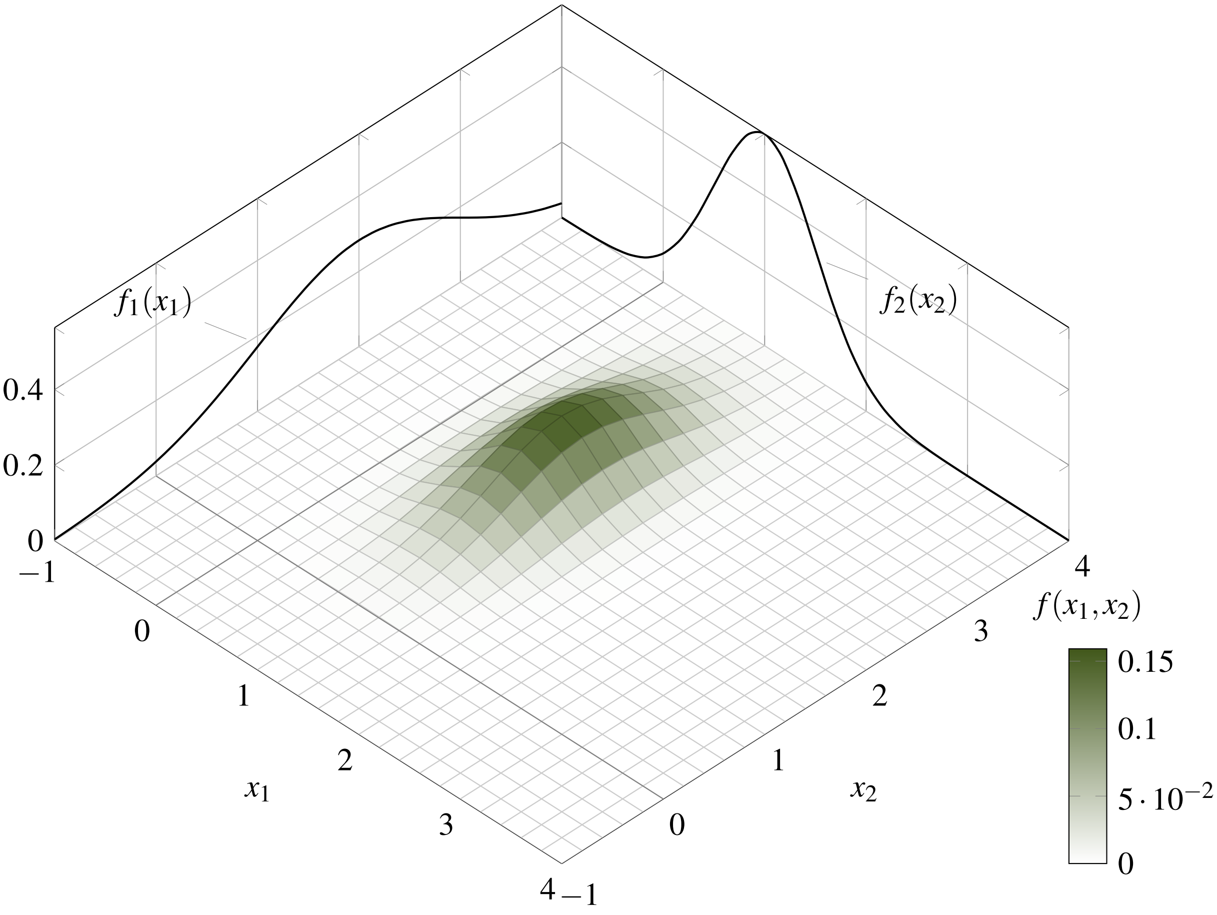

Figure 2.4: Illustration of bidimensional joint density.

In a way totally analogous to the discussion of ideas of conditional probability in Section 5.8, 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.

Definition 2.10 Consider a discrete random variable X. Then, 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 $P(Y \in B) = \sum_{x \in \overline{X}} P(\{X = x\} \cap \{Y \in B\})$, we see that $\sum_{x \in \overline{X}} f_{X|B}(x) = 1$, 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.

Definition 2.11 Consider a continuous random variable X. Then, 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:

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

Since $P(Y \in B) = \int_{\overline{X}} P(\{X \in A\} \cap \{Y \in B\})$, we observe, by letting $A = \overline{X}$ 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, $f_{X|y}(x) = f(x,y)/f_2(y)$.

■ **Example 2.5** [☕☕] Consider a continuous random pair (X,Y) with joint density

$$f(x,y) = \begin{cases} ax & \text{if } 1 \le x \le y \le 2, \\ 0 & \text{otherwise,} \end{cases}$$

where a is an undetermined parameter. Then, we first would like to know which value must take parameter a for $f(x,y)$ to be actually a density. By integrating over the described support we would need to obtain $\int\int_{(\overline{X,Y})} f(x,y)dxdy = 1$. Then, we can compute

$$\begin{aligned} \int_{x=1}^{x=2} \Big(\int_{y=x}^{y=2} axdy \Big) dx = \int_{x=1}^{x=2} ax(2-x)dx \\ &= a\big[x^2 - x^3/3\big]_{x=1}^{x=2} = a\big[2^2 - 2^3/3 - 1^2 + 1^3/3\big] \\ &= a[2/3] = 1 \;\; \Rightarrow \;\; a = 3/2. \end{aligned}$$

Secondly, we would like to obtain the marginal densities of X and Y. We can then compute, for the support $x \in [1,2]$,

$$f_1(x) = \int_{y=2}^{y=x} (3/2)x dy = (3/2)\left[y\right]_{y=2}^{y=x} = (3/2)\left(x-2\right).$$

Similarly, for the support $y \in [1,2]$, we obtain

$$f_2(y) = \int_{x=1}^{x=y} (3/2)x dx = (3/2)\left[x^2/2\right]_{x=1}^{x=y} = (3/4)\left(y^2-1\right).$$

Thirdly, we would like to obtain the conditional density of X given $Y = y$. Using the definition $f_{X|y}(x) = f(x,y)/f_2(y)$, we can compute

$$f_{X|y}(x) = \frac{(3/2)x}{(3/4)\left(y^2-1\right)} = \frac{x}{2(y^2-1)}$$

for the support $1 \leq x \leq y \leq 2$. Note that the density above is not well-defined for $x = y = 1$. ■

■ **Example 2.6** [☕☕] Consider a person that earns a random wage of $X \in [0,40]$ dollars each day according to a density $f_1(x) = x/800$. This person takes all the earned wage $X = x$ to the casino each night and ends up with a random amount Y. At worst, this person can lose all the entire wage wage x and thus end up with nothing, $y = 0$. At best, this person can earn twice the wage $y = 2x$. Ending up with each possible outcome $y \in [0,2x]$ is equally likely. We would like to know the probability that, on a given night, this person wins a positive amount of money from the casino.

To answer this question, note first that, since ending up with each outcome $y \in [0,2x]$ is equally likely, this person ends the night with an amount of Y dollars that follows a density $f_{Y|x}(y) = 1/(2x)$ for $y \in [0,2x]$. Then, we would like to compute the probability that $P(Y > X)$. To compute this probability, we need first to derive the joint density $f(x,y)$. Using the definition $f_{Y|x}(y) = f(x,y)/f_1(x)$, we obtain

$$f(x,y) = f_{Y|x}(y)f_1(x) = (1/2x) \cdot (x/800) = 1/1600$$

for the support $x \in [0,40]$ and $0 \leq y \leq 2x$. Then, we can compute

$$\begin{aligned} P(Y > X) &= \int_{x=0}^{x=40} \int_{y=x}^{y=2x} (1/1600) dy dx \\ &= (1/1600) \int_{x=0}^{x=40} 2x dx = (1/1600)\left[x^2/2\right]_{x=0}^{x=40} = 1/2. \end{aligned}$$

■

2.5 Functions of a Random Variables (☕☕)

In many applications, we wish to obtain the probability distribution of some transformation $Y = g(X)$ of a random variable X. To pursue this approach, we must verify first that $Y = g(X)$ is indeed a random variable. Then, given the requirement of the definition of a random variable, it turns helpful to to restrict attention to cases where the transformation g corresponds to a one-to-one function.

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

▪ **Example 2.7** [☕] 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

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

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.

Observation 2.4 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. 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

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

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

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

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

$$H(y) = P(g(X) \le y) = P(g(X) < y) = P(X > T(y)) = 1 - F\big(T(y)\big),$$

so that

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

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

$$h(y) = f\big(T(y)\big)\left|T'(y)\right|.$$

The above arguments serve as an informal proof of the following useful result.

Theorem 2.3 [☕] 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 2.3 can be applied.

■ **Example 2.8** [☕] 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 2.3, we can use the reasoning followed in its proof to obtain the density of a transformation $Y = g(X)$ even in cases where g is not one-to-one, as the following example shows.

■ **Example 2.9** [☕☕] 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

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

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 $\overline{Y} = \mathbb{R}_+$, we obtain that $Y = X^2$ has density

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

As we shall see later, the density $h(y)$ obtained above corresponds to a distribution known as *Chi-square* with parameter 1. ■

Even further, when the transformation g is neither one-to-one nor continuous, the logic behind the result of the Theorem 2.3 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 2.4 [☕☕] 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 $\overline{X}$ 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 ($i = 1,2,\dots,k$);

(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 $i = 1,2,\dots,k$;

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

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

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

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 2.3, 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 $X_1, \dots, X_n$, with joint density function f, and transform them into "new" random variables $Y_1, \dots, Y_n$ using the functions

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

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

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

Under these conditions we can state the following useful result, which is simply a generalization of Theorem 2.3 for the multidimensional case.

Theorem 2.5 [☕☕] 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 2.5 above

▪ **Example 2.10** [☕] 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 2.5 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,

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

▪

2.6 Independence of Random Variables (☕)

The idea of independent random events applies as well to random phenomena captured by a multidimensional random variable. In many applications, the odds of occurrence of the outcomes described by a random variable do not affect the odds of outcomes described by another random variable. This case can be formally characterized by several, seemingly different, conditions. One of the traditional definitions of independence of random variables is the following.

Definition 2.12 A set of random variables $X_1,\dots,X_n$ are *independent* if the probability of the simultaneous occurrence of Borel sets with the from $(a_i,b_i]$, $i=1,\dots,n$, can be decomposed as the product of the probabilities of these events, that is, if

$$P\big(X_1\in(a_1,b_1],\dots,X_n\in(a_n,b_n]\big)=P\big(X_1\in(a_1,b_1]\big)\times\cdots\times P\big(X_n\in(a_n,b_n]\big). \quad (2.3)$$

This definition covers 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 (Eq. (2.3)) above. To see this, suppose that we wish to verify whether a subset of random variables $\{X_{j_1},\dots,X_{j_k}\}\subset\{X_1,\dots,X_n\}$ meets the required multiplicative decomposition of probabilities. Then, we only need to consider the condition in the definition in (Eq. (2.3)) 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 $m\notin\{1,\dots,k\}$.

As mentioned above, other alternative formulations of independence or random variable are common in probability theory and its applications. The message conveyed by the following Theorem plays a central role around all such alternative formulations.

Theorem 2.6 [☕☕] Consider a set of elementary events Ω and a set of algebras $\mathscr{A}_1,\dots\mathscr{A}_n$ on Ω. If each collection of events $A_1,\dots,A_n$, with $A_i\in\mathscr{A}_i$ for $i=1,\dots,n$, is independent, then each collection of events $B_1,\dots,B_n$, with $B_i\in\sigma(\mathscr{A}_n)$ for $i=1,\dots,n$ 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 (Eq. (2.3)) 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 (Eq. (2.3)) 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 2.6, 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 (Eq. (2.3)) 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

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

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

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

$$P(X_1 = x_1,\dots,X_n = x_n) = P(X_1 = x_1) \times \cdots \times P(X_n = x_n)$$
$$\Leftrightarrow f(x_1,\dots,x_n) = f_1(x_1) \times \cdots \times f_n(x_n).$$

As for the case where the random variables $X_1,\dots,X_n$ are continuous, we can simply resort to the multiplicative decomposition of the corresponding joint distribution function and to the result of the famous Fubbini'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 $X_1,\dots,X_n$ are independent if and only if

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

Beyond these quite common formulations, the result in Theorem 2.6 enables us to state a general necessary condition[2] of independence of random variables. In particular, notice that, if we know that the random variables $X_1,\dots,X_n$ are independent, then it must be the case that

$$P(X_1 \in B_1,\dots,X_n \in B_n) = P(X_1 \in B_1) \times \cdots \times P(X_n \in B_n)$$

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

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 2.5 would not apply, as the following example illustrates.

■ **Example 2.11** [☕] 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

[2] 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.

$Y = \min\{X_1, X_2, X_3\}$. 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 $y > 0$. ■

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 2.12** [☕☕] 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 2.5 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 $V = Y$:

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

■

Observation 2.5 The notion of independency of 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 $i = 1,\dots,n$, then the random vectors $X_1,\dots,X_n$ are *independent* if, for each $x_1 \in \mathbb{R}^{k_1},\dots,x_n \in \mathbb{R}^{k_n}$, we have

$$P(X_1 \leq x_1,\dots,X_n \leq x_n) = P(X_1 \leq x_1) \times \cdots \times P(X_n \leq x_n).$$

2.7 Practice Exercises

Exercise 2.1 A die is rolled 12 times. Compute the probability of getting at most 3 fours.

Exercise 2.2 Let X be a random variable on some probability space $(\Omega, \mathscr{F}, P)$ and let $g : \mathbb{R} \to \mathbb{R}$ be a one-to-one function. Show that $Y = g(X)$ is a random variable on $(\Omega, \mathscr{F}, P)$.

Exercise 2.3 Let $F_1, \ldots, F_n$ be distribution functions on some probability space $(\Omega, \mathscr{F}, P)$. 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 $(\Omega, \mathscr{F}, P)$.

Exercise 2.4 Let X be a continuous random variable on some probability space $(\Omega, \mathscr{F}, P)$ with density $f(x) = 1/2e^{-|x|}$ for $x \in \mathbb{R}$. Compute $P(X \geq 0)$, $P(|X| \leq 2)$, and $P(1 \leq |X| \leq 2)$.

Exercise 2.5 Any point in the interval $[0,1)$ can be represented by its decimal expansion $.x_1 x_2 \ldots$. 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.

