Probability and Statistics Lecture Notes

Antonio Jiménez-Martínez

Probability spaces

In this chapter we introduce the theoretical structures that will allow us to assign probabilities in a wide range of probability problems.

1.1. Examples of random phenomena

Science attempts to formulate general laws on the basis of observation and experiment. The simplest and most used scheme of such laws is:

$$\text{if a set of conditions } B \text{ is satisfied} \implies \text{event } A \text{ occurs.}$$

Examples of such laws are the law of gravity, the law of conservation of mass, and many other instances in chemistry, physics, biology... . If event A occurs inevitably whenever the set of conditions B is satisfied, we say that A is **certain** or **sure** (under the set of conditions B). If A can never occur whenever B is satisfied, we say that A is **impossible** (under the set of conditions B). If A may or may not occur whenever B is satisfied, then A is said to be a **random phenomenon**.

Random phenomena is our subject matter. Unlike certain and impossible events, the presence of randomness implies that the set of conditions B do not reflect all the necessary and sufficient conditions for the event A to occur. It might seem them impossible to make any worthwhile statements about random phenomena. However, experience has shown that many random phenomena exhibit a statistical regularity that makes them subject to study. For such random phenomena it is possible to estimate the chance of occurrence of the random event. This estimate can be obtained from laws, called **probabilistic** or **stochastic**, with the form:

$$\begin{array}{l}\text{if a set of conditions } B \text{ is satisfied} \\ \text{repeatedly } n \text{ times}\end{array} \implies \begin{array}{l}\text{event } A \text{ occurs } m \text{ times} \\ \text{out of the } n \text{ repetitions.}\end{array}$$

Probabilistic laws play an important role in almost every field of science.

If one assumes that $n \to \infty$, then "the probability that event A occurs" (under the set of conditions B) can be estimated as the ratio m/n.

Now, how are probabilities assigned to random events? Historically, there have been two approaches to study random phenomena, the *relative frequency* (or statistical) method and the *classical* (or a priori) method. The relative frequency method relies upon observation of occurrence of the event A under a large number of repetitions of the fixed set

of conditions B. Then, one counts the number of times that event A has occurred. The classical method, whose introduction is credited to Laplace (1812), makes use of the concept of equal likelihood, which is taken as a primitive of the model. Under this approach, if an event is regarded as the aggregation of several mutually exclusive and equally likely elementary events, the probability of such event is obtained as the sum of the individual probabilities of the elementary events.

These two approaches were joined in the 20th century by the axiomatic approach, introduced by A. N. Kolmogorov, which is consistent with both of them and allows for a systematic and rigorous treatment of general classes of random phenomena. This is the approach that we shall follow.

1.2. Probability spaces

We would like to use a development that allows us to assign probabilities to as many events as possible. We begin with an arbitrary nonempty set Ω of **elementary events** or **(sample) points** ω. In decision theory, the elementary events are also known as **states of the world**. Consider a collection or family $\mathcal{F}$ of subsets of Ω with a generic element $A \in \mathcal{F}$ (i.e., $A \subseteq \Omega$). Elements $A \in \mathcal{F}$ are called **(random) events**. We wish to impose formally conditions on $\mathcal{F}$ that allow us to assign probabilities in general probability problems. To do this, we use measurable structures, which are at the foundations of probability and statistics.

Intuitively, suppose that random events are described by sentences and we wish to assign probabilities to them. Given events A and B, it makes sense to connect such sentences so as to form new sentences like "A and B," "A or B," and "not A." Then, the family of events should be closed under intersections, unions, and complements. It should also include the entire set of elementary events. Such a family of sets is called an algebra (or a field) of sets. If we also wish to discuss results along the "law of averages," which have to do with the average behavior over an infinite sequence of trials, then it is useful to add closure under countable unions and intersections to our list of desiderata. An algebra that is closed under countable unions is a σ-algebra (or a σ-field). A nonempty set equipped with a σ-algebra of subsets is a measurable space and elements of this σ-algebra are called measurable sets or (random) events.

Definition 1. Let Ω be an arbitrary nonempty set. A nonempty family $\mathcal{F}$ of subsets of Ω is an **algebra** on Ω if $A, B \in \mathcal{F}$ implies:

(a) $A \cup B \in \mathcal{F}$;

(b) $\Omega \setminus A \in \mathcal{F}$.

A **$\boldsymbol{\sigma}$-algebra** on Ω is an algebra on Ω that is also closed under countable unions, i.e., for each sequence $\{A_n\}_{n=1}^{\infty} \subseteq \mathcal{F}$, we have $\cup_{n=1}^{\infty} A_n \in \mathcal{F}$. The elements of a σ-algebra are called **(random) events**.

Definition 2. A **measurable space** is a pair $(\Omega, \mathcal{F})$ where Ω is an arbitrary nonempty set and $\mathcal{F}$ is a σ-algebra on Ω.

Each algebra $\mathcal{F}$ on Ω contains $\emptyset$ and Ω.[1] Since $\mathcal{F}$ is nonempty by definition, there exists some $\emptyset \neq A \in \mathcal{F}$, so $\Omega \setminus A \in \mathcal{F}$ follows from (b) in the definition above. Hence, $\Omega = A \cup (\Omega \setminus A) \in \mathcal{F}$ and $\emptyset = \Omega \setminus \Omega \in \mathcal{F}$, respectively, from (a) and from (b) in the definition above. The events Ω and $\emptyset$ are called, respectively, the **certain** (or **sure**) event and the **impossible** event.

Notice that, by DeMorgan's law, $A \cap B = \Omega \setminus [(\Omega \setminus A) \cup (\Omega \setminus B)]$ and $A \cup B = \Omega \setminus [(\Omega \setminus A) \cap (\Omega \setminus B)]$. Thus, if $\mathcal{F}$ is a σ-algebra on Ω, then, since it is closed under complementation, we obtain that $\mathcal{F}$ is closed under the formation of finite unions if and only if it is closed under the formation of finite intersections. Then, we can replace requirement (a) in the definition above by (a)$'$ $A \cap B \in \mathcal{F}$. Analogously, by applying the infinite form of DeMorgan's law, we can replace the statement above to define a σ-algebra by the requirement that a σ-algebra is an algebra that is also closed under countable intersections, i.e., for each sequence $\{A_n\}_{n=1}^{\infty} \subseteq \mathcal{F}$, we have $\cap_{n=1}^{\infty} A_n \in \mathcal{F}$.

Example 1. Using set operations, formal statements regarding events are expressed as:

(1) "Event A does not occur": $\Omega \setminus A$;
(2) "Both events A and B occur": $A \cap B$;
(3) "Either event A, event B, or both occurs": $A \cup B$;
(4) "Either event A or event B occurs, but not both of them": $(A \setminus B) \cup (B \setminus A) =: A \Delta B$ (symmetric difference set operation);
(5) "Events A and B are mutually exclusive": $A \cap B = \emptyset$;
(6) "Event A occurs and event B does not occur": $A \setminus B$;
(7) "If event A occurs, then event B also occurs": $A \subseteq B$;
(8) "Neither event A nor event B occur": $\Omega \setminus (A \cup B)$.

Example 2. Consider the experiment of rolling a die once. Then $\Omega = \{1, \ldots, 6\}$. If we wish to be able to discern among all possible subsets of Ω, then we would take 2^{Ω} as our σ-algebra. However, suppose that we wish to model the pieces of information obtained by a person who is only told whether or not 1 has come up. Then, 2^{Ω} would not be the most appropriate σ-algebra. For instance, $\{1, 2, 3\} \in 2^{\Omega}$ is the event "a number less than 4 has come up," a piece of information that this person does not receive in this experiment. In this experiment, it makes more sense to choose a σ-algebra like $\{\emptyset, \Omega, \{1\}, \{2, \ldots, 6\}\}$.

Example 3. Consider the experiment of rolling a die arbitrarily many times. Then, $\Omega = \{1, \ldots, 6\} \times \{1, \ldots, 6\} \times \cdots = \{1, \ldots, 6\}^{\infty}$. Suppose that we wish to talk about the event "number 2 comes up in the ith roll of the die". Then, we should certainly choose a

[1] Some textbooks include explicitly $\Omega \in \mathcal{F}$ as a requirement in the definition of algebra.

σ-algebra on Ω that contains all sets of the form

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

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

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

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

The simplest example of a σ-algebra is $\{\emptyset, \Omega\}$, which is the smallest (with respect to set inclusion) σ-algebra on Ω. The largest possible σ-algebra on Ω is the **power class** 2^{Ω}, the collection of all subsets of Ω. In many probability problems, one often wishes to do the following. Beginning with a collection $\mathcal{F}$ of subsets of Ω, one searches for a family of subsets of Ω that (a) contains $\mathcal{F}$, (b) is a σ-algebra on Ω itself, and (c) is in a certain sense as small as possible. The notion of σ-algebra generated by $\mathcal{F}$ gives us precisely this.

Definition 3. Let Ω be an arbitrary nonempty set and let $\mathcal{F}$ be a nonempty family of subsets of Ω. The ***σ*-algebra generated by $\mathcal{F}$** is the family of subsets of Ω

$$\sigma(\mathcal{F}) := \bigcap_{i \in I} \left\{ \mathcal{F}_i \subseteq 2^{\Omega} : \text{ each } \mathcal{F}_i \supseteq \mathcal{F} \text{ is a } \sigma\text{-algebra on } \Omega \right\},$$

where I is an arbitrary index set.

From the definition above one immediately observes that

$$\sigma(\mathcal{F}) = \inf \left\{ \mathcal{G} \subseteq 2^{\Omega} : \mathcal{G} \supseteq \mathcal{F} \text{ and } \mathcal{G} \text{ is a } \sigma\text{-algebra on } \Omega \right\}.$$

Theorem 1. *Given a nonempty collection $\mathcal{F}$ of subsets of a nonempty set Ω, $\sigma(\mathcal{F})$ satisfies the following properties:*

(a) $\sigma(\mathcal{F})$ is a σ-algebra on Ω;
(b) $\mathcal{F} \subseteq \sigma(\mathcal{F})$;
(c) if $\mathcal{F} \subseteq \mathcal{G}$ and $\mathcal{G}$ is a σ-algebra on Ω, then $\sigma(\mathcal{F}) \subseteq \mathcal{G}$.