Exercise 2.6 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.

Exercise 2.7 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.

Exercise 2.8 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) $P(X \geq 3|X > 0)$.

Exercise 2.9 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.

Exercise 2.10 Let X be a random variable on some probability space $(\Omega, \mathscr{F}, P)$ such that $P(|X-1| = 2) = 0$. Express $P(|X-1| \geq 2)$ in terms of the distribution function F of X.

Exercise 2.11 Show that the distribution function F of a random variable is continuous from the right and that

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

Exercise 2.12 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.

Exercise 2.13 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 $F(m) = 1/2$.

Exercise 2.14 Let X be a random variable (on some probability space) with distribution function

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

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

Exercise 2.15 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 $F'(x) = f(x)$?

Exercise 2.16 Let X be a continuous random variable with density f. Find a formula for the density of $Y = |X|$.

Exercise 2.17 Let X be a positive continuous random variable with density f. Find a formula for the density of $Y = 1/(X+1)$.

Exercise 2.18 Let T be a positive continuous random variable on some probability space $(\Omega, \mathscr{F}, P)$. 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 $s,t>0$,

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

Exercise 2.19 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 $X\le 0$ and $g(X)=1-X$ for $X>0$.

Exercise 2.20 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.

Exercise 2.21 Let be a (X_1,X_2) random vector with joint density

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

supported on the set $U=\{(x_1,x_2)\in\mathbb{R}^2 : x_1>0, 0<x_2<2\pi\}$. 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 $g(\mathbb{R}_{++})=\{(y_1,y_2)\in\mathbb{R}^2 : y_1>0, 0<y_2<2\pi\}$. 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) $h(y_1,y_2)=(2\pi)^{-1}y_1e^{-y_1^2/2}$, supported in $\{(y_1,0)\in\mathbb{R}^2 : y_1\ge 0\}$;
(b) $h_1(y_1)=y_1e^{-y_1^2/2}$, supported in $(0,+\infty)$;
(c) $h_2(y_2)=(2\pi)^{-1}$, supported in $(0,2\pi)$.

Exercise 2.22 Let X and Y be two 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 $\mathbb{R}$.

Exercise 2.23 Suppose that X and Y are independent continuous random variables. Derive formulas for the joint density for $(X+Y, X)$, the density of $X+Y$, and the density of $Y-X$.

Exercise 2.24 Let X and Y be 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.

Exercise 2.25 Let X and Y be two independent continuous random variables (on some probability space $(\Omega, \mathscr{F}, P)$) having the same density each, $f(x) = g(y) = 1$ for $x, y \in (0, 1]$. Find

(a) $P(|X-Y| \leq 0.5)$;

(b) $P\left(\left|\frac{X}{Y} - 1\right| \leq .05\right)$;

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

Exercise 2.26 Let X and Y be 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.

Exercise 2.27 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.

Exercise 2.28 Let X, Y and Z be 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?

3. Expected Values and Moments of Distributions (☕)

The information contained in the probability distribution of a random variable can often be summarized by some characteristics of the distribution, such as its shape and location. Such characteristics are in most cases described by numbers known as the *moments of the distribution*. One of the moments used in many applications is 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.

Definition 3.1 Suppose that X is a discrete random variable with discrete density f. Then, its *expected value* is

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

Suppose that X is a continuous random variable with density f. Then, its *expected value* is

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

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.

Observation 3.1 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

$$E[X|Y=y] = \int_{\mathcal{X}} x f_{X|y}(x)dx.$$

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

■ **Example 3.1** [☕] Let us go back to the random pair (X,Y) with joint density

$$f(x,y) = \begin{cases} 3x/2 & \text{if } 1 \leq x \leq y \leq 2, \\ 0 & \text{otherwise,} \end{cases}$$

which we considered earlier in Example 2.5. Suppose that we want to compute the conditional expectation of the inverse of the random variable X given that the random variable Y takes the value $3/2$, that is, $E[1/X \mid Y = 3/2]$. Recall that we obtained in Example 2.5 the conditional density

$$f_{X|y}(x) = \frac{(3/2)x}{(3/4)\left(y^2-1\right)} = \frac{x}{2(y^2-1)}$$

for the support $1 \leq x \leq y \leq 2$. Then, we can compute

$$E[1/X \mid Y = 3/2] = \int_{x=1}^{x=3/2} \frac{1}{x} \cdot \frac{x}{2((3/2)^2-1)} dx = 8/5\left[x\right]_{x=1}^{x=3/2} = 4/5.$$

■

■ **Example 3.2** [☕☕☕] Suppose that X and Y are two continuous random variables with joint density

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

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

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

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

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

Thus, we obtain

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

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 3.3** [☕] Let us go back to the pair of random variables specified in Example 2.4. These were two discrete random variables (X,Y) with joint density function

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

so that

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

Also, let us consider the event $A =$ "$X \geq Y$." We wish to obtain $E[X]$, $\text{Var}[X]$, $E[X \cdot Y]$, $E[X\,|\,A]$, and $\text{Var}[X\,|\,A]$. First, recall from Example 2.4 that

$$f_1(x) = \begin{cases} \frac{12}{72} & \text{for } x = 1 \\ \frac{18}{72} & \text{for } x = 2 \\ \frac{42}{72} & \text{for } x = 4. \end{cases}$$

Then, we have

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

and

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

To compute $E[X \cdot Y]$, note first that $E[X \cdot Y] = E[X] \cdot E[Y]$ does not follow necessarily.[1] 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 $E[X\,|\,A]$ and $\text{Var}[X\,|\,A]$, we have first to compute the conditional density

$$f_{X|A}(x) = \frac{P(\{X = x\} \cap A)}{P(A)} \quad \text{for} \quad x \in \{1,2,4\}.$$

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

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,

$$f_{X|A}(x) = \begin{cases} \frac{(2/72)}{(49/72)} & \text{for } x = 1 \\ \frac{(5/72)}{(49/72)} & \text{for } x = 2 \\ \frac{(42/72)}{(49/72)} & \text{for } x = 4 \end{cases} = \begin{cases} \frac{2}{49} & \text{for } x = 1 \\ \frac{5}{49} & \text{for } x = 2 \\ \frac{42}{49} & \text{for } x = 4. \end{cases}$$

Then, we obtain

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

and

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

so that

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

■

Observation 3.2 Note that the expected value $E[X \mid Y]$ is a random variable itself, where Y captures an uncertain outcome. It makes sense to compute the expected value of such a random variable over the support of Y. In particular, we obtain

$$\begin{aligned} E[E[X \mid Y]] &= \int_{\mathbb{Y}} E[X \mid Y = y] f_2(y) dy \\ &= \int_{\mathbb{Y}} \Big[\int_{\mathbb{X}} x f_{X|y}(x) dx \Big] f_2(y) dy \\ &= \int_{\mathbb{X}} x \Big[\int_{\mathbb{Y}} \frac{f(x,y)}{f_2(y)} f_2(y) dy \Big] dx \\ &= \int_{\mathbb{X}} x f_1(x) dx = E[X] \end{aligned}$$

since $\int_{\overline{Y}} f(x,y)dy = f_1(x)$. The implication $E[E[X \mid Y]] = E[X]$ derived above is known as the *law of iterated expectations*. Its message is intuitive. Suppose that we start from the expected value of a random variable X, computed as conditional on another random variable Y. Suppose that we then average such a conditional expected value over all possible realizations y of the random variable Y. Then, naturally we end up with the unconditional expected value of the random variable X.

■ **Example 3.4** [☕☕☕] Let us go back to Example 3.2, where we considered the pair (X,Y) of continuous random variables with joint density

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

where $n > 2$ is an integer. In Example 3.2, we derived the conditional expected value

$$E[Y \mid X = x] = \frac{n-1+x}{n}.$$

Then, the law of iterated expectations tells us that, by computing the expected valued of the conditional expected value $E[Y \mid X = x]$ over $X = x$, we should obtain the unconditional expectation $E[Y]$. Let us then verify this implication for this particular random pair.

On the one hand, note that the marginal density of X is given by

$$\begin{aligned} f_2(y) &= \int_{\overline{X}} f(x,y)dx = n(n-1)\int_{x=0}^{x=y} (y-x)^{n-2}dx \\ &= n(n-1)\left[-\frac{(y-x)^{n-1}}{n-1}\right]_{x=0}^{x=y} = ny^{n-1} \end{aligned}$$

for $0 \leq y \leq 1$. Therefore,

$$E[Y] = \int_{y=0}^{y=1} nyy^{n-1}dy = \frac{n}{n+1}\left[y^{n+1}\right]_{y=0}^{y=1} = \frac{n}{n+1}.$$

On the other hand, recall that we derived in Example 3.2 the marginal density $f_1(x) = n(1-x)^{n-1}$ of X. Then, we can compute

$$E[E[Y \mid X = x]] = \int_{x=0}^{x=1} (n-1+x)(1-x)^{n-1}dx.$$

To compute the integral above, note that

$$(n-1+x)(1-x)^{n-1} = n(1-x)^{n-1} - (1-x)^n.$$

Therefore,

$$\begin{aligned} E[E[Y \mid X=x]] &= n\int_{x=0}^{x=1}(1-x)^{n-1}dx - \int_{x=0}^{x=1}(1-x)^{n}dx \\ &= \Big[-(1-x)^n\Big]_{x=0}^{x=1} + \frac{1}{n+1}\Big[(1-x)^{n+1}\Big]_{x=0}^{x=1} \\ &= 1-\frac{1}{n+1} = \frac{n}{n+1}. \end{aligned}$$

We thus observe that $E[E[Y \mid X=x]] = E[Y] = n/(n+1)$. ■

More generally, the expected value is just one out of a (countable) family of indicators 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

$$E[g(X)] = \int_{\overline{X}} g(x)f(x)dx = \int_{g(\overline{X})} yh(y)dy.$$

To obtain the expression of our family of quantities of interest, let us consider $g(X) = X^r$, where r is any positive integer.

Definition 3.2 The *moment of order* $r \in \{1,2,\dots\}$ of the random variable X with density f is the number

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

Observation 3.3 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 $\int_{\overline{X}} |x|^r f(x)dx < \infty$ implies that $\int_{\overline{X}} |x|^k f(x)dx < \infty$ for each $k = 1,\dots,r$. Then, since $|x|^r f(x) \geq x^r f(x)$, we obtain that

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

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*.

Definition 3.3 The *variance* of a random variable X with density f is the number

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

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.

Observation 3.4 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

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

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 (the *skewness of the distribution*) and the moment of order 4 can be used to measure the weight to the tail of the distribution (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 4.2 in Subsection chapter 4 will provide us with an interesting characterization result that can be viewed as a formal statement of such an intuition.

3.1 Covariance and Correlation (☕)

Subsection section 2.6 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

$$\mathrm{Cov}[X,Y] = \int_{\overline{X}}\int_{\overline{Y}}(x-\mu_X)(y-\mu_Y)f(x,y)dxdy. \tag{3.1}$$

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

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

If $\mathrm{Cov}[X_1,X_2]=0$, 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*.

Observation 3.5 What is the relationship 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 (Eq. (3.1)) above, we obtain

$$\begin{aligned}\mathrm{Cov}[X,Y] &= \int_{\overline{X}}\int_{\overline{Y}}(x-\mu_X)(y-\mu_Y)f(x,y)dxdy\\ &= \int_{\overline{X}}\int_{\overline{Y}}xyf(x,y)dxdy+\mu_X\mu_Y\\ &\quad -\mu_X\int_{\overline{Y}}yf_2(y)\,dy-\mu_Y\int_{\overline{X}}xf_1(x)\,dx\\ &= E[X\cdot Y]-\mu_X\mu_Y.\end{aligned} \tag{3.2}$$

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 (Eq. (3.2)) above, it necessarily follows that $\mathrm{Cov}[X,Y]=0$. Therefore, independence of two random variables implies that they are uncorrelated. However, two uncorrelated random variables need not be independent in general.

On the one hand, positive values of $\mathrm{Cov}[X,Y]$ indicate that, according to their odds of occurrence, Y tends to increase as X does. On the other hand, negative values of $\mathrm{Cov}[X,Y]$ indicate that Y tends to decrease when X increases. Higher values of $\mathrm{Cov}[X,Y]$ in absolute terms reflect higher degrees of relation in the odds of occurrence of the events described by X and Y.

Observation 3.6 To gain intuition about the idea behind the correlation of two random variables, consider the special case where X and Y are indicator functions, respectively, of whether any of two events A and B occurs. Thus, $X = 1$ 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 (Eq. (3.2)) above to this discrete case, we see that

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

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

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

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 3.1 — Schwarz's Inequality. [☕] Let W and Z be two random variables, then

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

Proof of Theorem 3.1. 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

$$
\begin{aligned}
0 \leq & E\left[\left(W - \frac{E[W\cdot Z]}{E[Z^2]}\cdot Z\right)^2\right] = E\left[W^2 + \frac{\left(E[W\cdot Z]\right)^2}{(E[Z^2])^2}\cdot Z^2 - 2\frac{E[W\cdot Z]}{E[Z^2]}\cdot W\cdot Z\right] = \\
& = E[W^2] - \frac{\left(E[W\cdot Z]\right)^2}{E[Z^2]} \quad \Rightarrow \left(E[W\cdot Z]\right)^2 \leq E[W^2]E[Z^2],
\end{aligned}
$$

as stated. ■

Observation 3.7 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 3.1 above to the variables W, Z. We then obtain

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

3.2 Practice Exercises

Exercise 3.1 Let X be a 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.

Exercise 3.2 Let (X,Y) be a 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

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

Exercise 3.3 Let X be a random variable on some probability space $(\Omega, \mathscr{F}, P)$ which takes only the values 0, 1, 2, ... Show that $E[X] = \sum_{n=1}^{\infty} P(X \geq n)$.

Exercise 3.4 Let X be a continuous random with $\overline{X} = [0,b]$, where $b > 0$, with distribution function F, and with density function f. Show that

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

Exercise 3.5 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 $E[X|Y = y]$.

Exercise 3.6 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

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

Exercise 3.7 Let X and Y be two random variables (on some probability space $(\Omega, \mathscr{F}, P)$) such that $P(|X - Y| \leq a) = 1$ for some constant $a \in \mathbb{R}$. Show that $|E[X] - E[Y]| \leq a$.