Proof. (a) First, take $A \in \sigma(\mathcal{F})$, then $A \in \mathcal{F}_i$ for each $\mathcal{F}_i \supseteq \mathcal{F}$, $i \in I$. Since each $\mathcal{F}_i$ is a σ-algebra on Ω, we have $\Omega \setminus A \in \mathcal{F}_i$ for each $\mathcal{F}_i \supseteq \mathcal{F}$, $i \in I$, and, therefore, $\Omega \setminus A \in \sigma(\mathcal{F})$. Second, take a sequence $\{A_n\}_{n=1}^{\infty} \subseteq \sigma(\mathcal{F})$, then $\{A_n\}_{n=1}^{\infty} \subseteq \mathcal{F}_i$ for each $\mathcal{F}_i \supseteq \mathcal{F}$, $i \in I$. Since each $\mathcal{F}_i$ is a σ-algebra on Ω, we have $\cup_{n=1}^{\infty} A_n \in \mathcal{F}_i$ for each $\mathcal{F}_i \supseteq \mathcal{F}$, $i \in I$. Therefore, $\cup_{n=1}^{\infty} A_n \in \sigma(\mathcal{F})$.
(b) The result follows directly from the definition of $\sigma(\mathcal{F})$, taking into account the set operations of inclusion and intersection.
(c) Take a σ-algebra $\mathcal{G}$ on Ω such that $\mathcal{G} \supseteq \mathcal{F}$. Then, it must be the case that $\mathcal{G} = \mathcal{F}_k$ for some $k \in I$ so that $\sigma(\mathcal{F}) \subseteq \mathcal{G}$. ∎

Definition 4. A **topology on a set** is a collection of subsets of the set that contains the empty set and the set itself, and that is closed under finite intersections and arbitrary (maybe uncountable!) unions. A member of a topology on a set is called an **open set** in such a set.

Notice that a σ-algebra on a countable set is also a topology on that set but the converse is not true.

Example 4. Consider 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, $\Omega \setminus \{a\} = \{b, c, d\} \notin \gamma$. 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. If we look for the σ-algebras generated by γ and τ, separately, we obtain

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

Let us now comment on a generated σ-algebra that is extensively used in probability and statistics.

Definition 5. Let Ω be an arbitrary nonempty set endowed with a topology τ. The **Borel σ-algebra on the topological space** (Ω, τ) is the σ-algebra generated by such topology $\mathcal{B}_\Omega := \sigma(\tau)$. The members $A \in \mathcal{B}_\Omega$ are called Borel sets in (Ω, τ) (or, simply, in Ω when the topology is understood from the context).

Therefore, the Borel σ-algebra of a set is a concept relative to the topology that one considers for this set. Of particular interest in this course is the Borel σ-algebra $\mathcal{B}_\mathbb{R}$ on $\mathbb{R}$. When no particular topology on $\mathbb{R}$ is provided, it is usually understood that the Borel σ-algebra on $\mathbb{R}$ is generated by the topology of *all* open sets in $\mathbb{R}$, that is,

$$\mathcal{B}_\mathbb{R} := \sigma\left(\{S \subseteq \mathbb{R} : S \text{ is open in } \mathbb{R}\}\right).$$

This notion of Borel σ-algebra can be directly extended to Euclidean spaces.

Example 5. Consider the collection of open intervals in $\mathbb{R}$,

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

We now show that $\sigma(\beta) = \mathcal{B}_\mathbb{R}$. Let α denote the collection of all open sets in $\mathbb{R}$. Since each open interval is an open set in $\mathbb{R}$, we have that $\beta \subseteq \sigma(\alpha)$. Then, $\sigma(\beta) \subseteq \sigma(\alpha)$ because $\sigma(\alpha)$ is a σ-algebra on $\mathbb{R}$. On the other hand, since each open set in $\mathbb{R}$ can be expressed as the result of the union of countably many open intervals, we know that $\alpha \subseteq \sigma(\beta)$. This

is so because, as a σ-algebra that contains β, $\sigma(\beta)$ must contain the unions of countably arbitrarily many open intervals. But, then $\sigma(\alpha) \subseteq \sigma(\beta)$ follows from the fact that $\sigma(\beta)$ is a σ-algebra on $\mathbb{R}$. Therefore, $\sigma(\alpha) = \sigma(\beta)$.

Example 6. Consider the collection of all bounded right-semiclosed intervals of $\mathbb{R}$,

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

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

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

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

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

Then, by an argument totally analogous to the previous ones, we have $\beta \subseteq \sigma(\delta)$ and, then, $\sigma(\beta) \subseteq \sigma(\delta)$. Therefore, $\sigma(\beta) = \sigma(\delta)$.

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

(1) the collection of all closed intervals,
(2) the collection of all bounded left-semiclosed intervals,
(3) the collection of all intervals of the form $(-\infty, a]$,
(4) the collection of all intervals of the form $[b, +\infty)$, and
(5) the collection of all closed sets.

The fact in (5) above that the collection of all closed sets induces $\mathcal{B}_{\mathbb{R}}$ implies that singletons and countable sets in $\mathbb{R}$ are members of its Borel σ-algebra (of course, provided that no reference to a particular topology different from the collection of all open sets is given!).

Up to now we have imposed conditions on the set of possible random events on which we can take probabilities. But how do we actually assign probabilities to such events? A. N. Kolmogorov (1929, 1933) developed the axiomatic approach to probability theory, which relates probability theory to set theory and to the modern developments in measure theory. Historically, mathematicians have been interested in generalizing the notions of length, area and volume. The most useful generalization of these concept is provided by the notion of a measure. Using these tools from measure theory, Kolmogorov's axioms follow from the definition of probability measure below.

Definition 6. A **set function** is an (extended) real function defined on a family of subsets of a measurable space.

In its abstract form, a measure is a set function with some additivity properties that reflect the properties of length, area and volume.

Definition 7. Let $(\Omega, \mathcal{F})$ be a measurable space, then a **measure** P on $(\Omega, \mathcal{F})$ is a set function $P : \mathcal{F} \to \mathbb{R}^*$, where $\mathbb{R}^* := \mathbb{R} \cup \{-\infty, +\infty\}$, satisfying:

(a) $P(\emptyset) = 0$ and $P(A) \geq 0$ for each $A \in \mathcal{F}$;

(b) (**σ-additivity**) if $\{A_n\}_{n=1}^{\infty} \subseteq \mathcal{F}$ is a sequence of pairwise disjoint events in $\mathcal{F}$, then

$$P\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \sum_{n=1}^{\infty} P(A_n).$$

If a measure P on $(\Omega, \mathcal{F})$ also satisfies $P(\Omega) = 1$, then it is called a **probability measure**.[2]

Definition 8. A **probability space** is a triple $(\Omega, \mathcal{F}, P)$ where Ω is an arbitrary nonempty set, $\mathcal{F}$ is a σ-algebra of subsets of Ω, and P is a probability measure on $(\Omega, \mathcal{F})$.

From the definition of probability measure above we can deduce some properties of probabilities that will be used extensively throughout the course. Let $(\Omega, \mathcal{F}, P)$ be a probability space. Then,

(P1) $P(\emptyset) = 0$, as stated directly in the definition of probability measure.
(P2) for each $A \in \mathcal{F}$, $P(\Omega \setminus A) = 1 - P(A)$.

Proof. Since $\Omega = (\Omega \setminus A) \cup A$ and $(\Omega \setminus A) \cap A = \emptyset$, σ-additivity implies

$$1 = P(\Omega) = P(\Omega \setminus A) + P(A) \Rightarrow P(\Omega \setminus A) = 1 - P(A). \qquad \blacksquare$$

(P3) for each $A \in \mathcal{F}$, $0 \leq P(A) \leq 1$.

Proof. Note that $P(A) \geq 0$ is stated directly in the definition of probability measure. Also, $P(A) \leq 1$ follows from the result $P(A) = 1 - P(\Omega \setminus A)$ shown above and from the fact that $P(\Omega \setminus A) \geq 0$ stated in the definition of probability measure. $\blacksquare$

(P4) for $A, B \in \mathcal{F}$, if $A \subseteq B$, then $P(A) \leq P(B)$.

Proof. Since we can write $B = A \cup [(\Omega \setminus A) \cap B]$ and $A \cap [(\Omega \setminus A) \cap B] = \emptyset$, σ-additivity implies

$$P(B) = P(A) + P((\Omega \setminus A) \cap B) \geq P(A). \qquad \blacksquare$$

[2] Given some set function $P : \mathcal{F} \to \mathbb{R}^*$, Kolmogorov's axioms are:
(i) $P(A) \geq 0$ for each $A \in \mathcal{F}$;
(ii) $P(\Omega) = 1$;
(iii) for each set $\{A_1, \ldots, A_n\}$ of pairwise disjoint events, we have $P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$.
Compare these axioms with the given definition of probability measure.

(P5) for each $A, B \in \mathcal{F}$, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Proof. Consider the following identities:

$$A \cup B = A \cup [B \setminus (A \cap B)],$$

$$B = [A \cap B] \cup [B \setminus (A \cap B)].$$

Since the sets united in each identity are disjoint, it follows from σ-additivity that

$$P(A \cup B) = P(A) + P(B \setminus (A \cap B)),$$

$$P(B) = P(A \cap B) + P(B \setminus (A \cap B)).$$

Then, the result follows by combining these two identities. ∎

(P6) for each $A, B \in \mathcal{F}$, $P(A \cup B) \leq P(A) + P(B)$.

Proof. The result follows directly from the property above (P5) combined with the fact that $P(A \cap B) \geq 0$, as stated directly in the definition of probability measure. ∎

(P7) for each sequence $\{A_n\}_{n=1}^{\infty} \subseteq \mathcal{F}$ of events (not necessarily disjoint!), $P(\cup_{n=1}^{\infty} A_n) = 1 - P(\cap_{n=1}^{\infty} \Omega \setminus A_n)$, that is, the probability of at least one of the events A_n will occur is 1 minus the probability that none of the events will occur.

Proof. By applying the infinite form of DeMorgan's law, we have

$$\cup_{n=1}^{\infty} A_n = \Omega \setminus [\cap_{n=1}^{\infty} \Omega \setminus A_n].$$

Then, the result follows using Property (P2). ∎

Definition 9. A **Borel measure** is a measure defined on the Borel sets of a topological space.

One of the most important examples of measures is the Lebesgue measure on the real line and its generalizations to Euclidean spaces. It is the unique measure on the Borel sets whose value on each interval is its length.

Definition 10. The **Lebesgue measure on** $\mathbb{R}$ is the set funtion $\lambda : \mathcal{B}_{\mathbb{R}} \to [0, +\infty)$ specified by $\lambda((a, b]) := b - a$ for each $(a, b] \in \mathcal{B}_{\mathbb{R}}$.

Similarly, we can define the Lebesgue measure on $\mathbb{R}^n$ by assigning to each rectangle its n-dimensional "volume". For $a, b \in \mathbb{R}^n$ such that $a_i \leq b_i$, $i = 1, \ldots, n$, let $(a, b] := \times_{i=1}^{n} (a_i, b_i]$. The **Lebesgue measure on** $\mathbb{R}^n$ is then the set funtion $\lambda : \mathcal{B}_{\mathbb{R}^n} \to [0, +\infty)$ defined by $\lambda((a, b]) := \prod_{i=1}^{n} (b_i - a_i)$.

Sometimes, one starts with a probability measure defined on a small σ-algebra (or even on an algebra!) and then wishes to extend it to a larger σ-algebra. For example,

Lebesgue measure can be constructed by defining it first on the collection of finite unions of right-semiclosed intervals and extending it to the collection of Borel sets of $\mathbb{R}$, which is generated by the former collection. The following example illustrates why extending probability measures should interests us.

Example 7. Consider the experiment of tossing a coin three times. Let 0="heads" and 1="tails." Then, $\Omega = \{0,1\}^3$ and $|\Omega| = 2^3$. Since Ω is finite, we can take $\mathcal{F} = 2^\Omega$ right away. Now, notice that the finiteness of Ω allows us to assign probabilities (in quite an intuitive way) to the events in $\mathcal{F}$ using the notion of relative frequency. Therefore, for $S \in \mathcal{F}$, we can take $P(S) = |S| / |\Omega|$. For instance, the probability of the event S ="at least one of the last two tosses does not come up heads" would be computed by noting that

$$S = \{(0,1,0),(0,0,1),(0,1,1),(1,1,0),(1,0,1),(1,1,1)\},$$

so that $P(S) = 6/2^3 = 3/4$. Now, how should we take the space of events and how should we compute probabilities if the sample set were infinite? Suppose that the coin is tossed infinitely many times. Then, $\Omega = \{0,1\}^\infty$. To consider the space of events, let us deal first with "easy" events. For instance, consider the set

$$A_S = \{(\omega_i)_{i=1}^\infty \in \{0,1\}^\infty \,:\, (\omega_1,\omega_2,\omega_3) \in S\},$$

where $S \subset \{0,1\}^3$ is the event obtained above. This type of sets is known as a cylinder set and it is specified by imposing conditions only on a finite number of the coordinates of a generic element of it. More generally, given a set $T \subseteq \{0,1\}^k$ for some finite k, a cylinder set is a set of the form

$$A_T = \{(\omega_i)_{i=1}^\infty \in \{0,1\}^\infty \,:\, (\omega_1,\ldots,\omega_k) \in T\}.$$

Notice that this cylinder set is nothing but the event "the outcome of the first k tosses belongs to T." Then, if we use the notion of relative frequency to compute probabilities, we would like to assign to it the probability $|T| / 2^k$. For instance, using the earlier event S to construct the cylinder set A_S, we would compute the probability of the event "out of first three tosses, at least one of its last two tosses does not come up heads" as $P(A_S) = |S| / 2^k = 3/4$. In this manner, as k increases, we would be able to compute probabilities of events when the number of tosses becomes arbitrary. Then, it makes sense that we include the set of all cylinder sets

$$\begin{aligned}\mathcal{A} &= \cup_{k=1}^\infty \left\{A_T \,:\, T \subseteq \{0,1\}^k\right\} \\ &= \bigcup_{k=1}^\infty \left\{\{(\omega_i)_{i=1}^\infty \in \{0,1\}^\infty \,:\, (\omega_1,\ldots,\omega_k) \in T\} \,:\, T \subseteq \{0,1\}^k\right\}\end{aligned}$$

in our event space. Therefore, we would like to consider $\mathcal{A}$ as the nucleus of our event space so that $\sigma(\mathcal{A})$ would be the chosen σ-algebra. We have that $\mathcal{A}$ is an algebra (check

it!) but need not be a σ-algebra. So it seems worrisome that we only know how to assign probabilities to the elements in $\mathcal{A}$. What do we do with the events in $\sigma(\mathcal{A}) \setminus \mathcal{A}$?

A general method for extending measures was developed by C. Carathéodory (1918) and is known as the **Carathéodory extension procedure**. We present briefly the approach followed by this procedure and some useful consequences of it. Intuitively, the Carathéodory extension method allows us to construct a probability measure on a σ-algebra by specifying it only on the algebra that generates that σ-algebra. Furthermore, under fairly general conditions, such a construction is done in a unique way. We begin with a formulation of Carathéodory's Theorem which reflects precisely this intuition.

Theorem 2. *Let $\mathcal{A}$ be an algebra on a nonempty set Ω and let $P : \mathcal{A} \to \mathbb{R}^*$. If P is σ-additive, then there exists a measure P^* on $\sigma(\mathcal{A})$ such that $P^*(A) = P(A)$ for each $A \in \mathcal{A}$. Moreover, if $P(\Omega) < \infty$, then P^* is unique.*

Now, we get into the formal details of the procedure and present another formulation of Carathéodory's Theorem.

Definition 11. 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 12. Let P be a measure on a measurable space $(\Omega, \mathcal{F})$. The measure P generates a set function $P^* : 2^\Omega \to \mathbb{R}^*_+$ defined by

$$P^*(A) := \inf \left\{ \sum_{n=1}^\infty P(A_n) \, : \, \{A_n\}_{n=1}^\infty \subset \mathcal{F} \text{ and } A \subset \cup_{n=1}^\infty A_n \right\}, \qquad (*)$$

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 $(\Omega, \mathcal{F})$ and the σ-algebra generated by $\mathcal{F}$, $\mathcal{F}^* = \sigma(\mathcal{F})$. Then, $\mathcal{F}^*$ is a σ-algebra larger than $\mathcal{F}$ (in the sense that $\mathcal{F}^* \supseteq \mathcal{F}$). The formulation of Carathéodory's Theorem stated below asserts that there exists an outer measure P^* on $(\Omega, \mathcal{F}^*)$ such that:
(a) $P^*(A) = P(A)$ for each $A \in \mathcal{F}$;
(b) if Q is another measure on $(\Omega, \mathcal{F}^*)$ such that $Q(A) = P(A)$ for each $A \in \mathcal{F}$, then it must be the case that $Q(A) = P^*(A)$ for each $A \in \mathcal{F}$.

Theorem 3. *A measure P on a measurable space $(\Omega, \mathcal{F})$ such that $P(\Omega) < \infty$ has a unique extension P^* (i.e., conditions (a) and (b) above are satisfied), defined by equation (*) above, to the generated σ-algebra $\sigma(\mathcal{F})$. 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 $(\Omega, \mathcal{F}, P)$, 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 $A \in \mathcal{F}$ 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 $P^*(\{a \in A : f(a) \neq f(a)\}) = 0$. We use the symbol $=_{\text{a.e.}}$ to denote "equal almost everywhere"

1.3. Conditional probability and Bayes' theorem

In many problems there is some information available about the outcome of the random phenomenon at the moment at which we assing probalities to events. Hence, one may face questions of the form: "what is the probability that event A occurs given that another event B has occurred?"

Definition 13. Let $(\Omega, \mathcal{F}, P)$ be a probability space. Consider two events $A, B \in \mathcal{F}$ such that $P(B) > 0$, then **the conditional probability of A given B** is given by:

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

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

Using the definition of conditional probability, we obtain the **chain-rule formulas**

$$P(A \cap B) = P(B)P(A|B),$$
$$P(A \cap B \cap C) = P(B)P(A|B)P(C|B \cap A),$$

and so on. Furthermore, if $\{B_n\}_{n=1}^{\infty}$ partitions Ω, then, for $A \in \mathcal{F}$, it can be shown that

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

This property is known as the **law of total probability**.

If one begins with a probability space $(\Omega, \mathcal{F}, P)$ and considers an event $B \in \mathcal{F}$ such that $P(B) > 0$, then it can be shown that the set function $P(\cdot|B) : \mathcal{F} \to [0, 1]$ specified in the definition above is a well defined probability measure (check it!) on the measure space $(\Omega, \mathcal{F})$ so that $(\Omega, \mathcal{F}, P(\cdot|B))$ constitutes a probability space.

Theorem 4 (Bayes' rule). *Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\{A_1, \ldots, A_n\} \subseteq \mathcal{F}$ a set of n mutually disjoint events such that $P(A_i) > 0$ for each $i = 1, \ldots, n$ and $\cup_{i=1}^n A_i = \Omega$. If $B \in \mathcal{F}$ is an event such that $P(B) > 0$, then*

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

Proof. Since

$$B = B \cap \Big(\bigcup_{i=1}^n A_i\Big) = \bigcup_{i=1}^n \Big(B \cap A_i\Big)$$

and $\{B \cap A_1, \ldots, B \cap A_n\}$ is a set of disjoint events, we obtain

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

However, by applying the definition of conditional probability, we have

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

Therefore, using the definition of conditional probability again, we can write, for each $k = 1, \ldots, n$,

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

as stated. ∎

Example 8. A ball is drawn from one of two urns depending on the outcome of a roll of a fair die. If the die shows 1 or 2, 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, the ball is drawn from Urn II which contains 7 red balls and 3 white balls. We ask ourselves: given that a white ball is drawn, what is the probability that it came from Urn I? from Urn II?

Let I (II) denote the event "the ball comes from Urn I (Urn II)." Let w (r) denote the event "the drawn ball is white (red)." We compute $P(I|w)$ and $P(II|w)$ by applying Bayes' rule:

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

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

1.4. Independence

Let A and B be two events in a probability space. An interesting case occurs when knowledge that B occurs does not change the odds that A occurs. We may think intuitively that A and B occur in an *independent way* if $P(A|B) = P(A)$ when $P(B) > 0$. Hence, the definition of conditional probability leads to the following definition

Definition 14. Let $(\Omega, \mathcal{F}, P)$ be a probability space. **Two events** $A, B \in \mathcal{F}$ are **independent** if $P(A \cap B) = P(A)P(B)$.

More generally, **a finite collection** $A_1, \ldots, A_n$ of events is **independent** if

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

for each $2 \leq j \leq n$ and each $1 \leq k_1 < \cdots < k_j \leq n$. That is, a finite collection of events is independent if each of its subcollections is. Analogously, **an infinite (perhaps uncountable) collection of events** is defined to be **independent** if each of its finite subcollections is.