Exercise 3.8 Show that $\mathrm{Var}[aX] = a^2 \mathrm{Var}[X]$ for any random variable X and constant $a \in \mathbb{R}$.

Exercise 3.9 Let X and Y be two continuous random variables with joint density

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

where $\rho > 0$. Find the conditional density $f_{Y|x}(y)$.

Exercise 3.10 Let X and Y be two 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 $X = x$, $E[Y \mid X = x]$.
Hint: Use the Gaussian integral identity: $\int_{-\infty}^{+\infty} e^{-z^2} dz = \sqrt{\pi}$.

Exercise 3.11 Let X and Y be two continuous random variables with joint density

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

Find the conditional expected value of X given $Y = y$, $E[X \mid Y = y]$.

4. Moment Generating Functions (☕☕)

In chapter 3, some arguments hinted 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). Furthermore, each of these two function completely characterizes its distribution. In consequence, working with such any of these transformations is equivalent to having access to the distribution itself. The main caveat of these functions is that they lack a intuitive interpretation. Nonetheless, they turn 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*.

Definition 4.1 The *moment generating function* of a random variable X with density f is a function $\phi_X : \mathbb{R} \to \mathbb{R}$ specified as

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

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.

Definition 4.2 The *characteristic function* of a random variable X with density f is a function $\varphi_X : \mathbb{R} \to \mathbb{R}$ specified as

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

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.

Observation 4.1 To see how the moment generating function can be used to easily compute the moments of the corresponding distribution, let us invoke the Taylor 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,

$$m_r(X) = \phi_X^{(r)}(t)\Big|_{t=0}. \tag{4.1}$$

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 (Eq. (4.1)), 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

$$\phi_X(t_1, \ldots, t_n) = \underset{\overline{X}}{\int \cdots \int} e^{(t_1 x_1 + \cdots + t_n x_n)} f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n$$

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

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

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

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

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 4.1 [☕] 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 4.1** [☕] 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 chapter 5, 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 4.2 — 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 4.2 above can be exploited to find a particular probability distribution.

■ **Example 4.2** 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

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

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

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

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

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

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

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

As we will see later, the above expression corresponds to the density of a distribution known as Chi-Square with parameter 1. In fact, this distribution already appeared in Example 2.9. ■

4.1 Practice Exercises

Exercise 4.1 Let (X_1, X_2) be a random vector. Using the concept of moment generating function, show that

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

Exercise 4.2 Let X be a random variable which takes only the values 0, 1, 2, ... Show that, for $t \in (-1,1)$, $\Phi_X(t) = E[t^x]$, $\Phi_X'(t) = E[xt^{x-1}]$, and $\Phi_X''(t) = E[x(x-1)t^{x-2}]$.

5. Some Important Distributions (☕)

5.1 Some Discrete Distributions (☕)

We begin with some discrete distributions that appear when we consider a class of experiments known as Bernoulli trials. 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*.

5.1.1 The binomial distribution

Let X be the random variable associated with the number of successes in the n outcomes of a sequence of Bernoulli 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 2.3. 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

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

It follows that

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

and

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

Observation 5.1 Consider 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

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

We can compute compute

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

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

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

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 obtain

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

and

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

Some insights of the observation above are summarized in the next two results.

Theorem 5.1 [☕] 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 5.1 Let $X_i \sim b(n, p)$, $i = 1, \ldots, k$, be independent random variables. Then,

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

This logic behind this result was already used in Example 2.3.

5.1.2 The negative binomial distribution

Consider now a sequence (maybe infinite) of Bernoulli trials and let X be the random variable that describes 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 5.1** [☕☕] 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 5.2** [☕☕☕] Consider two independent geometric random variables $X,Y \sim G(p)$ and suppose that we wish to compute the probability $P(X = m|X+Y = n)$ for $m \in \{1,2,\ldots,n-1\}$. First note that, by the definition of conditional probability, we have

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

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:

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

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

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

■

5.1.3 The multinomial distribution

The binomial distribution can be generalized 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*.

5.1.4 The Poisson distribution

Let us first consider the Taylor expansion of the function $h(r) = e^r$. In particular, 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 a real number $r > 0$, consider the function $f : \mathbb{R} \to \mathbb{R}_+$ defined by

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

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 $X \sim \mathscr{P}(r)$. 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 $X \sim \mathscr{P}(r)$, we have

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

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 5.2 [☕] Let $X_i \sim \mathscr{P}(r_i)$, $i = 1, \ldots, k$, be independent random variables. Then,

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

The following results relate the Poisson with the binomial distribution.

Theorem 5.3 [☕☕] Let $X \sim \mathscr{P}(r_x)$ and $Y \sim \mathscr{P}(r_y)$ 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 $X \sim \mathscr{P}(\theta p/(1-p))$ and $Y \sim \mathscr{P}(\theta)$ for an arbitrary $\theta > 0$.

Theorem 5.4 [☕] If $X \sim \mathscr{P}(r)$ and $(Y|X=x) \sim b(x,p)$, then $Y \sim \mathscr{P}(rp)$.

5.2 Some Continuous Distributions (☕)

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

5.2.1 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

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

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

■ **Example 5.3** Let X be a random variable with density

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

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

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

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

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

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 5.5 [☕] 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 $x \in \mathbb{R}$.

5.2.2 The Γ, χ^2, and Beta distributions

It is a well known that the integral

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

yields a finite positive number for $\alpha > 0$. Then, 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 a 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 gamma distribution is often used to model waiting times. 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,

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

Then, the corresponding distribution function can be computed as

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

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

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

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 a 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 $\mathrm{Var}[X] = 2r$.

Theorem 5.6 [☕] Let $X_i \sim \Gamma(\alpha_i, \beta)$, $i = 1, \dots, k$, be independent random variables. Then, $Y_k = \sum_{i=1}^{k} X_i \sim \Gamma(\sum_{i=1}^{k} \alpha_i, \beta)$.

Theorem 5.7 [☕] Let $X \sim U[0,1]$. Then, $Y = -2\ln X \sim \chi^2(2)$.

Theorem 5.8 [☕] 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 5.9 [☕] Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of independent random variables such that $X_n \sim exp(\beta)$ for each $n = 1, 2, \dots$. Let $Y_n = \sum_{i=1}^{n} X_i$ for $n = 1, 2, \dots$ and let Z be the random variable corresponding to the number of $Y_n \in [0,t]$ for $t > 0$. Then $Z \sim P(t/\beta)$.

The Beta distribution is 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

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

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

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

The marginal distribution of X is then

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

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 5.8 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

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

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 5.10 [☕] Let $U \sim \Gamma(\alpha, \gamma)$ and $V \sim \Gamma(\beta, \gamma)$ be two independent random variables. Then $X = U/(U+V) \sim B(\alpha, \beta)$.

5.2.3 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 5.1 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 $\sigma^2 = 1$.

Observation 5.2 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.

The following result is an implication of the above observation.

Theorem 5.11 [☕] 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.

Observation 5.3 Using the result in Theorem 5.11, 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

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

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

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

While many applications require us to work with normal distributions, normal density functions usually contain a factor of the type e^{-s^2}. In consequence, antiderivatives $\int e^{-s^2}ds$ cannot be obtained in closed form and, thus, we need to resort to numerical integration techniques instead. 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 the distribution function $H(z)$ of a standard normal $N(0,1)$. Such computations are provided by tables for the normal $N(0,1)$.

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

Theorem 5.12 [☕] Let X be a standard normal random variable. Then,

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

Theorem 5.13 [☕] If X and Y are independent normally distributed random variables, then $X+Y$ and $X-Y$ are independent.

Theorem 5.14 [☕] 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 5.15 [☕] If $X \sim N(\mu, \sigma^2)$, then $(X-\mu)^2/\sigma^2 \sim \chi^2(1)$.

The proof of the result above has already been illustrated 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 5.2 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.

Observation 5.4 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

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

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

Theorem 5.16 [☕] 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.

Observation 5.5 Let us consider the bivariate normal. From the theorem above it follows 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!

Unlike the implication in the observation above, 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 though only if these two random variables are not distributed jointly as bivariate normal.

■ **Example 5.4** [☕] 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

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

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

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

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 5.17 [☕] 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

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

However, if two random variables X and Y have both normal distribution, then this 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 5.18 [☕] 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 informs us about the relation between a multivariate normal distribution and its conditional distributions.

Theorem 5.19 [☕] 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

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

Observation 5.6 For the bivariate normal case, we use the expression given in the theorem above for the joint density of (X,Y) to obtain—upon 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:

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

5.2.4 The t and the F distributions

Definition 5.3 A random variable X is said to have the t *distribution* if its density function is given by

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

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 5.20 [☕] 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 5.21 [☕] 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) $(\overline{X}_n - \mu)\big/(S_n/\sqrt{n}) \sim t(n-1)$.

Definition 5.4 A random variable X is said to have the *F distribution* if its density function is given by

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

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 5.22 [☕] Let $X \sim \chi^2(\alpha)$ and $Y \sim \chi^2(\beta)$ be independent random variables. Then

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

5.3 Practice Exercises

Exercise 5.1 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.

Exercise 5.2 Let X be a random variable with density function $f(x) = (1/3)(2/3)^x$, $x = 0, 1, 2, \ldots$. Find the conditional density of X given that $X \geq 3$.

Exercise 5.3 Let X be a random variable with geometric distribution. Show that

$$P(X > k + j | X > k) = P(X > j).$$

Exercise 5.4 Let X be a random variable with moment generating function

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

Compute $P(X \leq 4)$.

Exercise 5.5 Let $X \sim \mathscr{P}(1)$. Compute, if it exists, the expected value $E[X!]$.

Exercise 5.6 Prove Theorem 5.6.

Exercise 5.7 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 $Y = \min\{X_1, X_2, X_3\}$.

Exercise 5.8 Let $X \sim U[0,1]$. Find the density function of $Y = -\ln X$.

Exercise 5.9 Prove Theorem 5.13.

Exercise 5.10 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 $P(X_1 > X_2 + X_3 + 2)$.

Exercise 5.11 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

$$E\left[|X|^{2n+1}\right] = 2^n n! \sqrt{\frac{2}{\pi}}.$$

Exercise 5.12 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?

Exercise 5.13 Let $X_i \sim N(0,1)$, $i = 1, \dots, 4$, be independent random variables. Show that $Y = X_1X_2 + X_3X_4$ has the density function $f(y) = (1/2)\exp\{-|y|\}$ for each $y \in \mathbb{R}$.

Exercise 5.14 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

$$h(x,y) = f(x)g(y)\left[1 + \alpha(2F(x) - 1)(2G(y) - 1)\right],$$

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.

Exercise 5.15 Give a closed expression for $E[X^r]$, $r = 1, 2, \dots$, where $X \sim F(\alpha, \beta)$.

Exercise 5.16 Let $X \sim \chi^2(n)$ and $Y \sim \chi^2(m)$ be independent random variables. Find the density of $Z = X\big/(X+Y)$.

Exercise 5.17 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 $\mathrm{Var}[X] = \mathrm{Var}[Y]$.

Exercise 5.18 Let $X \sim N(2,4)$. Compute $P(1 < X < 6)$ using only the function $\gamma(y) = 1\big/\sqrt{2\pi}\int_0^y e^{-s^2/2} ds$.

Exercise 5.19 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 $\mathrm{Cov}[X,Y]$ too.
(b) Find the conditional density of $Y|X = x$, $E[Y|X = x]$, and $\mathrm{Var}[Y|X = x]$.
(c) Find $P(4 \leq Y \leq 6 | X = 4)$.

Exercise 5.20 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 $x \in \mathbb{R}$.

II

Statistical Inference

6 Convergence of Distributions (☕) 109

6.1 Convergence in Distribution (☕) 109

6.2 Convergence in Probability and Almost Surely (☕) 111

6.3 Laws of Large Numbers (☕) 113

6.4 The Central Limit Theorem (☕) 115

6.5 Practice Exercises 116

7 Parametric Point Estimation (☕) 119

7.1 Consistent Estimation (☕) 120

7.2 Sufficient Estimation (☕) 120

7.3 Unbiased Estimation (☕) 122

7.4 Maximum Likelihood Estimation (☕) 123

7.5 Rao-Cramér Bound and Efficient Estimation (☕) 127

7.6 Practice Exercises 128

8 Hypothesis Testing 129

8.1 Neyman-Pearson Theory (☕) 129

8.2 Tests Based on the Likelihood Ratio (☕) . . 136

8.3 Practice Exercises 139

6. Convergence of Distributions (☕)

In this chapter we study convergence properties of sequences of random variables.

6.1 Convergence in Distribution (☕)

Convergence in distribution yields the weakest notion of convergence.

Definition 6.1 Given some probability space, let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables and let X be a random variable. Let $\{F_n\}_{n=1}^{\infty}$ and F be the corresponding sequence of distribution functions and the distribution function. We say that X_n *converges in distribution* to X or, equivalently, F_n *converges in law (or weakly)* to F if

$$\lim_{n\to\infty} F_n(x) = F(x)$$

for each point x at which F is continuous. We write $X_n \xrightarrow{L} X$ and $F_n \xrightarrow{w} F$.

From here onwards, let us use "i.i.d." as abbreviation for "independent and identically distributed."

■ **Example 6.1** [☕] Let $X_1, X_2, \ldots, X_n$ i.i.d. random variables with (common) density function

$$f(x) = \begin{cases} 1/\theta & \text{if } 0 \le x < \theta; \\ 0 & \text{otherwise,} \end{cases}$$

where $0 < \theta < \infty$. Let $Y_n = \max\{X_1, X_2, \ldots, X_n\}$ for $n = 1, 2, \ldots$. Then, the distribution function of Y_n is given by

$$\begin{aligned} G_n(y) &= P(Y_n \leq y) = P(Y_1 \leq y, \ldots, Y_n \leq y) = [F(y)]^n \\ &= \begin{cases} 0 & \text{if } y < 0; \\ (y/\theta)^n & \text{if } 0 \leq y < \theta; \\ 1 & \text{if } y \geq \theta. \end{cases} \end{aligned}$$

Then, given $y \geq 0$,

$$\lim_{n\to\infty} G_n(y) = G(y) = \begin{cases} 0 & \text{if } y < \theta; \\ 1 & \text{if } y \geq \theta. \end{cases}$$

Therefore, $Y_n \xrightarrow{L} Y$, where Y is the random variable associated to a random experiment that yields θ with certainty. ■

The following example shows that convergence in distribution does not imply convergence of the moments of the distribution. This illustrates that convergence in distribution gives us a weak form of convergence.

■ **Example 6.2** [☕] Let $\{F_n\}_{n=1}^{\infty}$ be a sequence of distribution functions defined by

$$F_n(x) = \begin{cases} 0 & \text{if } x < 0; \\ 1 - 1/n & \text{if } 0 \leq x < n; \\ 1 & \text{if } x \geq n. \end{cases}$$

Note that, for each $n = 1, 2, \ldots$, F_n is the distribution function of a discrete random variable X_n, supported on the set $\{0, n\}$, with density function

$$P(X_n = 0) = 1 - \frac{1}{n}, \quad P(X_n = n) = \frac{1}{n}.$$

We have, for each given $x \geq 0$,

$$\lim_{n\to\infty} F_n(x) = \begin{cases} 0 & \text{if } x < 0; \\ 1 & \text{if } x \geq 0. \end{cases}$$

Note that F is is the distribution function of a random variable X degenerate at $x = 0$ so that, clearly, for $r = 1, 2, \ldots$, one obtains $E[X^r] = 0$. However, we have

$$E[X_n^r] = (0)^r \left(1 - \frac{1}{n}\right) + (n)^r \left(\frac{1}{n}\right) = n^{r-1},$$

so that, evidently, $\lim_{n\to\infty} E[X^r] \neq E[X]$. ■

6.2 Convergence in Probability and Almost Surely (☕)

This section formalizes a pair of notions of converge in which a sequence of random variables approaches another random variable. We will see subtle differences between the two ideas of convergence presented in this section. First, convergence in probability is not related with convergence of random variables in the sense typically understood in real analysis. In particular, convergence in probability tells us something about the convergence of a sequence of probabilities.

Definition 6.2 Given some probability space, let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables and let X be a random variable. We say that X_n *converges in probability* to X if for each $\varepsilon > 0$, we have

$$\lim_{n\to\infty} P\big(|X_n - X| > \varepsilon\big) = 0.$$

We write $X_n \xrightarrow{P} X$.

Note that the condition in the definition above can be equivalently rewritten as

$$\lim_{n\to\infty} P\big(|X_n - X| \leq \varepsilon\big) = 1.$$

■ **Example 6.3** [☕] Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables with (discrete) density function

$$P(X_n = 0) = 1 - \frac{1}{n}, \quad P(X_n = 1) = \frac{1}{n}.$$

Then,

$$P\big(|X_n| > \varepsilon\big) = \begin{cases} 1/n & \text{if } 0 < \varepsilon < 1; \\ 0 & \text{if } \varepsilon \geq 1, \end{cases}$$

so that $\lim_{n\to\infty} P\big(|X_n| > \varepsilon\big) = 0$ and, therefore, $X_n \xrightarrow{P} 0$. ■

■ **Example 6.4** [☕☕] Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of i.i.d. random variables with (common) density function

$$f(x) = \begin{cases} e^{-(x-\theta)} & \text{if } x > \theta; \\ 0 & \text{if } x \leq \theta, \end{cases}$$

where $\theta \in \mathbb{R}$, and let $Y_n = \min\{X_1, \ldots, X_n\}$ for each $n = 1, 2, \ldots$. We can now show that $Y_n \xrightarrow{P} \theta$. To do this, note that, for any given real number $y > \theta$, we have

$$\begin{aligned} F_n(y) &= P(\min\{X_1,\ldots,X_n\} \leq y) = 1 - P(\min\{X_1,\ldots,X_n\} > y) \\ &= 1 - P(X_1 > y, \cdots, X_n > y) = 1 - \left(\int_y^{\infty} e^{-(x-\theta)} dx\right)^n \\ &= 1 - e^{-n(y-\theta)}. \end{aligned}$$

Therefore, for a given $\varepsilon > 0$, we obtain

$$\begin{aligned} P(|Y_n - \theta| \leq \varepsilon) &= P(\theta - \varepsilon \leq Y_n \leq \theta + \varepsilon) \\ &= F_n(\theta + \varepsilon) - F_n(\theta - \varepsilon) \\ &= 1 - e^{-n(\theta+\varepsilon-\theta)}, \end{aligned}$$

where we have taken into account that $F_n(\theta - \varepsilon) = 0$ since $\theta - \varepsilon < \theta$. Finally, we trivially obtain $1 - e^{-n\varepsilon} \to 1$ as $n \to \infty$, as required. ■

Convergence in probability is preserved through continuous transformations and under some common algebraic operations.

Theorem 6.1 [☕] Suppose $X_n \xrightarrow{P} X$ and let $g : \mathbb{R} \to \mathbb{R}$ be a continuous function. Then $g(X_n) \xrightarrow{P} g(X)$.

Theorem 6.2 [☕] Suppose $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$. Then:
(i) $\alpha X_n + \beta Y_n \xrightarrow{P} \alpha X + \beta Y$ for each $\alpha, \beta \in \mathbb{R}$;
(ii) $X_n \cdot Y_n \xrightarrow{P} X \cdot Y$.

Theorem 6.3 [☕] Suppose $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{L} X$.

There exists another notion of converge that strengthens further the idea of converge in probability. Unlike convergence is probability, this notion does have a flavor similar to the pointwise converge of a sequence of function. Almost surely convergence requires that converge occurs in a set that (possibly) only excludes a set that has probability zero of occurrence. This message of something happing except in a set that has probability zero of occurrence is more formally developed in chapter chapter 11.

Definition 6.3 Given some probability space, let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables and let X be a random variable. We say that X_n *converges almost surely* to X if for each $\varepsilon > 0$, we have

$$P\Big(\lim_{n\to\infty} |X_n - X| < \varepsilon\Big) = 1.$$

■ **Example 6.5** [☕] Consider the sample set $\Omega = [0,1]$ and a probability space $(\Omega, \mathscr{B}_\Omega, P)$ where P assigns equal probability of occurrence to each $\omega \in \Omega$. Then, take a random variable X defined as $X(\omega) = \omega$ so that X that distributes uniformly on the support $[0,1]$. In particular, X has density $f(x) = 1$ for each $x \in [0,1]$ and $P(X \in [a,b]) = b - a$ for each interval $[a,b] \subseteq [0,1]$ (with $b > a$). Also, consider a sequence of random variables on $(\Omega, \mathscr{B}_\Omega, P)$ defined as $X_n(\omega) = \omega + \omega^n$. Then, notice $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ that for each $\omega \in [0,1)$ whereas $\lim_{n\to\infty} X_n(1) = 2 \neq X(1) = 1$. Since $P(1) = 0$, we have that X_n converges to X for each $\omega \in \Omega$, except for a set $\{1\}$ of zero probability. Therefore X_n converges almost surely to X. ■

Observation 6.1 The differences between the ideas of convergence in probability and almost surely convergence are subtle but fundamental. To appreciate such differences between the two notions, recall that a random variable X is a (measurable) function on a certain sample set Ω. Therefore, a sequence of random variables X_n converges almost surely to random variable X if the functions $X_n(\omega)$ converge to the function $X(\omega)$ for each $\omega \in \Omega$ except (possibly) for some $\omega \in A$ such that $P(A) = 0$. On the other hand X_n converges almost surely to random variable X is the sequence of probabilities $\{P\big(|X_n - X| \geq \varepsilon\big)\}_{n=1}^{\infty}$ converges to zero. Note, that any of such probabilities $P\big(|X_n - X| \geq \varepsilon\big)$ is a number which does not depend on ω.

6.3 Laws of Large Numbers (☕)

The weak law of large numbers informs us about the converge in probability of the mean of a large number of random variables.

Theorem 6.4 — WLLN. [☕] Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of i.i.d. random variables with $E[X_n] = \mu$ and $\text{Var}[X_n] = \sigma^2 < \infty$, and let $\overline{X}_n = \sum_{i=1}^{n} X_i/n$. Then, $\overline{X}_n \xrightarrow{P} \mu$.

Proof of Theorem 6.4. Using the inequality of Chebychev-Bienaymé, for $\varepsilon > 0$, we have

$$0 \leq P\left(\left|\overline{X}_n - \mu\right| \geq \varepsilon\right) \leq \frac{1}{\varepsilon^2}\text{Var}\left[\overline{X}_n\right] = \frac{\sigma^2}{n\varepsilon^2}.$$

Then, the result follows since $\sigma^2/n\varepsilon^2 \to 0$ as $n \to \infty$. ■

There is also a well known result, called the Strong Law of Large Numbers, where the requirements of the theorem above allows us to achieve almost surely convergence of the sample mean to theoretical mean.

Theorem 6.5 — SLLN. [☕] Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of i.i.d. random variables with $E[X_n] = \mu$ and $\text{Var}[X_n] = \sigma^2 < \infty$, and let $\overline{X}_n = \sum_{i=1}^{n} X_i/n$. Then, the sample mean $\overline{X}_n$ converges almost surely to theoretical mean μ. That is, for each $\varepsilon > 0$, we have

$$P\Big(\lim_{n\to\infty} |\overline{X}_n - \mu| < \varepsilon\Big) = 1.$$

■ **Example 6.6** [☕] Consider a sequence $\{X_n\}_{n=1}^{\infty}$ of i.i.d. random variables with $E[X_n] = \mu$ and $\text{Var}[X_n] = \sigma^2 < \infty$. Let

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

The WLLN states that $\overline{X}_n \xrightarrow{P} \mu$. Let us now ask about convergence results concerning the *sample variance* S_n^2. Assume that $E\left[X_i^4\right] < \infty$ for each $i = 1, 2, \ldots$, so that $\text{Var}[S_n^2] < \infty$ for each $n = 1, 2, \ldots$. We obtain

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{n}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2\right).$$

Using the WLLN, we obtain that $\overline{X}_n \xrightarrow{P} \mu$. Also, by taking $Y_i = X_i^2$, the WLLN tells us that

$$\overline{Y}_n = \sum_{i=1}^{n} X_i^2/n \xrightarrow{P} E[Y_k] = E[X_k^2]$$

for any given $k = 1, 2, \ldots$. Then, combining the results in the Theorems above, we obtain

$$S_n^2 = \frac{n}{n-1} \cdot \left(\frac{1}{n} \sum_{i=1}^{n} X_i^2 - \overline{X}_n^2 \right) \xrightarrow{P} 1 \cdot \left(E[X_k^2] - \mu^2 \right) = \sigma^2.$$

■

6.4 The Central Limit Theorem (☕)

The central limit theorem is a key implication in probability and statistics that convey the striking message that averages of large samples of arbitrary distributions end up behaving as normal. This key implication thus gives us appealing foundations to consider normal distributions as governing random phenomena that concerns large populations. At a practical level, the central limit theorem can be used to approximate the computation of probabilities when we consider the realizations of sequences of random variables. The central limit theorem has a number of (slightly) different formulations. In practices, all that it requires is finite variances of the random variables in the sequence.

We begin by introducing the notion of random sample, which is a key tool in estimation.

Definition 6.4 A *random sample of size n* from a distribution with distribution function F is a set $\{X_1, X_2, \ldots, X_n\}$ of i.i.d. random variables whose (common) distribution function is F.

Using Theorem 5.11, we can verify that if $X_1, X_2, \ldots, X_n$ are i.i.d. normal random variables with mean μ and variance σ^2, then the random variable

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$$

has the standard normal distribution.

Now, suppose that $X_1, X_2, \ldots, X_n$ are the observations (not necessarily independent!) of a random sample of size n obtained from *any* distribution with finite variance $\sigma^2 > 0$ and, therefore, finite mean μ. The important result stated below says that the random variable $\sqrt{n}(\overline{X}_n - \mu)/\sigma$ converges in distribution to a random variable distributed according to the standard normal. It will be then possible to use this approximation to the normal distribution to compute approximate probabilities concerning $\overline{X}_n$. In the statistical problem where μ is unknown, we shall use this approximation of $\overline{X}_n$ to estimate μ.

Theorem 6.6 — Lindeberg-Lévy Central Limit Theorem. [☕] Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables with $E[X_n] = \mu$ and $0 < \text{Var}[X_n] = \sigma^2 < \infty$, and let $\overline{X}_n = \sum_{i=1}^{n} X_i / n$. Then, the sequence of random variables $\{Y_n\}_{n=1}^{\infty}$ defined by

$$Y_n = \frac{(\sum_{i=1}^{n} X_i - n\mu)}{\sqrt{n}\sigma} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$$

satisfies $Y_n \xrightarrow{L} Y \sim N(0,1)$.

■ **Example 6.7** [☕] Consider a set $\{X_1, \ldots, X_{75}\}$ of random variables with $X_i \sim U[0,1]$ for each $i = 1, \ldots, 75$. We are interested in computing $P(0.45 < \overline{X}_n < 0.55)$, where $\overline{X}_n = \sum_{i=1}^{75} X_i / 75$. Such computation maybe complicated to obtain directly. However, using the theorem above together with the fact that $\mu = 1/2$ and $\sigma^2 = 1/12$, one obtains

$$\begin{aligned} P(0.45 < \overline{X}_n < 0.55) &= P\left(\frac{\sqrt{75}(0.45 - 0.5)}{1/\sqrt{12}} < \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} < \frac{\sqrt{75}(0.55 - 0.5)}{1/\sqrt{12}}\right) \\ &= P(-1.5 < 30(\overline{X}_n - 0.5) < 1.5) \approx 0.866, \end{aligned}$$

since $30(\overline{X}_n - 0.5)$ is approximately distributed according to the standard normal distribution. ■

6.5 Practice Exercises

Exercise 6.1 Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables with $X_i \sim b(n,p)$, for each $i = 1, \ldots, n$ $(0 < p < 1)$. Obtain the probability distribution of a random variable X such that $X_n \xrightarrow{L} X$.

Exercise 6.2 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*:

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

Exercise 6.3 Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables with mean $\mu < \infty$ and variance a/n^p, where $a \in \mathbb{R}$ and $p > 0$. Show that $X_n \xrightarrow{P} \mu$.

Exercise 6.4 Let $\overline{X}$ be the mean of a random sample of size 128 from a Gamma distribution with $\alpha = 2$ and $\beta = 4$. Approximate $P(7 < \overline{X} < 9)$.