Example 9. Consider an experiment with a sample set $\Omega = \{a, b, c, d\}$ and suppose that the probability of each outcome $\omega \in \Omega$ is $1/4$. Consider the following three events

$$A = \{a, b\}\,,\ B = \{a, c\}\,,\ C = \{a, d\}\,.$$

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 must say that events A, B, and C are **pairwise independent** but all three of them are not independent.

Example 10. Let $\Omega = \{(x, y) \in \mathbb{R}^2 : 0 \leq x, y \leq 1\}$ and consider the probability space $(\Omega, \mathcal{B}_\Omega, \lambda)$ where λ is the Lebesgue measure on $\mathbb{R}^2$. It can be shown that the events

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

are independent. To do this, let us compute the area of the respective rectangles. First, notice that

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

Then, one obtains $\lambda(A) = 1/2$, $\lambda(B) = 1/4$, and $\lambda(A \cap B) = 1/8$, as required.

Consider now the event

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

We have $\lambda(C) = 1/2 - (1/2)^3 = 3/8$ and $\lambda(C \cap B) = 1/2(1/4)^2 = 1/32$ so that $\lambda(C)\lambda(B) = 3/32 \neq 1/32 = \lambda(C \cap B)$, that is, C and B are not independent.

Problems

1. [Infinite form of DeMorgan's law] Let $A_1, A_2, \ldots$ be an infinite sequence of distinct subsets of some nonempty set Ω. Show that
(a) $\Omega \setminus (\cup_{n=1}^{\infty} A_n) = \cap_{n=1}^{\infty}(\Omega \setminus A_n)$.
(b) $\Omega \setminus (\cap_{n=1}^{\infty} A_n) = \cup_{n=1}^{\infty}(\Omega \setminus A_n)$.

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

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

4. Let $(\Omega, \mathcal{F})$ be a measurable space. For $A \in \mathcal{F}$, let $\mathcal{F}(A)$ be the collection of subsets of Ω of the form $A \cap B$, where $B \in \mathcal{F}$. Show that, for a given $A \in \mathcal{F}$, $\mathcal{F}(A)$ is a σ-algebra on Ω.

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

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

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

6. Let $(\Omega, \mathcal{F}, P)$ be a probability space and, for $A \in \mathcal{F}$, let $P_A : \mathcal{F} \to [0, 1]$ be a set function defined by $P_A(B) := P(A \cap B)$ for each $B \in \mathcal{F}$.
(a) Show that, for a given $A \in \mathcal{F}$, P_A is a probability measure on $(\Omega, \mathcal{F})$.
(b) Show that, for a given $A \in \mathcal{F}$ such that $P(A) > 0$, the set function Q_A on $\mathcal{F}$ defined by $Q_A(B) := P_A(B)/P(A)$ for each $B \in \mathcal{F}$ is a probability measure on $(\Omega, \mathcal{F})$.

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

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

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

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

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

10. 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?

11. 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.

12. 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.

13. 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.
Suppose now that it is known that at least one of the three removed balls is black.
(b) Compute the probability that all three of the removed balls is black.

14. 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.

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

Combinatorics

Consider a finite sample set Ω and suppose that its elementary events are equally likely (as considered by the classical approach to probability theory). Then, using the relative frequency interpretation of probability, we can compute the probability of an event $A \subseteq \Omega$ simply by dividing the cardinality of A over the cardinality of Ω.

The aim of this chapter is to introduce several combinatorial formulas which are commonly used for counting the number of elements of a set. These methods rely upon special structures that exist in some common random experiments.

2.1. Ordered samples

Consider a finite set $S := \{1, 2, \ldots, s\}$ and suppose that we are interested in drawing a sequence of $m \leq s$ elements from this set. In this case, we care about the order of the draws. Then, the outcome of the draws can then be regarded as an m-tuple or sequence $\omega = (\omega_1, \omega_2, \ldots, \omega_m)$, where ω_i is the element in the ith draw.

Suppose first 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 called **random sampling with replacement**. Here, we have $\Omega = S^m$ so that $|\Omega| = s^m$.

On the other hand, suppose that we do not return the elements into the set before the following draw. Then, $\Omega = \{(\omega_i)_{i=1}^m : \omega_i \neq \omega_j \text{ for each } i \neq j\}$. This procedure is known as **random sampling without replacement**. Here, $|\Omega| = s(s-1)(s-2)\cdots(s-m+1)$. Let $(s)_m := s(s-1)(s-2)\cdots(s-m+1)$ denote the number of different possible m-tuples when there is no replacement. Also, if the elements from S are drawn $m = s$ times without replacement, then there are $(s)_s =: s!$ possible outcomes of the experiment.

Example 11. Suppose that we are asked to compute the probability that there shows up at least one head when a fair coin is tossed n times. We can regard flipping the coin n times as drawing a random sample with replacement of size n from the set $\{0, 1\}$, where head is indicated by 0. Therefore, $|\Omega| = 2^n$. Let C be the event "there shows up at least one zero" and let B_i be the event that "the ith toss results in a zero." Then $C = \cup_{i=1}^n B_i$, and note that the sets $B_1, \ldots, B_n$ are not pairwise disjoint! However, using property (P2) of probabilities and DeMorgan's law, we have

$$P(C) = 1 - P(\Omega \setminus C) = 1 - P(\Omega \setminus \cup_{i=1}^n B_i) = 1 - P\Big(\cap_{i=1}^n (\Omega \setminus B_i)\Big).$$

But $\cap_{i=1}^n (\Omega \setminus B_i)$ consist of the event "the n tosses yield ones," i.e., $\cap_{i=1}^n (\Omega \setminus B_i) =$

$\{(1,\ldots,1)\}$. Then, $P(\cap_{i=1}^{n}(\Omega \setminus B_i)) = 2^{-n}$, and the sought probability is $P(C) = 1 - 2^{-n}$.

Let us now consider the following problem: a random sample of size m is chosen from a set S of s distinct objects with replacement. We ask ourselves about the probability of the event A="in the sample no element appears twice." Note that the cardinality of the sample set 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 the set

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

But this is precisely the cardinality of the sample set (or sure event) associated with an experiment of random sampling without replacement from that set S! So, the sought probability can be computed as

$$\frac{s(s-1)(s-2)\cdots(s-m+1)}{s^m} = \left(1-\frac{1}{s}\right)\left(1-\frac{2}{s}\right)\cdots\left(1-\frac{m-1}{s}\right).$$

A typical problem of this sort is known as the birthday problem.

Example 12. Suppose that we are asked to compute the probability of the event A ="no two people from a group of five friends have a common birthday." Here we shall ignore the leap years and the fact that birth rates are not exactly equally likely over the year. Using the expression obtained above, here $m = 5$ and $s = 365$, so that we can easily compute

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

2.2. Permutations

Suppose that we have a set S of s distinct objects, which we permute randomly among themselves. Then, we ask about the final positions of some pre-specified objects. Here, we should identify each position i after the rearrangement with the element ω_i drawn from the set S. Also, notice that, since two distinct objects from S cannot end up in the same position, we are considering random sampling without replacement and, consequently, the number of possible ways of distributing the s objects into the s final positions is $(s)_s = s!$. This is the cardinality of the sample set of our experiment, i.e., $|\Omega| = s!$.

Now, let M be a strict subset of size m of S, and consider the event A ="$m < s$ pre-specified objects from S end up in m pre-specified positions." Given that m elements from S are required to end up in fixed positions, the number of sequences with $s - m$ coordinates that can be extracted without replacement from the set $S \setminus M$ is $(s-m)!$. Therefore, $|A| = (s-m)!$ and the probability that m specified objects from S end up in m specified positions after permuting randomly among themselves the s distinct objects

is

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

On the other hand, notice that the number of sequences that can be obtained from S if only $m < s$ objects can be used in the permutations is $(s)_m = s(s-1)\cdots(s-m+1)$. This is the number of different permutations with m objects that can be made from a set of $s > m$ objects if the final order of the elements matters. In other words, we are computing the number of ways of obtaining an ordered subset of m elements from a set of $s > m$ elements. We denote this number by P_m^s,

$$P_m^s = (s)_m = s(s-1)\cdots(s-m+1),$$

which is nothing but the inverse of the probability, calculated above, that m pre-specified elements from S end up in m pre-specified positions after permuting randomly the s objects.

Example 13. A committee of 5 members, consisting of a president, a secretary and three officials is selected from a club of 50 members. The officials are ranked as official 1, 2 and 3, according to the degree of their importance within the club. The presidency and the secretary position are 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. We are interested in computing 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 $\{1, \ldots, 50\}$ 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 $1/P_3^{48} = 1/(48 \cdot 47 \cdot 46)$.

2.3. Combinations

Suppose that we have a set S of s distinct objects, which we permute randomly among themselves. Again, we should identify each position i after the rearrangement with the element ω_i drawn from the set S. Also, since two distinct objects from S cannot end up in the same position, we are considering random sampling without replacement, as in the section above.

Now we are interested in computing the number of different subsets of size $m \leq s$ that can be extracted from S. In other words, we wish to compute the number of sequences that can be obtained if the order of its coordinates does not matter. First, notice that there are $(s)_m$ different sequences of size m that can be drawn from S without replacement. But, instead of focusing on ordered m-tuples, we care indeed about the subsets of size m that can be drawn from S disregarding the order in which its elements are selected. Notice 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 the same. Therefore, there are $(s)_m/m!$ different samples of size m that can be drawn from S without replacement and regardless the order of its elements. Using the binomial operator, we usually write

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

Example 14. Consider again the problem in Example 11 above, where we were interested in computing the probability of event $C =$"at least one zero (head) shows up" when a fair coin is tossed n times. Consider the event $A_i =$"there shows up *exactly* i zeroes." Then, $C = \cup_{i=1}^n A_i$ and $A_i \cap A_j = \emptyset$ for each $i, j = 1, \ldots, n$ such that $i \neq j$. So, $P(C) = \sum_{i=1}^n P(A_i)$. To compute each $P(A_i)$, note that $\binom{n}{i}$ gives us the number of subsets of size i that can be extracted from $1, \ldots, n$. In other words, $\binom{n}{i}$ gives us the cardinality of the event that "i tosses result in zero shows up while, at the same time, the remaining $n - i$ tosses yield one." This is precisely the cardinality of A_i. Therefore,

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

which coincides with $1 - 2^{-n}$, as obtained in Example 11.

Example 15. The economics department consists of 8 full professors, 14 associate professors, and 18 assistant professors. A committee of 5 is selected at random from the faculty of the department. Suppose that we are asked 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, the committee of five can be chosen from the forty in $\binom{40}{5}$ ways. There are 18 assistant professors so that the committee of five can be chosen from them in $\binom{18}{5}$ ways. Therefore, the sought probability is $\binom{18}{5} \Big/ \binom{40}{5}$.