Exercise 6.5 Let $f(x) = 1/x^2$ for $1 < x < \infty$ and $f(x) = 0$ for $x \leq 1$. Consider a random sample of size 72 from the probability distribution of a random variable X which has f as density function. Compute approximately the probability that more than 50 observations of the random variable are less than 3.

Exercise 6.6 Let $\overline{X}_1$ and $\overline{X}_2$ be the means of two independent random samples of size n from a population with variance σ^2. Find the value of n such that $P\left(\left|\overline{X}_1 - \overline{X}_2\right| < \sigma/5\right) \approx 0.99$ and justify your answer.

Exercise 6.7 Let $X_n \xrightarrow{L} X$ and let Y_n be a sequence of random variables with the property that, for any finite number c,

$$\lim_{n\to\infty} P(Y_n > c) = 1.$$

Show that, for any finite number c, we have

$$\lim_{n\to\infty} P(X_n + Y_n > c) = 1.$$

7. Parametric Point Estimation (☕)

Sometimes we are interested in working with a random variable X but we do not know its distribution function F. The distribution function F describes the behavior of a phenomenon or population (whose individuals are, accordingly, the realizations of the random variable X). Basically, this not knowing a distribution function can take two forms. Either we ignore completely the form of $F(x)$ or we do know the functional form of F but ignore a set of parameters upon which F depends. The problem of point estimation is of the second type. For instance, we may know that a certain population has a normal distribution $N(\mu,\sigma^2)$ but ignore one of the parameters, say σ^2. Then, after drawing a random sample $\{X_1,X_2,\ldots,X_n\}$ from the distribution $N(\mu,\sigma^2)$, the problem of point estimation consists of choosing a number $T(X_1,X_2,\ldots,X_n)$ that depends only on the sample and best estimates the unknown parameter σ^2. If both parameters μ and σ^2 are unknown, then we need to seek for a pair

$$T(X_1,X_2,\ldots,X_n) = \big(T_1(X_1,X_2,\ldots,X_n),T_2(X_1,X_2,\ldots,X_n)\big) \in \mathbb{R}^2$$

such that T_1 estimates μ and T_2 estimates σ^2.

Let us be specific about the estimation problem. Consider that the random variable X has a distribution function F_θ and a density function f_θ which depend on some unknown parameter $\theta = (\theta_1,\ldots,\theta_k) \in \mathbb{R}^k$. Let Θ denote the subset $\Theta \subseteq \mathbb{R}^k$ of possible values for the parameter and let $\mathscr{X}$ denote the set of possible random samples of size n. Thus, we are indeed considering a family $\{F_\theta : \theta \in \Theta\}$ of distribution functions parameterized by θ. A *point estimator (or statistic)* for θ is any function $T : \mathscr{X} \to \Theta$.

Next, we introduce certain desirable properties of estimators. The criteria that we discuss are consistency, sufficiency, unbiasedness, and efficiency.

7.1 Consistent Estimation (☕)

Consider a random sample and suppose that we want to choose an estimator T of an unknown parameter θ. Then, the consistency criterion requires that it becomes very likely that the estimator approaches the true value of the parameter as the size of the random sample increases.

Definition 7.1 Let $\{X_1, X_2, \dots, X_n\}$ be a random sample from F_θ. A point estimator $T(X_1, X_2, \dots, X_n)$ is *consistent* for $\theta \in \Theta$ if $T(X_1, X_2, \dots, X_n) \xrightarrow{P} \theta$.

Intuitively, if a estimator T is consistent for a parameter θ, then we may interpret it as "T being close to θ on average" as n increases.

▪ **Example 7.1** [☕] Let $\{X_1, X_2, \dots, X_n\}$ be a random sample from a binomial distribution $b(1, p)$. Then, $E[X_k] = p$ for each $k = 1, 2, \dots, n$. From the WLLN, we know that

$$\overline{X}_n = T(X_1, X_2, \dots, X_n) = \sum_{i=1}^{n} X_i/n \xrightarrow{P} p$$

so that the *sample mean* is consistent to estimate p. Now it can be easily checked that

$$\frac{\sum_{i=1}^{n} X_i + 1}{n+2} = \frac{\sum_{i=1}^{n} X_i}{n} \cdot \frac{n}{n+2} + \frac{1}{n+2} \xrightarrow{P} p.$$

Thus, a consistent estimator for a certain parameter need not be unique. Finally, as shown earlier,

$$S_n^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2\right) \xrightarrow{P} \operatorname{Var}[X_k]$$

for each $k = 1, 2, \dots, n$, so that the sample variance is consistent to estimate the variance of the distribution. It can be easily checked that S_n^2 is not the unique consistent estimator for the variance of the population. ▪

7.2 Sufficient Estimation (☕)

The desiderata associated to the sufficiency criterion can be summarized as requiring that the only information obtained about the unknown parameter is that provided by the sample itself. Thus, we want to rule out possible relations between the proposed estimator and the parameter. Under this criterion, we seek for estimators that make "full use" of the information contained in the sample.

Definition 7.2 Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from F_θ. A point estimator $T(X_1, X_2, \ldots, X_n)$ is *sufficient* for $\theta \in \Theta$ if the conditional density of $(X_1, X_2, \ldots, X_n)$, given $T(X_1, X_2, \ldots, X_n) = t$, does not depend on θ (except perhaps for a set A of zero measure, $P_\theta[X \in A] = 0$).

■ **Example 7.2** [☕] Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a binomial distribution $b(1, p)$ and consider the estimator $T(X_1, X_2, \ldots, X_n) = \sum_{i=1}^n X_i$. Then, by considering $t = \sum_{i=1}^n x_i$, we obtain

$$
\begin{aligned}
f_\theta(x_1, \ldots, x_n | T = t) &= \frac{P\Big(X_1 = x_1, \ldots, X_n = x_n, \sum_{i=1}^n X_i = \sum_{i=1}^n x_i\Big)}{P\Big(\sum_{i=1}^n X_i = t\Big)} \\
&= \frac{p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}}{\binom{n}{t} p^t (1-p)^{n-t}} = \frac{1}{\binom{n}{t}},
\end{aligned}
$$

which does not depend on p. So, $\sum_{i=1}^n X_i$ is sufficient to estimate p. ■

Often times, it turns out to be difficult to use the definition of sufficiency to check whether a estimator is sufficient or not. The following result is then helpful in many applications.

Theorem 7.1 — Fisher-Neyman Factorization Criterion. [☕] Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from F_θ and let f_θ denote the joint density function of $(X_1, X_2, \ldots, X_n)$. Then, a estimator $T(X_1, X_2, \ldots, X_n)$ is sufficient for a parameter θ if and only if $f_\theta(x_1, \ldots, x_n)$ can be factorized as follows:

$$
f_\theta(x_1, \ldots, x_n) = h(x_1, \ldots, x_n) \cdot g_\theta(T(x_1, \ldots, x_n)),
$$

where h is a nonnegative function of $x_1, \ldots, x_n$ only and does not depend on θ, and g_θ is a nonnegative nonconstant function of θ and $T(x_1, \ldots, x_n)$ only.

■ **Example 7.3** [☕] As in the previous example, let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a binomial distribution $b(1, p)$ and consider the estimator $T(X_1, X_2, \ldots, X_n) = \sum_{i=1}^n X_i$. Then, we can write

$$
f_p(x_1, \ldots, x_n) = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} = 1 \cdot (1-p)^n \left(\frac{p}{1-p}\right)^{\sum_{i=1}^n x_i}.
$$

Therefore, by taking $h(x_1,\dots,x_n)=1$ and $g_p(\sum_{i=1}^n x_i)=(1-p)^n\left[p/(1-p)\right]^{\sum_{i=1}^n x_i}$, we obtain that $\sum_{i=1}^n X_i$ is sufficient to estimate p. ■

■ **Example 7.4** [☕] Let $\{X_1,X_2,\dots,X_n\}$ be a random sample from a normal distribution $N(\mu,\sigma^2)$ and suppose that we are interested in estimating both μ and σ^2. Then, we can write

$$\begin{aligned} f_{(\mu,\sigma^2)}(x_1,\dots,x_n) &= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n}\exp\left\{\frac{-\sum_{i=1}^n(x_i-\mu)^2}{2\sigma^2}\right\} \\ &= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n}\exp\left\{\frac{\mu\sum_{i=1}^n x_i}{\sigma^2}-\frac{\sum_{i=1}^n x_i^2}{2\sigma^2}-\frac{n\mu^2}{2\sigma^2}\right\}. \end{aligned}$$

Then, using the factorization theorem above, it follows that

$$T(X_1,X_2,\dots,X_n)=\left(\sum_{i=1}^n X_i,\sum_{i=1}^n X_i^2\right)$$

is a sufficient estimator for (μ,σ^2). ■

7.3 Unbiased Estimation (☕)

Another desirable criterion for choosing point estimators is that the expected value of the estimator gives us the true parameter.

Definition 7.3 Let $\{X_1,X_2,\dots,X_n\}$ be a random sample from F_θ. A point estimator $T(X_1,X_2,\dots,X_n)$ is *unbiased* for $\theta\in\Theta$ if

$$E_\theta\left[T(X_1,X_2,\dots,X_n)\right]=\theta.$$

At this point, we can derive a useful result about the moment of second order of any distribution. Let X be a random variable such that $E[X]=\mu$ and $\text{Var}[X]=\sigma^2$. Using the general definition of expected value given in chapter 3, we obtain

$$E[X^2]=\int_\Omega X^2dP=\int_\Omega\left[(X-\mu)^2-\mu^2+2\mu X\right]dP=\sigma^2+\mu^2.$$

With this result in hand, we can now show that the sample mean and the sample variance are unbiased estimators for the population mean and the population variance, respectively. Consider a random variable X with $E[X]=\mu$, $\text{Var}[X]=\sigma^2$, and distribution function $F_{(\mu,\sigma^2)}$.

Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from $F_{(\mu,\sigma^2)}$. First, we easily obtain

$$E\left[\overline{X}_n\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \mu.$$

Secondly, we have

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}_n^2\right],$$

so that

$$E[S_n^2] = \frac{1}{n-1}\left[\sum_{i=1}^{n} E[X_i^2] - nE[\overline{X}_n^2]\right].$$

Then, since $E[Z^2] = \big(E[Z]\big)^2 + \mathrm{Var}[Z]$ for any random variable Z, we obtain

$$\begin{aligned} E[S_n^2] &= \frac{1}{n-1}\left[n\mu^2 + n\sigma^2 - n\left[\left(E[\overline{X}_n]\right)^2 + \mathrm{Var}[\overline{X}_n]\right]\right] \\ &= \frac{1}{n-1}\left[n\mu^2 + n\sigma^2 - n\left[\mu^2 + \frac{\sigma^2}{n}\right]\right] \\ &= \sigma^2. \end{aligned}$$

Hence, the sample variance is unbiased to estimate the population variance. On the other hand, notice that the estimator $\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2/n$ is biased to estimate σ^2.

7.4 Maximum Likelihood Estimation (☕)

In the previous sections we have considered several desirable properties that can be used to search for appropriate estimators. Now, we introduce another method which has a more constructive approach. The basic tool of this method is the *likelihood function* of a random sample, which is nothing but its joint density function. To follow the usual notation, given a random sample $\{X_1, X_2, \ldots, X_n\}$ from a distribution F_θ, we rename its joint density function as

$$L(x_1, \ldots, x_n; \theta) \equiv f_\theta(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i).$$

Furthermore, for tractability reasons, in many applications it is convenient to work with the log transformation of the likelihood function:

$$\Pi(\theta) = \ln L(x_1, \ldots, x_n; \theta) = \sum_{i=1}^{n} \ln f_\theta(x_i).$$

At this point, we need to make some assumptions on our working benchmark.

Assumption 7.1 — Regularity Conditions. 1. For $\theta, \theta' \in \Theta$, we have

$$\theta \neq \theta' \Rightarrow f_\theta \neq f_{\theta'};$$

2. the support of f_θ does not depend on θ for each $\theta \in \Theta$.

Now, suppose that the actual value of the unknown parameter θ is θ_0. The following result gives us theoretical reasons for being interested in obtaining the maximum of the function $\Pi(\theta)$. It tells us that the maximum of $\Pi(\theta)$ asymptotically separates the true model at θ_0 from any other model $\theta \neq \theta_0$.

Theorem 7.2 [☕] Given Assumption 7.1, if θ_0 is the true value of the unknown parameter θ, then

$$\lim_{n\to\infty} P_{\theta_0}\left[L(X_1, \ldots, X_n; \theta_0) \geq L(X_1, \ldots, X_n; \theta)\right] = 1 \quad \text{for each} \quad \theta \in \Theta.$$

Proof of Theorem 7.2. By taking logs, the inequality $L(X_1, \ldots, X_n; \theta_0) \geq L(X_1, \ldots, X_n; \theta)$ can be rewritten as

$$\sum_{i=1}^{n} \ln f_\theta(X_i) \leq \sum_{i=1}^{n} \ln f_{\theta_0}(X_i) \Leftrightarrow Y_n = \frac{1}{n}\sum_{i=1}^{n} \ln\left(\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}\right) \leq 0.$$

Then, from the WLLN it follows that

$$\frac{1}{n}\sum_{i=1}^{n} \ln\left(\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}\right) \xrightarrow{P} E_{\theta_0}\left[\ln\left(\frac{f_\theta(X_1)}{f_{\theta_0}(X_1)}\right)\right].$$

Now, using the fact that $\ln(s)$ is a strictly concave function in s, we can use Jensen's inequality to obtain

$$E_{\theta_0}\left[\ln\left(\frac{f_\theta(X_1)}{f_{\theta_0}(X_1)}\right)\right] < \ln\left(E_{\theta_0}\left[\frac{f_\theta(X_1)}{f_{\theta_0}(X_1)}\right]\right).$$

However, notice that

$$E_{\theta_0}\left[\frac{f_\theta(X_1)}{f_{\theta_0}(X_1)}\right] = \int_{-\infty}^{+\infty} \frac{f_\theta(x_1)}{f_{\theta_0}(x_1)} dF_{\theta_0}(x_1) = \int_{-\infty}^{+\infty} \frac{f_\theta(x_1)}{f_{\theta_0}(x_1)} f_{\theta_0}(x_1)dx_1 = 1.$$

Since $\ln(1) = 0$, we have obtained that

$$Y_n = \frac{1}{n}\sum_{i=1}^{n} \ln\left(\frac{f_\theta(X_i)}{f_{\theta_0}(X_i)}\right) \xrightarrow{P} Z < 0.$$

Therefore, for any $\varepsilon > 0$, from the definition of convergence in probability, we know that

$$\lim_{n\to\infty} P_{\theta_0}\left[Z - \varepsilon \leq Y_n \leq Z + \varepsilon\right] = 1.$$

Since $Z < 0$, by choosing $\varepsilon > 0$ small enough so as to have $Z + \varepsilon = 0$, the equality above implies (considering only one of the inequalities within the probability operator)

$$\lim_{n\to\infty} P_{\theta_0}\left[Y_n \leq 0\right] = 1,$$

as desired. ■

In short, the likelihood function is asymptotically maximized at the true value θ_0.

Definition 7.4 Let $\{X_1, X_2, \ldots, X_n\}$ be random sample from F_θ and let $(x_1, x_2, \ldots, x_n)$ be a realization of that sample. The value $T(x_1, x_2, \ldots, x_n) = \widehat{\theta}$ is a *maximum likelihood estimate* for θ if

$$\Pi(\widehat{\theta}) \geq \Pi(\theta') \quad \text{for each} \quad \theta' \in \Theta.$$

■ **Example 7.5** [☕] Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a binomial distribution $b(1, p)$. Then,

$$f_p(x_1, \ldots, x_n) = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

and, consequently,

$$\Pi(p) = \left(\sum_{i=1}^{n} x_i\right)\ln p + \left(n - \sum_{i=1}^{n} x_i\right)\ln(1-p).$$

Then,

$$\Pi'(p) = 0 \;\Rightarrow\; (1-p)\sum_{i=1}^{n} x_i = p\left(n - \sum_{i=1}^{n} x_i\right) \;\Rightarrow\; \widehat{p} = \sum_{i=1}^{n} x_i/n.$$

Thus, the sample mean is the maximum likelihood estimator of p. ■

■ **Example 7.6** [☕] Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a uniform distribution $U[0,\theta]$. Since the parameter θ is in the support of the distribution, differentiation is not helpful here. Notice instead that the corresponding likelihood function can be written as

$$L(x_1,\ldots,x_n;\theta) = \left(\frac{1}{\theta^n}\right)\phi(\max\{x_i : i = 1,\ldots,n\},\theta),$$

where $\phi(a,b) = 1$ if $a \leq b$ and $\phi(a,b) = 0$ if $a > b$. So, $L(x_1,\ldots,x_n;\theta)$ is decreasing in θ for $\theta \geq \max\{x_i : i = 1,\ldots,n\}$ and equals zero for $\theta < \max\{x_i : i = 1,\ldots,n\}$. Furthermore, notice that, despite being decreasing in θ for $\theta \geq \max\{x_i : i = 1,\ldots,n\}$, its maximum is attained at $\widehat{\theta} = \max\{x_i : i = 1,\ldots,n\}$ since $L(x_1,\ldots,x_n;\widehat{\theta}) = 0$ for $\widehat{\theta} < \max\{x_i : i = 1,\ldots,n\}$. ■

■ **Example 7.7** [☕] Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a normal distribution $N(0,\sigma^2)$. The likelihood function is obtained as

$$L(x_1,\ldots,x_n;\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left\{\frac{-\sum_{i=1}^{n} x_i^2}{2\sigma^2}\right\}$$

so that

$$\Pi(\sigma^2) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{\sum_{i=1}^{n} x_i^2}{2\sigma^2}.$$

Therefore,

$$\Pi'(\sigma^2) = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n} x_i^2}{2\sigma^4} = 0 \;\Rightarrow\; \widehat{\sigma}^2 = \frac{\sum_{i=1}^{n} x_i^2}{n}.$$

■

7.5 Rao-Cramér Bound and Efficient Estimation (☕)

This section presents an important inequality which establishes a lower bound for the variance of any unbiased estimator. First, we need to restrict further our benchmark by imposing a few requirements additional to those given by Assumption 7.1.

Assumption 7.2 — Additional Regularity Conditions.

1. The point θ_0 is an interior point in Θ;
2. f_θ is twice differentiable with respect to θ;
3. the integral $\int f_\theta(x_i)dx_i$ can be differentiated twice (under the integral sign) with respect to θ.

Theorem 7.3 — Rao-Cramér Lower Bound. [☕] Under Assumptions 7.1 and 7.2, if $\{X_1, X_2, \ldots, X_n\}$ is a random sample from F_θ and $T(X_1, X_2, \ldots, X_n)$ is a point estimator for θ with mean $E\left[T(X_1, X_2, \ldots, X_n)\right] = \tau(\theta)$, then

$$\operatorname{Var}\left[T(X_1, X_2, \ldots, X_n)\right] \geq \frac{\left[\tau'(\theta)\right]^2}{nI(\theta)},$$

where

$$nI(\theta) = E_\theta\left[\frac{\partial \ln f_\theta(x_1, \ldots, x_n)}{\partial \theta}\right]^2$$

is a quantity called Fisher information of the random sample.

Note that if $T(X_1, X_2, \ldots, X_n)$ is an unbiased estimator of θ, then the Rao-Cramér inequality becomes

$$\operatorname{Var}\left[T(X_1, X_2, \ldots, X_n)\right] \geq \frac{1}{nI(\theta)}$$

The Rao-Cramér lower bound gives us another criterion for choosing appropriate estimators.

Definition 7.5 Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from F_θ. A point estimator $T(X_1, X_2, \ldots, X_n)$ is *efficient* for $\theta \in \Theta$ if its variance attains the Rao-Cramér lower bound.

7.6 Practice Exercises

Exercise 7.1 Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from F_θ and let $T(X_1, X_2, \ldots, X_n)$ be a point estimator of θ. Show that if T is unbiased for θ and $\lim_{n\to\infty} \operatorname{Var}[T] = 0$, then T is consistent for θ.

Exercise 7.2 Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a distribution with density function

$$f_\theta(x) = \theta x^{\theta-1}, \quad \text{for} \quad 0 < x < 1,$$

where $\theta > 0$. Argue whether the product $X_1 X_2 \cdots X_n$ is a sufficient estimator for θ or not.

Exercise 7.3 Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a Poisson distribution with mean r. Propose a maximum likelihood estimator for r.

Exercise 7.4 Let X and Y be two random variables such that $E[Y] = \mu$ and $\operatorname{Var}[Y] = \sigma^2$. Let $T(x) = E[Y|X = x]$. Show that $E[T(X)] = \mu$ and $\operatorname{Var}[T(X)] \le \sigma^2$.

Exercise 7.5 What is a sufficient estimator for θ if the random sample is drawn from a beta distribution with $\alpha = \beta = \theta > 0$?

Exercise 7.6 Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a distribution with density function

$$f_\theta(x) = \frac{e^{-(x-\theta)}}{\left[1 + e^{-(x-\theta)}\right]^2}, \quad \text{for} \quad -\infty < x < +\infty,$$

where $\theta \in \mathbb{R}$. Show that there exists a unique maximum likelihood estimator for θ.

Exercise 7.7 Let X_1 and X_2 constitute a random sample from a Poisson distribution with mean r. Show that $X_1 + X_2$ is a sufficient estimator for r and that $X_1 + 2X_2$ is not a sufficient estimator for r.

8. Hypothesis Testing

8.1 Neyman-Pearson Theory (☕)

In the previous chapter we analyzed the problem of using sample information to estimate unknown parameters of a probability distribution. In this chapter we follow a slightly different approach. We use sample information to test hypotheses about the unknown parameters. The treatment of this problem is as follows. We have a distribution function F_θ that depends on some unknown parameter (or vector of parameters) θ and our objective is to use a random sample $\{X_1, X_2, \ldots, X_n\}$ from this distribution to test hypotheses about the value of θ. As in the previous chapter, we assume that the functional form of F_θ, except for the parameter θ itself, is known. Suppose that we think, from preliminary information, that $\theta \in \Theta_0$ where $\Theta_0 \subset \Theta$. This assertion is usually known as the *null hypothesis*, $H_0 : \theta \in \Theta_0$, while the statement $H_1 : \theta \in \Theta_1 = \Theta \setminus \Theta_0$ is known as the *alternative hypothesis*. We write

$$H_0 : \theta \in \Theta_0;$$
$$H_1 : \theta \in \Theta_1.$$

There are two types of hypotheses: if Θ_0 (Θ_1) contains only one point, the hypothesis is *simple*, otherwise the hypothesis is *composite*. Note that if a hypothesis is simple, then the distribution function F_θ becomes completely specified under that hypothesis. For example, consider a random variable $X \sim N(\mu, \sigma^2)$. Then, we might propose the test

$$\begin{aligned} H_0 &: \mu \leq -1, \;\; \sigma^2 > 2; \\ H_1 &: \mu > -1, \;\; \sigma^2 \leq 2, \end{aligned}$$

where both the null and the alternative hypotheses are composite. Here, under any of those hypotheses, the distribution of X remains not fully specified.

The procedure that we follow to test hypotheses is as follows. Given the sample space $\mathscr{X}$, we search for a decision rule that allows us, for each realization $(x_1, \ldots, x_n)$ of the random sample, to either "accept" (roughly speaking) or reject the null hypothesis. More specifically, for $\Theta \subseteq \mathbb{R}^k$, we consider a statistic $T : \mathscr{X} \to \Theta$ and partition the sample space of that statistic into two sets $C \subset \mathbb{R}^k$ and $C^c = \mathbb{R}^k \setminus C$. Now, if $T(x_1, \ldots, x_n) \in C$, then we reject H_0 while if $T(x_1, \ldots, x_n) \in C^c$, then we fail to reject H_0. When $T(x_1, \ldots, x_n) \in C^c$ and, consequently, we fail to reject H_0, then we shall write from here onwards "accept" H_0. However, we emphasize that this does not necessarily mean that H_0 can be granted our stamp of approval. It rather means that the sample does not provide us with sufficient evidence against H_0.

Alternatively, we can partition the space of the random sample itself (instead of the set of possible values taken by the statistic) into $A \subset \mathbb{R}^n$ and $A^c = \mathbb{R}^n \setminus A$. Then, we can use the same reasoning as before, that is, if $(x_1, \ldots, x_n) \in A$, then we reject H_0 and "accept" it otherwise.

The set C (resp., A) such that if $T(x_1, \ldots, x_n) \in C$ (resp., $(x_1, \ldots, x_n) \in A$), then H_0 is rejected (with probability 1) is called the *critical region* of the test. There are four possibilities that can arise when one uses this procedure:

1. H_0 is accepted when it is correct;
2. H_0 is rejected when it is correct;
3. H_0 is accepted when it is incorrect (and, thus, H_1 is correct);
4. H_0 is rejected when it is incorrect (and, thus, H_1 is correct).

Possibilities 2. and 3. above are known, respectively, as *type I* and *type II* errors.

We can now present the basic theory underlying hypothesis testing.

Definition 8.1 A Borel-measurable function $\varphi : \mathbb{R}^n \to [0, 1]$ is a *test function*. Further, a test function φ is a *test* of hypothesis $H_0 : \theta \in \Theta_0$ against the alternative $H_1 : \theta \in \Theta_1$, with *error probability* (or *significance level*) α, if

$$E_\theta \left[\varphi(X_1, \ldots, X_n)\right] \leq \alpha \qquad \text{for each} \quad \theta \in \Theta_0.$$

The function (as a function of θ) $E_\theta \left[\varphi(X_1, \ldots, X_n)\right]$ is known as the *power function of the test* φ and the least upper bound $\sup_{\theta \in \Theta_0} E_\theta \left[\varphi(X_1, \ldots, X_n)\right]$ is known as the *size of the test* φ.

The interpretation of the concepts above is as follows. A test φ allows us to assign to each sample realization $(x_1,\ldots,x_n) \in \mathbb{R}^n$ a number $\varphi(x_1,\ldots,x_n) \in [0,1]$, which is to be interpreted as the probability of rejecting H_0. Thus, the inequality $E_\theta\left[\varphi(X_1,\ldots,X_n)\right] \leq \alpha$ for $\theta \in \Theta_0$ says that if H_0 were true, then the test φ rejects it with probability

$$\begin{aligned} E_\theta\left[\varphi(X_1,\ldots,X_n)\right] &= P\left(\text{reject } H_0 \mid H_0 \text{ is true}\right) \\ &= P\left(T(X_1,\ldots,X_n) \in C \mid H_0\right) = P\left((X_1,\ldots,X_n) \in A \mid H_0\right) \leq \alpha. \end{aligned}$$

In other words, the definition of test requires that the probability of the type I error exceeds not the amount α.

There is an intuitive class of tests, used often in applications, called *nonrandomized tests*, such that $\varphi(x_1,\ldots,x_n) = 1$ if $(x_1,\ldots,x_n) \in A$ and $\varphi(x_1,\ldots,x_n) = 0$ if $(x_1,\ldots,x_n) \notin A$ for some set $A \subset \mathbb{R}^n$ (i.e., φ is the indicator function I_A for a subset A of sample realizations). In the sequel, we will make use of this class of tests.

Given an error probability equal to α, let us use $(\alpha, \Theta_0, \Theta_1)$ as short notation for our hypothesis testing problem. Also, let Φ_α be the set of all tests for the problem $(\alpha, \Theta_0, \Theta_1)$.