Example 16. A die is rolled 12 times and suppose first that we are interested in computing the probability of getting exactly 2 fives. Let A denote that event of interest. Here note that $\Omega = \{1, \ldots, 6\}^{12}$ so that $|\Omega| = 6^{12}$. Now consider the event $A_{(i,j)}$, where $i, j = 1, \ldots, 12$, $i < j$, which describes the outcome such that number 5 shows up *only* in the ith and jth rolls. Then, we have $\left|A_{(i,j)}\right| = 5^{10}$ regardless the value of the pair (i, j). Also, we know that $A_{(i,j)} \cap A_{(k,l)} = \emptyset$ whenever $(i, j) \neq (k, l)$ and

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

where Q is the set specified as

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

Therefore, we know that

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

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

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

Suppose now that we wish to compute the probability that at least 1 one shows up. Let B denote that event of interest and consider the event B_k, where $k = 0, 1, 2, \ldots, 12$, which describes the outcome such that number 1 shows up exactly k times. Then, we have $B = \cup_{k=1}^{12} B_k$ and $B_k \cap B_l = \emptyset$ whenever $k \neq l$. Therefore, we know that $P(B) = \sum_{k=1}^{12} P(B_k)$. Following the same reasoning as above, we finally obtain

$$P(B) = \frac{\sum_{k=0}^{12} \binom{12}{k} 5^{12-k}}{6^{12}}.$$

Example 17. A set of n balls is distributed randomly into n boxes and we are asked to compute the probability that only the first box ends up being empty. Here, an elementary event should be identified with the final position of the balls so that ω_i should be interpreted as the box where the ith ball ends up. Then, the sample space is $\Omega = \{1, \ldots, n\}^n$ so that $|\Omega| = n^n$. Notice that we are considering random sampling with replacement since two different balls may end up in the same box.

Consider the event $A =$"only box 1 ends up being empty." Notice that this can happen if and only if exactly one of the remaining $n-1$ boxes contains two balls and all the other $n-2$ boxes have exactly one ball each. Consider then the event $B_i =$"box 1 ends up empty, box i ends up with two balls, and the remaining $n-2$ boxes end up with exactly one ball each." We have $A = \cup_{i=2}^{n} B_i$ and $B_i \cap B_j = \emptyset$ whenever $i \neq j$.

To compute $P(B_i)$, notice first that the number of subsets that can be extracted from $\{1, \ldots, n\}$ containing two balls is $\binom{n}{2}$. Then, the remaining $(n-2)$ balls can be rearranged in the remaining $(n-2)$ boxes in $(n-2)!$ different ways. Therefore, the number of distinct ways in which one can put no ball in box 1, two balls into box i, and exactly one ball in each of the remaining boxes is $\binom{n}{2}(n-2)!$. So, we obtain

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

Consequently, the sought probability is

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

2.4. Partitions

Many combinatorial problems involving unordered samples are of the following type. There is a box that contains r red balls and b black balls. A random sample of size m is drawn from the box without replacement. What is 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 is partitioned into two classes. A random sample of a certain size is taken and we inquire about the probability that the sample contains a specified number of elements of the two classes.

First, note 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 collection of all samples of size m drawn from a set of $b + r$ without replacement and without regard to order. As argued before, the probability that we must assign to each of these samples is

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

We must now compute 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

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

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 sought probability can be computed as

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

Example 18. We have a box containing r numbered balls. A random sample of size $n < r$ is drawn without replacement and the numbers of the balls are noted. These balls are then returned to the box, and a second random sample of size $m < r$ is drawn without

replacement. Suppose that we are asked to compute the probability that the two samples have exactly l balls in common. To solve this problem, we can proceed as follows. The first sample partitions the balls into two classes, those n selected and those $r-n$ not chosen. The problem then turns out to consists of computing the probability that the sample of size m contains exactly l balls from the first class. So, the sought probability is

$$\binom{n}{l}\binom{r-n}{m-l}\binom{r}{m}^{-1}.$$

Problems

1. 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.

2. 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.

3. 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.

4. 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.

5. 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.

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

7. 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.

8. 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.

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

10. 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.

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

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

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

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

Random variables and their distributions

3.1. Definitions

From here onwards we shall use the following bracket notation for inverse images. Given two sets Ω and Ω' and a function $f : \Omega \to \Omega'$, for $A \subseteq \Omega'$, we write $[f \in A]$ to denote $f^{-1}(A) = \{\omega \in \Omega : f(\omega) \in A\}$. Also, for simplicity, we shall write throughout $P[\cdot]$ instead of $P([\cdot])$ when using that bracket notation for inverse images.

Definition 15. Let $\mathcal{F}_\Omega$ and $\mathcal{F}_{\Omega'}$ be two nonempty families of subsets of two arbitrary nonempty sets Ω and Ω', respectively. A function $f : \Omega \to \Omega'$ is **$(\mathcal{F}_\Omega, \mathcal{F}_{\Omega'})$-measurable** if

$$f^{-1}(B) = \{\omega \in \Omega \,:\, f(\omega) \in B\} = [f \in B] \in \mathcal{F}_\Omega$$

for each $B \in \mathcal{F}_{\Omega'}$. When the families $\mathcal{F}_\Omega$ and $\mathcal{F}_{\Omega'}$ are understood from the context, we simply say that f is **measurable**.

Even though it is not required by the definition above, usually $\mathcal{F}_\Omega$ and $\mathcal{F}_{\Omega'}$ are σ-algebras. In the special case of a real-valued function $f : \Omega \to \mathbb{R}$, given a σ-algebra $\mathcal{F}$ of subsets of Ω, we say that f is **$\mathcal{F}$-measurable** if it is $(\mathcal{F}, \mathcal{B}_\mathbb{R})$-measurable.

Definition 16. Let $(\Omega, \mathcal{F})$ be a measurable space, a **random variable** on $(\Omega, \mathcal{F})$ is a real-valued $\mathcal{F}$-measurable function $X : \Omega \to \mathbb{R}$. If a random variable X assumes finitely many values, then it is called a **simple random variable**.

The point of the definition of random variable X is to ensure that, for each Borel set $B \in \mathcal{B}_\mathbb{R}$, $X^{-1}(B)$ can be assigned a measure or probability. The notion of random variable is a key concept in probability theory. It gives us a transformation through which we can assign probabilities on subsets of arbitrary sets by means of doing so on subsets of the real line. Note that the definition of a random variable is not attached to a probability measure. However, we need a probability measure to speak of the distribution of a random variable.

3.2. Probability distribution of a random variable

Definition 17. Let $(\Omega, \mathcal{F}, P)$ be a probability space and X a random variable on $(\Omega, \mathcal{F})$. The **probability distribution** of the random variable X is the probability measure ψ

on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ defined by

$$\psi(B) := P(X^{-1}(B)) = P(\{\omega \in \Omega \,:\, X(\omega) \in B\}) = P[X \in B], \quad \text{for each} \quad B \in \mathcal{B}_{\mathbb{R}}.$$

Moreover, the **distribution function** of the random variable X is the function $F : \mathbb{R} \to \mathbb{R}$ defined by

$$F(x) := \psi((-\infty, x]) = P(\{\omega \in \Omega \,:\, X(\omega) \leq x\}) = P[X \leq x].$$

Example 19. Consider the experiment of rolling three dice together and suppose that we are interested in the sum of the numbers that show up. In principle, we could take as our primitive the probability space $(\Omega, \mathcal{F}, P)$, where $\Omega = \{1, \ldots, 6\}^3$ so that $\omega = (\omega_1, \omega_2, \omega_3)$ is a generic elementary event, $\mathcal{F} = 2^{\Omega}$ and P is specified by $P(A) = |A|/6^3$ for each $A \in \mathcal{F}$. But this approach would not be particularly useful since we are interested only on the sum of the numbers that show up. Alternatively, we can make use of a function $X : \Omega \to \mathbb{R}$ specified as

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

It can be checked that this function X is measurable and, therefore, a random variable. Now, consider the event $A = (3, 5] \in \mathcal{B}_{\mathbb{R}}$. Using the concept of probability distribution of X, we can compute the probability that "the sum of the numbers that show up is larger that three but no larger than five" as

$$\psi(A) = P[X^{-1}(A)] = \frac{|\{(\omega_1, \omega_2, \omega_3) \in \Omega \,:\, 3 < \omega_1 + \omega_2 + \omega_3 \leq 5\}|}{6^3} = \frac{9}{6^3}.$$

Using the definition of distribution function, we can apply limits, for some $\varepsilon > 0$, to compute

$$P[X < x] = \lim_{\varepsilon \to 0} P[X \leq x - \varepsilon] = \lim_{t \to x^-} F(t).$$

Also,

$$P[X = x] = F(x) - \lim_{t \to x^-} F(t).$$

The following result gives us a useful characterization of a distribution function. Some textbooks take this result as the definition of distribution function.

Theorem 5. *Let $F : \mathbb{R} \to \mathbb{R}$ be a nondecreasing, right-continuous function satisfying*

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

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

Proof. Let the probability space $(\Omega, \mathcal{F}, P)$ be $((0,1), \mathcal{B}_{(0,1)}, \lambda)$, where λ is the Lebesgue measure on $((0,1), \mathcal{B}_{(0,1)})$. To understand the proof, suppose first that F is strictly increasing and continuous. Define the mapping $\theta := F^{-1}$. Since $F : \mathbb{R} \to (0,1)$ is a one-to-one function[3], $\theta : (0,1) \to \mathbb{R}$ is an increasing function. For $\omega \in (0,1)$, let $X(\omega) := \theta(\omega)$. Since θ is increasing, then X is $\mathcal{B}_{(0,1)}$-measurable. Now take a given $\omega \in (0,1)$. Then, we have $\theta(\omega) \leq x \Leftrightarrow \omega \leq F(x)$. Since λ is Lebesgue measure, we obtain

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

as required.

Now, if F has discontinuities or is not strictly increasing, define, for $\omega \in (0,1)$,

$$\theta(\omega) := \inf\{x \in \mathbb{R} : \omega \leq F(x)\} \equiv \inf\{x \in \mathbb{R} : \theta(\omega) \leq x\}.$$

Note that, since F is nondecreasing and right-continuous, then $\{x \in \mathbb{R} : \omega \leq F(x)\}$ is an interval with the form $[\theta(\omega), +\infty)$ for some $\omega \in (0,1)$ (i.e, it is closed on the left and stretches to ∞). Therefore, we obtain again $\theta(\omega) \leq x \Leftrightarrow \omega \leq F(x)$ so that, by defining $X(\omega) := \theta(\omega)$ for $\omega \in (0,1)$ and by applying the same argument as above, we obtain that X is a random variable on $((0,1), \mathcal{B}_{(0,1)}, \lambda)$ and $P[X \leq x] = F(x)$. ∎

Consider a probability space $(\Omega, \mathcal{F}, P)$ and a random variable X on it. Then, given a Borel set $(a, b] \subset \mathbb{R}$, it follows immediately from the definition of distribution function that, if F is the distribution function of X, then

$$P[a < X \leq b] = P[X \leq b] - P[X \leq a] = F(b) - F(a).$$

3.3. Discrete random variables

Definition 18. Let P be a probability measure on some measurable space $(\Omega, \mathcal{F})$. A **support** for P is an event $\text{supp}(P) \in \mathcal{F}$ such that $P(\text{supp}(P)) = 1$.

Definition 19. A random variable X on a probability space $(\Omega, \mathcal{F}, P)$ with probability distribution ψ is **discrete** if ψ has a countable support $\text{supp}(\psi) = \{x_1, x_2, \dots\}$.[4] In this case, ψ is completely determined by the values $\psi(\{x_i\}) = P[X = x_i]$ for $i = 1, 2, \dots$. If a random variable is discrete, then we say that its probability distribution and its distribution function are discrete as well.

We observe from the definition above that the support of a probability measure need not be unique. Without loss of generality, we shall restrict attention from here onwards

[3] A function $f : A \to B$ is **one-to-one**, or an **injection**, if for each $y \in B$ there is at most one $x \in A$ satisfying $f(x) = y$.

[4] Notice that each simple random variable is a discrete random variable but the converse is not true.

to a particular support. Intuitively, in this support we shall consider only events with positive probability. More precisely, for a probability measure P on $(\Omega, \mathcal{F})$, we will focus on the support $\text{supp}(X)$ such that $A \in \mathcal{F}$ and $P(A) = 0$ implies that A is not a subset of $\text{supp}(X)$. The support of the probability distribution of a random variable X is often called the **support of the random variable** X itself, which we denote by $\text{supp}(X)$. In this case, the support of the random variable X is identified with the range of X considered as a function. If X is a discrete random variable on a probability space $(\Omega, \mathcal{F}, P)$ with support $\{x_1, x_2, \dots\}$, then we can compute

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

3.4. Integration

Before presenting the notion of continuous random variable, we need to introduce the concept of integral of a function. Integration is the approach used in modern mathematics to compute areas and volumes. Hence, integration is a tool closely related to the notion of measure and, in particular, to the Lebesgue measure. A detailed treatment of the notion of integral is not within the scope of this course and, therefore, in this section we just introduce briefly its formal definition.

Consider a probability space $(\Omega, \mathcal{F}, P)$. Let us take first a $\mathcal{F}$-measurable function $g : \Omega \to \mathbb{R}$ that assumes a finite number of values, say $\{y_1, y_2, \dots, y_n\} \subset \mathbb{R}$, such that $A_i := g^{-1}(\{y_i\}) \in \mathcal{F}$ for each $i = 1, \dots, n$. This type of functions are known as **$\mathcal{F}$-simple functions**. Further, assume that $P(A_i) < \infty$ for each $i = 1, \dots, n$. Then, we say that function g is a ***P*-step function**. In this case, g admits the following **standard representation**:

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

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 $\omega \notin A_i$.

The **integral of the step function** g is then defined as

$$\int g(\omega) dP(\omega) = \int g(\omega) P(d\omega) := \sum_{i=1}^{n} y_i P(A_i).$$

The integration problem consists of enlarging this definition so that it may be applied to more general classes of functions, and not only to step functions.

So, let us take now a bounded $\mathcal{F}$-measurable function $f : \Omega \to \mathbb{R}$. We say that f is ***P*-integrable** (or ***P*-summable**) if

$$\sup \left\{ \int g(\omega) dP(\omega) : g \in L_P \text{ and } g \leq f \right\} = \inf \left\{ \int g(\omega) dP : g \in L_P \text{ and } f \leq g \right\},$$

where L_P is the space of P-step functions. The common value is called the **integral** of f with respect to P and is denoted either $\int f dP$, $\int f(\omega) dP(\omega)$, or $\int f(\omega) P(d\omega)$. There are several approaches to the abstract concept of integral. For example, the approach most closely related to the notion of measure is that of **Lebesgue integral** while the most used approach in Euclidean spaces, within the elementary calculus benchmark, is that of **Riemann integral**.

Let $(\Omega, \mathcal{F}, P)$ be a probability space and let X be a random variable on $(\Omega, \mathcal{F})$ with probability distribution ψ and distribution function F, both associated to the probability measure P. Throughout this course, we shall use the notation $dP(\omega)$, $dF(x)$, or $d\psi(x)$ interchangeably.

3.5. Continuous random variables

Definition 20. Let P and Q be two measures on a measurable space $(\Omega, \mathcal{F})$. Measure P is said to have **density** $f : \Omega \to \mathbb{R}$ with respect to measure Q if f is a nonnegative $\mathcal{F}$-measurable function and the following condition holds

$$P(A) = \int_A f dQ = \int_{\omega \in A} f(\omega) dQ(\omega) = \int_{\omega \in A} f(\omega) Q(d\omega) \quad \text{for each} \quad A \in \mathcal{F}.$$

Using the general definition above of a density of a measure, we can particularize it so as to introduce the concept of a density of a random variable, which is simply a density of the probability distribution of the random variable with respect to Lebesgue measure on the real line.

Definition 21. A random variable X on a probability space $(\Omega, \mathcal{F}, P)$, with probability distribution ψ, has **density function** $f : \mathbb{R} \to \mathbb{R}$ if ψ has density f with respect to Lebesgue measure on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$. That is, if f is nonnegative $\mathcal{B}_{\mathbb{R}}$-measurable and the following condition is satisfied

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

Definition 22. A random variable X on a probability space $(\Omega, \mathcal{F}, P)$ is **absolutely continuous** if it has a density function $f : \mathbb{R} \to \mathbb{R}$. That is, for each Borel set $B \in \mathcal{B}_{\mathbb{R}}$, we have

$$P[X \in B] = \int_{x \in B} f(x) dx.$$

Carathéodory's Theorem about the unique extension of a measure implies that if a random variable X with distribution function F is absolutely continuous then

$$F(b) - F(a) = \int_a^b f(x) dx$$

holds for every $a, b \in \mathbb{R}$ such that $a \leq b$. Therefore, we obtain that, if f is a density of the random variable X, then

$$F(b) - F(a) = P[a < X \leq b] = \int_a^b f(x)dx.$$

Also, f and F are related by

$$F(x) = P[X \leq x] = \int_{-\infty}^{x} f(t)dt.$$

Notice that $F'(x) = f(x)$ need not hold for each $x \in \mathbb{R}$; all that is required is that f integrates properly, as expressed in the definition of absolutely continuous random variable above. On the other hand, if $F'(x) = f(x)$ holds for each $x \in \mathbb{R}$ and f is continuous, then it follows from the fundamental theorem of calculus that f is indeed a density of F.[5] Nevertheless, if X is an absolutely continuous random variable with distribution function F, then F can be differentiated almost everywhere, and at each continuity point x of f, we have $F'(x) = f(x)$.

Discrete and continuous random variables allow for an analogous treatment. For a discrete random variable X on some probability space $(\Omega, \mathcal{F}, P)$ with support $\{x_1, x_2, \dots\}$, the function $f : \{x_1, x_2, \dots\} \to \mathbb{R}$ defined by $f(x_i) := P[X = x_i]$ is called the **discrete density function** of X. Now, compare how we compute probabilities in the discrete case

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

with the case in which X is an absolutely continuous random variable with density f:

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

Throughout the remaining of the course we shall use integrals when no possible confusion arises. In general, the particular analysis of discrete cases will only require change integrals to sums in the appropriate formulae.

Example 20. A box contains $a > 0$ red balls and $b > 0$ black balls. A random size of n balls is drawn from the box. Let X be the number of red balls picked. We would like to compute the density of X, considered as a random variable on some probability space, if the sampling is with replacement. To answer this, let us specify the set of balls as

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

and let us follow the convention that $\{1, \dots, a\}$ are red balls and $\{a+1, \dots, a+b\}$ are black balls. Then, since there is replacement, our sample set is $\Omega = S^n$ so that $|\Omega| = (a+b)^n$. The random variable X can be specified as

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

[5] We make no formal statements on the relation between differentiation and integration.

Also, the discrete density function of X is defined as $f(x) = P[X = x]$. Therefore, we need to compute the 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 sampling is with replacement, 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 known as the probability of success in a sequence of n Bernoulli trials. Using this probability of success, we can rewrite

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

which is known as the density function of a Binomial distribution with parameter p.

3.4. Functions of a random variable

Given a random variable X on a probability space $(\Omega, \mathcal{F}, P)$, a typical problem is probability theory is that of finding the density of the random variable $g(X)$. Since there is a measurability requirement in the definition of a random variable, one usually restricts attention to the case where g is a one-to-one function.

The treatment of this problem is relatively simple in the discrete case.

Example 21. Consider a discrete random variable X supported on $\{1, 2, \ldots, n\}$ and with discrete density function $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$ for some $p \in (0, 1)$. 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 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 $y \in \{a + b, a + 2b, \ldots, a + nb\}$.

Suppose now that X is absolutely continuous with distribution function F and density f. Furthermore, assume that f is continuous. Consider a one-to-one function $g : \mathbb{R} \to \mathbb{R}$

such that g^{-1} is differentiable on $\mathbb{R}$. We ask ourselves about the density of $Y = g(X)$. Let H and h denote, respectively, the distribution function and the density of the random variable $Y := g(X)$. Let $T := g^{-1}$ Recall that for an absolutely continuous random variable X with distribution function F and density f, if f is continuous, then $F'(x) = f(x)$ holds for each $x \in \mathbb{R}$.

First, suppose that g is increasing. Then, for $y \in \mathbb{R}$,

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

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

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

On the other hand, if g is decreasing, then

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

so that

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

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

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

The above arguments constitute an sketch for the proof of the following useful result.

Theorem 6.—Change of Variables— *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 $y \in V$. If X is a random variable with density f supported in U, then the random variable $Y = g(X)$ has density h supported in V and given by*

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

Example 22. Consider a positive absolutely continuous random variable X with continuous density f. 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 $h(y) = f(1/y)/y^2$.