Definition 8.2 Given a random sample $\{X_1, X_2, \ldots, X_n\}$ from F_θ. A test $\widehat{\varphi} \in \Phi_\alpha$ is a *most powerful test* against an alternative $\theta' \in \Theta_1$ if

$$E_{\theta'}\left[\widehat{\varphi}(X_1,\ldots,X_n)\right] \geq E_{\theta'}\left[\varphi(X_1,\ldots,X_n)\right] \qquad \text{for each} \quad \varphi \in \Phi_\alpha.$$

If a test $\widehat{\varphi} \in \Phi_\alpha$ is a most powerful test (uniformly) against each alternative $\theta' \in \Theta_1$, then $\widehat{\varphi}$ is a *uniformly most powerful test.*

To gain intuition on this, suppose that both hypotheses are simple so that $(\{\theta_0\}, \{\theta_1\}, \alpha)$ is our hypotheses testing problem. Then, note first that

$$\begin{aligned} E_{\theta_1}\left[\varphi(X_1,\ldots,X_n)\right] &= P\left(\text{reject } H_0 \mid H_1 \text{ is true}\right) \\ &= P\left(T(X_1,\ldots,X_n) \in C \mid H_1\right) = P\left((X_1,\ldots,X_n) \in A \mid H_1\right) \\ &= 1 - P\left(\text{accept } H_0 \mid H_1 \text{ is true}\right). \end{aligned}$$

Note that the expected value $E_{\theta_1}\left[\varphi(X_1,\ldots,X_n)\right]$ is the power of the test evaluated at the alternative hypothesis. Then, when we seek for a most powerful test, we are indeed trying to solve the problem

$$\min_{A\subset\mathbb{R}^n} P_{\theta_1}[(X_1,\dots,X_n)\in A^c]$$
$$\text{s.t.:}\quad P_{\theta_0}[(X_1,\dots,X_n)\in A]\leq\alpha.$$

In other words, the method of a most powerful test lead us to minimize the probability of type II error subject to the restriction that the probability of type I error exceeds not α, as imposed by the definition of the test. This method then gives us the practical procedure to follow in choosing the critical region for testing a hypothesis: choose the critical region (and, therefore, the test) in such a way that, for a given size α (or probability of type I error), the power of the test is maximized (or, equivalently, the probability of type II error is minimized).

Note that, for a general hypotheses testing problem $(\Theta_0,\Theta_1,\alpha)$, finding a uniformly most powerful test is equivalent to proposing a critical region $A\subset\mathbb{R}^n$ that, for each $\theta_1\in\Theta_1$, minimizes the probability $P_{\theta_1}[(X_1,\dots,X_n)\in A^c]$ under the restriction

$$\sup_{\theta_0\in\Theta_0} P_{\theta_0}[(X_1,\dots,X_n)\in A]\leq\alpha.$$

■ **Example 8.1** [☕☕] Let $\{X_1,X_2,\dots,X_n\}$ be a random sample from a normal distribution $N(\mu,1)$. We know that $\mu\in\Theta=\{\mu_0,\mu_1\}$, $\mu_0<\mu_1$. Consider the test

$$H_0:\mu=\mu_0;$$
$$H_1:\mu=\mu_1,$$

so that both H_0 and H_1 are simple hypotheses. We choose the sample mean $\overline{X}_n$ as statistic so that, intuitively, one would accept H_0 if $\overline{X}_n$ is "closer" to μ_0 than to μ_1. That is, one would reject H_0 if $\overline{X}_n>c$, for some constant c, and would otherwise accept H_0. Then, for $0<\alpha<1$, we have

$$\begin{aligned}\alpha&=P\left(\text{reject }H_0\mid H_0\text{ is true}\right)=P(\overline{X}_n>c\,|\,\mu=\mu_0)\\&=P\left(\frac{\overline{X}_n-\mu_0}{1/\sqrt{n}}>\frac{c-\mu_0}{1/\sqrt{n}}\right)=1-F_Z\left(\frac{c-\mu_0}{1/\sqrt{n}}\right),\end{aligned}$$

where $Z\sim N(0,1)$. Therefore, the value c must solve the equation

$$F_Z\left(\frac{c-\mu_0}{1/\sqrt{n}}\right)=1-\alpha,$$

so that one obtains

$$c = \mu_0 + \frac{z_{(1-\alpha)}}{\sqrt{n}},$$

where $z_{(1-\alpha)}$ denotes the realization z of the random variable Z that such that $P(Z \leq z) = 1-\alpha$, i.e., the *quantile of order* $(1-\alpha)$ of the distribution of Z. Therefore, the corresponding nonrandomized test φ is specified as

$$\varphi(x_1,\ldots,x_n) = \begin{cases} 1 & \text{if } \sum_{i=1}^n x_i/n > \mu_0 + z_{(1-\alpha)}/\sqrt{n}; \\ 0 & \text{otherwise.} \end{cases}$$

Finally, the power of the test at μ_1 is

$$\begin{aligned} E[\varphi(x_1,\ldots,x_n)\,|\,\mu=\mu_1] &= P\left(\overline{X}_n > \mu_0 + \frac{z_{(1-\alpha)}}{\sqrt{n}} \,\middle|\, \mu=\mu_1\right) \\ &= P\left(\frac{\overline{X}_n - \mu_1}{1/\sqrt{n}} > (\mu_0-\mu_1)\sqrt{n} + z_{(1-\alpha)}\right) \\ &= 1 - F_Z\Big(z_{(1-\alpha)} - (\mu_1-\mu_0)\sqrt{n}\Big). \end{aligned}$$

■

The result below, due to Neyman and Pearson, gives us a general method for finding a most powerful test of a simple hypothesis against a simple alternative. Following the notation used in the previous chapter, let $L(x_1,\ldots,x_n;\overline{\theta})$ denote the likelihood function of the random sample $\{X_1,\ldots,X_n\}$ given that the true value of the parameter θ is $\overline{\theta}$.

Theorem 8.1 — Neyman-Pearson Fundamental Lemma. Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a distribution function F_θ. Let θ_0 and θ_1 be two distinct values of θ and let k be a positive number. Consider the following test of two simple hypotheses:

$$H_0 : \theta = \theta_0;$$
$$H_1 : \theta = \theta_1.$$

Let A and A^c be a subset of the set of sample realizations and its complement, respectively, such that

$$\begin{aligned}
&\frac{L(x_1,\ldots,x_n;\theta_0)}{L(x_1,\ldots,x_n;\theta_1)} \le k, \qquad \text{for each} \quad (x_1,\ldots,x_n) \in A,\\
&\frac{L(x_1,\ldots,x_n;\theta_0)}{L(x_1,\ldots,x_n;\theta_1)} \ge k, \qquad \text{for each} \quad (x_1,\ldots,x_n) \in A^c,\\
&\alpha = \underset{A}{\int \cdots \int} L(x_1,\ldots,x_n;\theta_0)\, dx_1 \cdots dx_n.
\end{aligned}$$

Then, A is a critical region for a most powerful test $\widehat{\varphi}$ against the alternative θ_1.

The most powerful test $\widehat{\varphi}$ identified in the theorem above must be necessarily specified as:

$$\widehat{\varphi} = \begin{cases} 1 & \text{if} \;\; f_{\theta_1}(x_1,\ldots,x_n) > q f_{\theta_0}(x_1,\ldots,x_n);\\ \gamma(x_1,\ldots,x_n) & \text{if} \;\; f_{\theta_1}(x_1,\ldots,x_n) = q f_{\theta_0}(x_1,\ldots,x_n);\\ 0 & \text{if} \;\; f_{\theta_1}(x_1,\ldots,x_n) < q f_{\theta_0}(x_1,\ldots,x_n), \end{cases}$$

for some $q \ge 0$ and $0 \le \gamma(x_1,\ldots,x_n) \le 1$. When $q \to \infty$, $\widehat{\varphi}$ is specified as:

$$\widehat{\varphi} = \begin{cases} 1 & \text{if} \;\; f_{\theta_0}(x_1,\ldots,x_n) = 0;\\ 0 & \text{if} \;\; f_{\theta_0}(x_1,\ldots,x_n) > 0. \end{cases}$$

Finally, it can be shown that there is a functional form for $\gamma(x_1,\ldots,x_n)$ such that γ indeed does not depend on $(x_1,\ldots,x_n)$ and the resulting $\widehat{\varphi}$ is as identified by the Neyman-Pearson Lemma.

■ **Example 8.2** [☕☕] As in a previous example, consider a random sample $\{X_1, X_2, \ldots, X_n\}$ from a normal distribution $N(\mu, 1)$. We know that $\mu \in \Theta = \{\mu_0, \mu_1\}$, $\mu_0 < \mu_1$. Consider the test

$$\begin{aligned}
&H_0 : \mu = \mu_0;\\
&H_1 : \mu = \mu_1,
\end{aligned}$$

so that both H_0 and H_1 are simple hypotheses. Then,

$$L(x_1,\ldots,x_n;\mu_s) = (2\pi)^{-n/2} \exp\left\{ -\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu_s)^2 \right\}, \qquad s = 0, 1,$$

so that

$$\frac{L(x_1,\dots,x_n;\mu_0)}{L(x_1,\dots,x_n\mu_1)} = \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu_0)^2+\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu_1)^2\right\} \leq k,$$

for some positive number k that depends on α. Taking logs in the expression above, we obtain

$$-\sum_{i=1}^{n}(x_i-\mu_0)^2+\sum_{i=1}^{n}(x_i-\mu_1)^2 \leq 2\ln(k).$$

From the equation above, using the fact that, for $s=0,1$,

$$\sum_{i=1}^{n}(x_i-\mu_s)^2 = \sum_{i=1}^{n}(x_i-\bar{x}_n)^2+n\big(\bar{x}_n-\mu_s\big)^2+2(\bar{x}_n-\mu_s)\sum_{i=1}^{n}(x_i-\bar{x}_n),$$

where $\sum_{i=1}^{n}(x_i-\bar{x}_n) = n\bar{x}_n - n\bar{x}_n = 0$, we get to

$$n\left[\big(\bar{x}_n-\mu_1\big)^2-\big(\bar{x}_n-\mu_0\big)^2\right] \leq 2\ln(k).$$

Then, by computing the squares of the terms in brackets and by rearranging terms, we obtain

$$\bar{x}_n(\mu_0-\mu_1) \leq \frac{1}{2}(\mu_0^2-\mu_1^2)+\frac{1}{n}\ln(k).$$

Therefore, the critical region identified by the Neyman-Pearson Lemma is

$$\bar{x}_n \geq \frac{1}{2}(\mu_0+\mu_1)-\frac{\ln(k)}{n(\mu_1-\mu_0)}.$$

Note that the statistic selected is the sample mean. Finally we set

$$\frac{1}{2}(\mu_0+\mu_1)-\frac{\ln(k)}{n(\mu_1-\mu_0)} =: c,$$

where c is nothing but the constant proposed in the previous example. We can then proceed as in that example to obtain

$$c = \mu_0+\frac{z_{(1-\alpha)}}{\sqrt{n}}. = \frac{1}{2}(\mu_0+\mu_1)+\frac{\ln c}{n(\mu_1-\mu_0)}.$$

■

We end this section by discussing briefly the application of the Neyman-Pearson approach to testing a simple hypothesis against a composite alternative. Using the Neyman-Pearson Lemma, one can conclude that a test is a most powerful test for a simple hypothesis against a single value of the parameter as alternative. To follow this approach for a set of alternatives which is not a singleton, we should check for the Neyman-Pearson criterion for each value of the parameter within the set of alternatives. Thus, we would be searching for a uniformly most powerful test. Unfortunately, it is typical that the a uniformly most powerful test does not exist for *all* values of the parameter. In such cases, we must seek for tests that are most powerful within a restricted class of tests. One such restricted class is, for instance, the class of unbiased tests.

8.2 Tests Based on the Likelihood Ratio (☕)

We present here a classical method for testing a simple or composite hypothesis against a simple or composite alternative. This method is based on the ratio of the sample likelihood function given the null hypothesis over the likelihood function given either the alternative or the entire parameter space. This method gives us a test which is based on a sufficient statistic, if one exists. Also, this procedure often (but not necessarily) leads to a most powerful test or a uniformly most powerful test, if they exist.

Definition 8.3 Given a hypothesis testing problem $(\alpha, \Theta_0, \Theta_1)$, the critical region

$$A = \{(x_1, \ldots, x_n) \in \mathbb{R}^n : \lambda(x_1, \ldots, x_n) < k\},$$

where $k \in \mathbb{R}$ is a constant and

$$\lambda(x_1, \ldots, x_n) = \frac{\sup_{\theta \in \Theta_0} L(x_1, \ldots, x_n; \theta)}{\sup_{\theta \in \Theta} L(x_1, \ldots, x_n; \theta)},$$

corresponds to a test called a *generalized likelihood ratio test*.

In addition, it can be shown that the critical region specified above gives us the same test as the region specified using the statistic

$$\rho(x_1, \ldots, x_n) = \frac{\sup_{\theta \in \Theta_1} L(x_1, \ldots, x_n; \theta)}{\sup_{\theta \in \Theta_0} L(x_1, \ldots, x_n; \theta)}.$$

The idea behind this method is as follows. The numerator in the ratio λ is the best *explanation* of $(X_1, \ldots, X_n)$ under H_0 while the denominator is the best possible *explanation* of $(X_1, \ldots, X_n)$. Therefore, this test proposes that H_0 be rejected if there is a much better explanation of $(X_1, \ldots, X_n)$ than the one provided by H_0.

For practical purposes, $0 \leq \lambda \leq 1$ and the value of the constant k is determined using the restriction of the size of the test

$$\sup_{\theta \in \Theta_0} P(\lambda(x_1, \ldots, x_n) < k) = \alpha,$$

where, accordingly, α is the significance level of the test.

Theorem 8.2 [☕] For a hypothesis testing problem $(\alpha, \Theta_0, \Theta_1)$, the likelihood ratio test is a function of each sufficient statistic for the parameter θ.

▪ **Example 8.3** [☕☕] Let $\{X\}$ be a random sample, consisting of a single random variable, from a binomial distribution $b(n,p)$. We seek a significance level α for the test

$$H_0 : p \leq p_0;$$
$$H_1 : p > p_0,$$

for some $0 < p_0 < 1$. Then, if we propose the monotone likelihood ratio test, we have

$$\lambda(x) = \frac{\sup_{p \leq p_0} \binom{n}{x} p^x (1-p)^{n-x}}{\sup_{0 \leq p \leq 1} \binom{n}{x} p^x (1-p)^{n-x}} = \frac{\max_{p \leq p_0} p^x (1-p)^{n-x}}{\max_{0 \leq p \leq 1} p^x (1-p)^{n-x}}.$$

Now, it can be checked that the function $p^x(1-p)^{n-x}$ first increases until it achieves its maximum at $p = x/n$ and from then on it decreases. Therefore,

$$\max_{0 \leq p \leq 1} p^x (1-p)^{n-x} = \left(\frac{x}{n}\right)^x \left(1 - \frac{x}{n}\right)^{n-x}$$

and

$$\max_{p \leq p_0} p^x (1-p)^{n-x} = \begin{cases} p_0^x (1-p_0)^{n-x} & \text{if } p_0 < x/n \\ \left(\frac{x}{n}\right)^x \left(1 - \frac{x}{n}\right)^{n-x} & \text{if } p_0 \geq x/n. \end{cases}$$

Consequently,

$$\lambda(x) = \begin{cases} \frac{p_0^x (1-p_0)^{n-x}}{(x/n)^x [1-(x/n)]^{n-x}} & \text{if } x > np_0 \\ 1 & \text{if } x \leq np_0. \end{cases}$$

It follows that $\lambda(x) \leq 1$ for $x > np_0$ and $\lambda(x) = 1$ for $x \leq np_0$, so that λ is a function not increasing in x. Therefore, $\lambda(x) < k$ if and only if $x > k'$ and we should reject $H_0 : p \leq p_0$ when $x > k'$.