We can use arguments similar to those above to obtain the density of a transformation $Y = g(X)$ even when g is not one-to-one.

Example 23. Suppose that X is an absolutely continuous random variable with density

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

which 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. Here notice that the transformation $g(X) = X^2$ is not one-to-one. However, note that

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

since the graph of f is symmetric around the origin. Now, since $\sqrt{y} = x$ and $dx = 1/2y^{-1/2}dy$, we have

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

Since $H(x) = \int_0^x h(y)dy$ and $\text{supp}(Y) = \mathbb{R}_+$, we obtain that $Y = X^2$ has density

$$h(y) = \begin{cases} 0 & \text{if } y \leq 0, \\ \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-y/2} & \text{if } y > 0, \end{cases}$$

which corresponds to a distribution known as chi-square with parameter 1.

Problems

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

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

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

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

5. 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.

6. 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.

7. 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) $P[X \geq 3 | X > 0]$.

8. 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.

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

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

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

11. 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.

12. 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 $F(m) = 1/2$.

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

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

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

14. 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 $F'(x) = f(x)$?

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

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

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

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

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

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

Random vectors

4.1. Definitions

Many random phenomena have multiple "aspects" or "dimensions." To study such phenomena we need to consider random vectors.

Definition 23. Let $(\Omega, \mathcal{F})$ be a measurable space, an n-dimensional **random vector** on $(\Omega, \mathcal{F})$ is a vector-valued $\mathcal{F}$-measurable function $X : \Omega \to \mathbb{R}^n$.

It can be shown that a vector-valued function $X = (X_1, \ldots, X_n)$ is a random vector on some measurable space $(\Omega, \mathcal{F})$ if and only if each X_i, $i = 1, \ldots, n$, is a random variable on $(\Omega, \mathcal{F})$. Therefore, a random vector is simply an n-tuple $X = (X_1, \ldots, X_n)$ of random variables.

Definition 24. Let $X = (X_1, \ldots, X_n)$ be a random vector on some probability space $(\Omega, \mathcal{F}, P)$. The **joint probability distribution** of the random vector X is the probability measure ψ on $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ defined by

$$\psi(B) := P\left(\left\{\omega \in \Omega : \big(X_1(\omega), \ldots, X_n(\omega)\big) \in B\right\}\right) = P[(X_1, \ldots, X_n) \in B],$$

for each $B \in \mathcal{B}_{\mathbb{R}^n}$.

Moreover, the **joint distribution function** of the random vector X is the function $F : \mathbb{R}^n \to \mathbb{R}$ defined by[6]

$$\begin{aligned} F(x_1, \ldots, x_n) := \psi(S_x) &= P\left(\left\{\omega \in \Omega : \big(X_1(\omega), \ldots, X_n(\omega)\big) \leq (x_1, \ldots, x_n)\right\}\right) \\ &= P[X_1 \leq x_1, \ldots, X_n \leq x_n], \end{aligned}$$

where $S_x := \{y \in \mathbb{R}^n : y_i \leq x_i,\ i = 1, \ldots, n\}$ is the set of points "southwest" of x.

Definition 25. A random vector $X = (X_1, \ldots, X_n)$ on a probability space $(\Omega, \mathcal{F}, P)$ with probability distribution ψ is **discrete** if ψ has a countable support. In this case, ψ is completely determined by $\psi(\{x\}) = P[X_1 = x_1, \ldots, X_n = x_n]$ for $x \in \text{supp}(X) \subseteq \mathbb{R}^n$. If a random vector is discrete, then we say that its joint probability distribution and its joint distribution function are discrete as well.

Definition 26. Let $X = (X_1, \ldots, X_n)$ be a discrete random vector on some probability space $(\Omega, \mathcal{F}, P)$. The vector-valued function $f : \text{supp}(X) \to \mathbb{R}$ defined by $f(x_1, \ldots, x_n) := P[X_1 = x_1, \ldots, X_n = x_n]$ is called the **discrete joint density function** of X.

[6]For $x, y \in \mathbb{R}^n$, notation $x \leq y$ means $x_i \leq y_i$ for each $i = 1, \ldots, n$.

Definition 27. A random vector $X = (X_1, \ldots, X_n)$ on a probability space $(\Omega, \mathcal{F}, P)$ with probability distribution ψ is **absolutely continuous** if it has a density $f : \mathbb{R}^n \to \mathbb{R}$ with respect to Lebesgue measure λ on $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ such that for each Borel set $B \in \mathcal{B}_{\mathbb{R}^n}$,

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

The density f is called **joint density function** of X.

If X is an n-dimensional random vector (on some probability space) with probability distribution ψ, joint distribution function F, and joint density f (which may be discrete as well) and $g_i : \mathbb{R}^n \to \mathbb{R}$ is the function defined by $g_i(x_1, \ldots, x_n) := x_i$, then $g_i(X)$ is the random variable X_i with (a) **marginal probability distribution** ψ_i given by

$$\psi_i(B) = \psi(\{(x_1, \ldots, x_n) \in \mathbb{R}^n : x_i \in B\}) = P[X_i \in B] \text{ for each } B \in \mathcal{B}_{\mathbb{R}},$$

(b) **marginal distribution function** given by

$$F_i(x_i) = P[X_i \leq x_i],$$

and (c) **marginal density** given by

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

in the absolutely continuous case, or

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

in the discrete case.

4.2. Functions of random vectors

A change of variables for a discrete random vector is done straightforwardly just as shown for the case of a discrete random variable. Therefore, we turn to analyze the case of absolutely continuous random vectors.

Let $g : U \to V$ be a one-to-one continuously differentiable function, where $U, V \subseteq \mathbb{R}^n$ are open sets. We begin with n random variables $X_1, \ldots, X_n$, with joint density function f, transform them into "new" random variables $Y_1, \ldots, Y_n$ through the functions

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

and ask about the joint density function of $Y_1, \ldots, Y_n$. Let $T := g^{-1}$ and suppose that its Jacobian never vanishes, i.e.,

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

Under these conditions we can state the following useful result.

Theorem 7.—Change of Variables— *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*

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

Example 24. Let (X_1, X_2) an absolutely continuous random vector with joint density function

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

Consider the transformation given by

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

We are asked to find the joint density function of (Y_1, Y_2). To answer 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 the theorem above, one obtains

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

and, consequently,

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

4.3. Independent random variables

Analogously to the analysis of independent random events, an interesting case appears when the random phenomenon (or dimension of a random phenomenon) described by a random variable occurs regardless that captured by another random variable.

Definition 28. The random variables $X_1, \ldots, X_n$ (on some probability space $(\Omega, \mathcal{F}, P)$) are **independent random variables** if

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

for all Borel sets $B_1, \ldots, B_n$ with the form $B_i = (a_i, b_i]$, where $a_i < b_i$, $i = 1, \ldots, n$.

Let $(X_1, \ldots, X_n)$ be a random vector (on some probability space) with joint distribution function F and joint density f (corresponding either to the absolutely continuous or the discrete case). It can be shown that the requirement in the definition of independence of random variables above is equivalent to the condition

$$P[X_1 \leq x_1, \ldots, X_n \leq x_n] = P[X_1 \leq x_1] \cdots P[X_n \leq x_n] \text{ for each } (x_1, \ldots, x_n) \in \mathbb{R}^n.$$

Then, using the definition of distribution function, one obtains that $X_1, \ldots, X_n$ are independent if and only if

$$F(x_1, \ldots, x_n) = F_1(x_1) \cdots F_n(x_n).$$

Furthermore, by Fubini's theorem, we also obtain that $X_1, \ldots, X_n$ are independent if and only if

$$f(x_1, \ldots, x_n) = f_1(x_1) \cdots f_n(x_n).$$

Example 25. Let X_1, X_2, X_3 be independent absolutely continuous random variables with common density

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

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

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

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

4.4. Probability distribution of independent random vectors

The theory of independent random variables can be extended readily to random vectors. If X_i is a k_i-dimensional random vector, then $X_1, \ldots, X_n$ are independent random vectors if the earlier definition of independence holds with the appropriate changes in the formula

so as to consider random vectors instead of random variables. In particular, we say that $X_1, \ldots, X_n$ are **independent random vectors** if

$$P[X_1 \leq x_1, \ldots, X_n \leq x_n] = P[X_1 \leq x_1] \cdots P[X_n \leq x_n] \text{ for each } x_1 \in \mathbb{R}^{k_1}, \ldots, x_n \in \mathbb{R}^{k_n}.$$

Let us analyze in more detail the probability distribution of two independent random vectors (or variables). Let X and Y be independent random vectors (on some probability space $(\Omega, \mathcal{F}, P)$) with distributions ψ_x and ψ_y in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. Then, (X, Y) has probability distribution $\psi_x \times \psi_y$ in $\mathbb{R}^{n+m}$ given by

$$(\psi_x \times \psi_y)(B) = \int_{\mathbb{R}^n} \psi_y(\{y : (x, y) \in B\})\psi_x(dx) \quad \text{for each} \quad B \in \mathcal{B}_{\mathbb{R}^{n+m}}.$$

The following result allows us to use the probability distribution of two independent random vectors (or variables) to compute probabilities.

Theorem 8. *Let X and Y be independent random vectors (on some probability space $(\Omega, \mathcal{F}, P)$) with distributions ψ_x and ψ_y in $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. Then,*

$$P[(X, Y) \in B] = \int_{\mathbb{R}^n} P[(x, Y) \in B]\psi_x(dx) \quad \textit{for each} \quad B \in \mathcal{B}_{\mathbb{R}^{n+m}},$$

and

$$P[X \in A, (X, Y) \in B] = \int_A P[(x, Y) \in B]\psi_x(dx) \quad \textit{for each } A \in \mathcal{B}_{\mathbb{R}^n} \textit{ and } B \in \mathcal{B}_{\mathbb{R}^{n+m}}.$$

Now, we can use the result above to propose an alternative approach to obtain the distribution of a particular function of random variables, namely, the sum of independent random variables.