Since X is a discrete random variable, it may be not possible to obtain the size α. We have

$$\alpha = \sup_{p \leq p_0} P_p[X > k'] = P_{p_0}[X > k'].$$

If such k' does not exist, then we should choose an integer k' such that

$$P_{p_0}[X > k'] \leq \alpha \quad \text{and} \quad P_{p_0}[X > k' - 1] > \alpha.$$

■

■ **Example 8.4** [☕☕] Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a normal distribution $N(\mu, \sigma^2)$ and consider the hypothesis testing problem

$$\begin{aligned} &H_0 : \mu = \mu_0; \\ &H_1 : \mu \neq \mu_0, \end{aligned}$$

where σ^2 is also unknown. Here we have $\theta = (\mu, \sigma^2)$,

$$\Theta = \left\{ (\mu, \sigma^2) \in \mathbb{R}^2 : -\infty < \mu < \infty, \sigma^2 > 0 \right\},$$

and

$$\Theta_0 = \left\{ (\mu_0, \sigma^2) \in \mathbb{R}^2 : \sigma^2 > 0 \right\}.$$

We obtain

$$\sup_{\theta \in \Theta_0} L(x_1, \ldots, x_n; \theta) = \frac{1}{(\widehat{\sigma}_0 \sqrt{2\pi})^n} \exp\left[-\frac{\sum_{i=1}^n (x_i - \mu_0)^2}{2\widehat{\sigma}_0^2} \right] =: \Phi_{\widehat{\sigma}_0^2}(x_1, \ldots, x_n),$$

where $\widehat{\sigma}_0^2 = (1/n) \sum_{i=1}^n (x_i - \mu_0)^2$ is nothing but the maximum likelihood estimator for σ^2 given that the mean of the distribution is μ_0. It follows that

$$\sup_{\theta \in \Theta_0} L(x_1, \ldots, x_n; \theta) = \frac{1}{(2\pi/n)^{n/2} \left[\sum_{i=1}^n (x_i - \mu_0)^2 \right]^{n/2}} e^{-n/2}.$$

Now, it can be checked that the maximum likelihood estimator for (μ, σ^2) when both μ and σ^2

are unknown is

$$(\widehat{\mu},\widehat{\sigma}^2)=\left(\overline{X}_n,\sum_{i=1}^{n}(X_i-\overline{X}_n)^2/n\right).$$

Then, we obtain

$$\sup_{\theta\in\Theta} L(x_1,\ldots,x_n;\theta)=\frac{1}{(2\pi/n)^{n/2}\left[\sum_{i=1}^n(x_i-\bar{x}_n)^2\right]^{n/2}}e^{-n/2}.$$

Therefore,

$$\begin{aligned}\lambda(x_1,\ldots,x_n)&=\left[\frac{\sum_{i=1}^n(x_i-\bar{x}_n)^2}{\sum_{i=1}^n(x_i-\mu_0)^2}\right]^{n/2}=\left[\frac{\sum_{i=1}^n(x_i-\bar{x}_n)^2}{\sum_{i=1}^n(x_i-\bar{x}_n)^2+n(\bar{x}_n-\mu_0)^2}\right]^{n/2}\\&=\left[\frac{1}{1+\left[n(\bar{x}_n-\mu_0)^2/\sum_{i=1}^n(x_i-\bar{x}_n)^2\right]}\right]^{n/2},\end{aligned}$$

which happens to be a decreasing function in $(\bar{x}_n-\mu_0)^2/\sum_{i=1}^n(x_i-\bar{x}_n)^2$. Thus,

$$\lambda(x_1,\ldots,x_n)<k\Leftrightarrow\left|\frac{(\bar{x}_n-\mu_0)/\sqrt{n-1}}{\sqrt{\sum_{i=1}^n(x_i-\bar{x}_n)^2/(n-1)}}\right|>k'\Leftrightarrow\left|\frac{\sqrt{n}(\bar{x}_n-\mu_0)}{s_n}\right|>k'',$$

where $s_n^2=[1/(n-1)]\sum_{i=1}^n(x_i-\bar{x}_n)^2$ is the sample variance and $k''=k'\sqrt{(n-1)/n}$. Also, we know that the statistic

$$T(X_1,\ldots,X_n)=\frac{\sqrt{n}(\overline{X}_n-\mu_0)}{S_n}$$

has a distribution t with $n-1$ degrees of freedom (recall Theorem 5.21 (iv)). So, given the symmetry of the function density of a random variable distributed according to a t, we should make use of the quantile $t_{n-1,\alpha/2}$ to specify k''. ■

8.3 Practice Exercises

Exercise 8.1 Let $\{X_1,\ldots,X_n\}$ be a random sample from a normal distribution $N(\mu,1)$. Use the result in the Neyman-Pearson Lemma to test the null hypothesis $H_0;\ \mu=0$ against the alternative $H_1;\ \mu=1$. For $n=25$ and $\alpha=0.05$, compute the power of this test when the alternative is true.

Exercise 8.2 Let $\{X_1, \ldots, X_n\}$ and $\{Y_1, \ldots, Y_m\}$ be independent random samples from normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively. Use a monotone likelihood ratio test to test the hypothesis $H_0;\ \sigma_1^2 = \sigma_2^2$ against $H_1;\ \sigma_1^2 \neq \sigma_2^2$.

III

Miscellanea on Probability

9 Combinatorics (☕☕) 143

9.1 Ordered Samples and Permutations 143

9.2 Combinations 145

9.3 Practice Exercises 150

10 Integration (☕☕) 153

11 Extension of Probability Measures (Continued) (☕☕☕) 155

9. Combinatorics (☕☕)

This chapter presents 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 $P(A) = |A| \,/\, |\Omega|$.

9.1 Ordered Samples and Permutations

Let us begin with a finite set $S = \{1, 2, \dots, s\}$ 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 $\omega = (\omega_1, \omega_2, \dots, \omega_m)$, 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

$$P(A) = \frac{P_m^s}{s^m} = \left(1 - \frac{1}{s}\right)\left(1 - \frac{2}{s}\right)\cdots\left(1 - \frac{m-1}{s}\right). \tag{9.1}$$

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

■ **Example 9.1** [☕] 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 (Eq. (9.1)) 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 9.2** [☕] 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 $1/P_3^{48} = 1/(48\cdot 47\cdot 46)$. ■

9.2 Combinations

In some problems, we are interested in computing the number of different subsets of size $m \le s$ 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

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

To illustrate how combinations can be used in computing probabilities, let us consider again the experiment, already analyzed in Section section 1.3 , 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 section 1.3, 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 9.3** [☕] 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 9.4** [☕☕☕] 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

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

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 $\{1,\dots,12\}$. 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 $k = 1,2,\dots,12$, 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 9.5** [☕☕☕] 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

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

so that $|\Omega| = n!$. Then, $P(A) = P(\{\omega\})$ for $\omega = (1,2,\dots,n)$ 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 $G \subset \{1,2,\dots,n\}$ 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 $\{F_G\}_{G\subset\{1,\dots,n\}}$ 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 $G \subset \{1,\dots,n\}$. Since there are $\binom{n}{m}$ of such events, we finally obtain

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

■

■ **Example 9.6** [☕☕☕] 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 $\Omega = \{1,\dots,n\}^n$ 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 $\{1,\dots,n\}$ 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

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

■

Finally, 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 9.7** [☕☕] 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

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

■

9.3 Practice Exercises

Exercise 9.1 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?

Exercise 9.2 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.

Exercise 9.3 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.

Exercise 9.4 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.

Exercise 9.5 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.

Exercise 9.6 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 $1/4$.

Exercise 9.7 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.

Exercise 9.8 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.

Exercise 9.9 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.

Exercise 9.10 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.

Exercise 9.11 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.

Exercise 9.12 Suppose that n balls are distributed at random into r boxes. Compute the probability that the box 1 contains exactly k balls, where $0 \leq k \leq n$.

Exercise 9.13 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.

Exercise 9.14 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.

Exercise 9.15 A set of n objects are permuted among themselves. Show that the probability that k pre-specified objects occupy k pre-specified positions is $(n-k)!/n!$.

Exercise 9.16 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.

10. Integration (☕☕)

This chapter deals briefly with 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 $(\Omega, \mathscr{F}, P)$ and let us begin by taking a simple random variable X on $(\mathscr{F}, P)$ so that $\overline{X} = \{x_1, x_2, \ldots, x_n\}$. Let us denote by $A_i = \{\omega \in \Omega \,|\, X(\omega) = x_i\} \in \mathscr{F}$, $i = 1, \ldots, n$, 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 $(\Omega, \mathscr{F}, P)$. 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 $\int XdP$, $\int X(\omega)dP(\omega)$, 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*.

11. Extension of Probability Measures (Continued) (☕☕☕)

This chapter deals in more detail with the problem of the 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 $\mathscr{A}$ 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 $\mathscr{A}$ and are *only* interested in computing probabilities of occurrence of events $A \in \mathscr{A}$, such a measure Q would be enough for our purposes. However, as an algebra, $\mathscr{A}$ might not contain some relatively more complicated events, such as the unions of arbitrary countable sequences of events in $\mathscr{A}$. 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 gain 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

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

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 $\mathscr{A}$ of such events such that be closed under complements, and *finite* unions and intersections. Such a family $\mathscr{A}$ would be an algebra on $\Omega = \{1,\dots,6\}^{\infty}$ so that one could propose $\sigma(\mathscr{A})$ 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 $\mathscr{A}$ 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 $A \in \sigma(\mathscr{A}) \setminus \mathscr{A}$? A formal procedure to construct an extension of the measure Q to the set $\sigma(\mathscr{A}) \setminus \mathscr{A}$ 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 11.1 — Carathéodory (1918). [☕☕☕] Let $\mathscr{A}$ be an algebra on a nonempty set Ω and let Q be a measure on $\mathscr{A}$. Then there exists a measure P on $\sigma(\mathscr{A})$ such that $P(A) = Q(A)$ for each $A \in \mathscr{A}$. Moreover, if Q is a probability measure, then P is unique.

Thus, Theorem 11.1 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 11.1 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.

Now, we get into the formal details of the procedure and present another formulation of Carathéodory’s Theorem.

Definition 11.1 An *outer measure* δ on an arbitrary nonempty set Ω is set function $\delta : 2^{\Omega} \to \mathbb{R}_+^*$, where $\mathbb{R}_+^* = \mathbb{R}_+ \cup \{+\infty\}$, that verifies

(a) $\delta(\emptyset) = 0$;
(b) (*monotonicity*) for $A, B \in 2^{\Omega}$, $A \subseteq B$ implies $\delta(A) \leq \delta(B)$;
(c) (σ-*subadditivity*) for each sequence $\{A_n\}_{n=1}^{\infty}$ of subsets of Ω, we have $\delta(\cup_{n=1}^{\infty} A_n) \leq \sum_{n=1}^{\infty} \delta(A_n)$.

Definition 11.2 Let P be a measure on a measurable space $(\Omega, \mathscr{F})$. The measure P generates a set function $P^* : 2^{\Omega} \to \mathbb{R}_{+}^{*}$ defined by

$$P^*(A) = \inf \left\{ \sum_{n=1}^{\infty} P(A_n) \ : \ \{A_n\}_{n=1}^{\infty} \subset \mathscr{F} \ \text{ and } \ A \subset \cup_{n=1}^{\infty} A_n \right\}, \tag{CExt}$$

which is called the *Carathédory extension* of P.

Intuitively, the Carathédory extension P^* of a measure P is constructed from P by approximating events *from the outside*. If $\{A_n\}_{n=1}^{\infty}$ forms a good covering of A in the sense that they not overlap one another very much or extend much beyond A, then $\sum_{n=1}^{\infty} P(A_n)$ should be a good outer approximation to the measure assigned to A. Then, this approach allows for the following. Consider a measure P on a measurable space $(\Omega, \mathscr{F})$ and the σ-algebra generated by $\mathscr{F}$, $\mathscr{F}^* = \sigma(\mathscr{F})$. Then, $\mathscr{F}^*$ is a σ-algebra larger than $\mathscr{F}$ (in the sense that $\mathscr{F}^* \supseteq \mathscr{F}$). The formulation of Carathéodory's Theorem stated below asserts that there exists an outer measure P^* on $(\Omega, \mathscr{F}^*)$ such that:
(a) $P^*(A) = P(A)$ for each $A \in \mathscr{F}$;
(b) if Q is another measure on $(\Omega, \mathscr{F}^*)$ such that $Q(A) = P(A)$ for each $A \in \mathscr{F}$, then it must be the case that $Q(A) = P^*(A)$ for each $A \in \mathscr{F}$.

Theorem 11.2 [☕☕] A measure P on a measurable space $(\Omega, \mathscr{F})$ such that $P(\Omega) < \infty$ has a unique extension P^* (i.e., conditions (a) and (b) above are satisfied), defined by Eq. (CExt) above, to the generated σ-algebra $\sigma(\mathscr{F})$. Moreover, the extension P^* is an outer measure of Ω.

The extension P^* of P identified in the theorem above is also known as the *outer measure generated by* P. Given a probability space $(\Omega, \mathscr{F}, P)$, the phrase *"P-almost everywhere"* (which is often substituted by just "almost everywhere" (or "almost surely") when the probability measure P is understood from the context) means "everywhere except possibly for a set $A \in \mathscr{F}$ with $P^*(A) = 0$", where P^* is the outer measure generated by P. For example, we say that two functions $f, g : A \to B$ are P-almost everywhere equal if $P^*(\{a \in A : f(a) \neq f(a)\}) = 0$.

Bibliography

P. Billingsley. *Probability and Measure*. Wiley and Sons, 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.