Let X and Y be two independent random variables with respective distributions ψ_x and ψ_y. Using the first result in the Theorem above, we obtain

$$P[X + Y \in B] = \int_{\mathbb{R}} \psi_y(B \setminus \{x\})\psi_x(dx) = \int_{\mathbb{R}} P[Y \in B \setminus \{x\}]\psi_x(dx).$$

The **convolution** of ψ_x and ψ_y is the measure $\psi_x * \psi_y$ defined by

$$(\psi_x * \psi_y)(B) := \int_{\mathbb{R}} \psi_y(B \setminus \{x\})\psi_x(dx) \qquad \text{for each} \quad B \in \mathcal{B}_{\mathbb{R}}.$$

That is, $X + Y$ has probability distribution $\psi_x * \psi_y$. Now, if F and G are the distribution functions corresponding to ψ_x and ψ_y, then $X + Y$ has distribution function $F * G$. Taking $B = (-\infty, y]$, the definition of convolution above gives us

$$(F * G)(y) = \int_{\mathbb{R}} G(y - x)dF(x).$$

Furthermore, if ψ_x and ψ_y have respective density functions f and g, then $Z = X + Y$ has density $f * g$, which, in turn, gives us the density h_{X+Y} of the sum of two independent random variables,

$$h_{X+Y}(z) = (f * g)(z) = \int_{\mathbb{R}} g(z - x) f(x) dx.$$

These arguments can be easily generalized to obtain the distribution of the sum of an arbitrary finite number n of independent random variables.

4.5. Covariance and correlation

Definition 29. Let X_1 and X_2 be two random variables (on a probability space $(\Omega, \mathcal{F}, P)$) with joint probability distribution ψ and joint distribution function F. The **covariance** of X_1 and X_2 is

$$\mathrm{Cov}[X_1, X_2] := \int_{\mathbb{R}} \int_{\mathbb{R}} (x_1 - \mu_1)(x_2 - \mu_2) dF(x_1, x_2),$$

where

$$\mu_i := \int_{\mathbb{R}} x_i dF_i(x_i) \quad \text{for each} \quad i = 1, 2.$$

If $\mathrm{Cov}[X_1, X_2] = 0$, then we say that the random variables X_1 and X_2 are **uncorrelated**. Moreover, the **correlation coefficient** between X_1 and X_2 is the number

$$\rho[X_1, X_2] = \frac{\mathrm{Cov}[X_1, X_2]}{\left(\int_{\mathbb{R}} (x_1 - \mu_1)^2 dF_1(x_1) \int_{\mathbb{R}} (x_2 - \mu_2)^2 dF_2(x_2)\right)^{1/2}}.$$

Obviously, if two random variables X_1 and X_2 are uncorrelated, then $\rho[X_1, X_2] = 0$. Also, if $\int_{\mathbb{R}} (x_i - \mu_i)^2 dF_i(x_i) < \infty$ for $i = 1, 2$, then $\rho[X_1, X_2] = 0$ implies that X_1 and X_2 are uncorrelated.

Let us now study the relation between independence and no correlation of two random variables. Consider two random variables X_1 and X_2 with joint distribution function F. Then, applying the definition of covariance above, one obtains

$$\begin{aligned}
\mathrm{Cov}[X_1, X_2] &= \int_{\mathbb{R}} \int_{\mathbb{R}} (x_1 - \mu_1)(x_2 - \mu_2) dF(x_1, x_2) \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} x_1 x_2 dF(x_1, x_2) - \mu_1 \int_{\mathbb{R}} x_2 dF_2(x_2) - \mu_2 \int_{\mathbb{R}} x_1 dF_1(x_1) + \mu_1 \mu_2 \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} x_1 x_2 dF(x_1, x_2) - \mu_1 \mu_2.
\end{aligned}$$

Now, suppose that the random variables X_1 and X_2 are independent. Then, since in that case $F(x_1, x_2) = F_1(x_1) F_2(x_2)$, the expression above becomes

$$\int_{\mathbb{R}} \int_{\mathbb{R}} x_1 x_2 dF(x_1, x_2) - \mu_1 \mu_2 = \mu_1 \mu_2 - \mu_1 \mu_2 = 0.$$

Therefore, if two random variables are independent, then they are uncorrelated too. However, two uncorrelated random variables need not be independent.

Problems

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

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

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

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

and $g(\mathbb{R}_{++}) = \{(y_1, y_2) \in \mathbb{R}^2 : y_1 > 0,\, 0 < y_2 < 2\pi\}$. 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) $h(y_1, y_2) = (2\pi)^{-1} y_1 e^{-y_1^2/2}$,
(b) $h_1(y_1) = y_1 e^{-y_1^2/2}$,
(c) $h_2(y_2) = (2\pi)^{-1}$.

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

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

and

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

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

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

supported in $\mathbb{R}$.

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

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

5. Let X and Y be two independent absolutely continuous random variables (on some probability space $(\Omega, \mathcal{F}, P)$) having the same density each, $f(x) = g(y) = 1$ for $x, y \in (0, 1]$. Find
(a) $P[|X - Y| \leq .5]$;
(b) $P\left[\left|\frac{X}{Y} - 1\right| \leq .5\right]$;
(c) $P[Y \geq X | Y \geq 1/3]$.

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

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

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

7. 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.

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

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

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

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

Moments and expected value of a distribution

5.1. Definition of moments of a distribution

The information contained in a probability distribution can be often summarized by some characteristics of the general shape of the distribution and its location. Such characteristics are in most cases described by numbers known as the moments of the distribution.

Definition 30. Let X be a random variable with distribution function F. Suppose that, given a positive integer r and a real number k, $\int_{\mathbb{R}} |(x-k)|^r \, dF(x) < \infty$. Then, the ***r*th moment about the point *k*** of the random variable is

$$m_{r,k}(X) := \int_{\mathbb{R}} (x-k)^r dF(x).$$

Moreover, if $k = 0$, then the moment $m_{r,0}(X)$ is called the ***r*th moment about the origin**. We usually write $m_r(X)$ instead of $m_{r,0}(X)$.

5.2. Definition of expected value

The concept of expected value of a random variable can be interpreted as the every-day notion of mean value, weighted arithmetic sum or arithmetic average, according to the frequency interpretation of probabilities.

Definition 31. Let X be a random variable on a probability space $(\Omega, \mathcal{F}, P)$. The **expected value** (or **expectation**) of X on $(\Omega, \mathcal{F}, P)$ is the integral of X with respect to measure P:

$$E[X] = \int_{\Omega} X dP = \int_{\omega \in \Omega} X(\omega) dP(\omega).$$

Also, suppose that the random variable X on $(\Omega, \mathcal{F}, P)$ has probability distribution ψ and distribution function F. Then, the expected value of X can accordingly be specified as

$$E[X] = \int_{\mathbb{R}} x d\psi = \int_{x \in \mathbb{R}} x dF(x).$$

Furthermore, if X is a discrete random variable with support $\{x_1, x_2, \dots\}$ and discrete density function f, we have

$$E[X] = \sum_{i=1}^{\infty} x_i P[X = x_i] = \sum_{i=1}^{\infty} x_i f(x_i).$$

On the other hand, if X is an absolutely continuous random variable with density f, then

$$E[X] = \int_{\mathbb{R}} xf(x)dx.$$

As in the definition of covariance between two random variables, we shall often use notation $\mu = E[X]$ when no possible confusion arises.

Notice that the expected value of a random variable $E[X]$ coincides with its first moment (about the origin), $m_1(X)$.

Given a random variable X with distribution function F and a function $g : \mathbb{R} \to \mathbb{R}$, we have

$$E[g(X)] = \int_{\mathbb{R}} g(x)dF(x).$$

Finally, consider a random variable X on a probability space $(\Omega, \mathcal{F}, P)$ distribution function F. If the first two moments of X (about the origin), m_1 and m_2, are finite, then we can define the **variance** of X as

$$\mathrm{Var}[X] := E[(X-\mu)^2] = \int_{\Omega} \Big(X(\omega) - \mu\Big)^2 dP(\omega) = \int_{\mathbb{R}} (x-\mu)^2 dF(x),$$

which coincides with the second moment of X about the point μ, $m_{2,\mu}(X)$. Intuitively, the variance of a random variable is a measure of the degree of dispersion of the corresponding distribution with respect to its expected value.

5.3. Properties of expected values

The following properties, which can be immediately derived from analogous properties of the integral operator, will prove useful in the remaining of the course. We shall not demonstrate formally the required properties of the integral. Consider two random variables X and Y whose expected values are well defined on some probability space and two real numbers α and β. Then, the following properties hold:

(P1) $E[\alpha X] = \alpha E[X]$;
(P2) $E[\alpha] = \alpha$;
(P3) $E[\alpha X + \beta Y] = \alpha E[X] + \beta E[Y]$;
(P4) if $X \leq Y$ almost everywhere, then $E[X] \leq E[Y]$.

As for the variance of a random variable X, using some of the properties above, we can obtain the following useful result.

$$\begin{aligned}
\mathrm{Var}[X] &= E[(X-\mu)^2] \\
&= E[X^2 + \mu^2 - 2X\mu] \\
&= E[X^2] + \mu^2 - 2\mu E[X] \\
&= E[X^2] + \mu^2 - 2\mu^2 \\
&= E[X^2] - \mu^2 = m_2(X) - m_1^2(X).
\end{aligned}$$

Finally, we can also use the moments of a distribution to obtain and expression for the covariance between two random variables. Consider two random variables X_1 and X_2 with joint distribution function F. In the previous chapter, we showed that

$$\text{Cov}[X_1, X_2] = \int_{\mathbb{R}} \int_{\mathbb{R}} x_1 x_2 dF(x_1, x_2) - \mu_1 \mu_2.$$

So, if we let $g : \mathbb{R}^2 \to \mathbb{R}$ be the function $g(X_1, X_2) = X_1 X_2$, then we have

$$\int_{\mathbb{R}} \int_{\mathbb{R}} x_1 x_2 dF(x_1, x_2) = E[X_1 X_2].$$

Therefore, we can write

$$\text{Cov}[X_1, X_2] = m_1(X_1 X_2) - m_1(X_1) m_1(X_2),$$

where $m_1(X_1 X_2)$ indicates the moment of order 1 of the random variable $X_1 X_2$.

5.4. Conditional expected value

In many applications we are interested in deriving the expected value of a random variable given that there is some information available about another random variable. Consider two random variables, X and Y, on a probability space $(\Omega, \mathcal{F}, P)$. We would like to answer questions of the form: "what is the expected value of the random variable X given that the realization of another random variable Y is y (i.e., $Y(\omega) = y$)?" We begin by introducing the theory of conditional expected value from first principles and then relate this concept to that of conditional distribution.

Consider a probability space $(\Omega, \mathcal{F}, P)$ and another σ-algebra $\mathcal{G} \in \mathcal{F}$ on Ω. We start by facing the question "what is the expected value of a random variable X on $(\Omega, \mathcal{F}, P)$ given that we know for each $B \in \mathcal{G}$ whether or not $\omega \in B$."

Definition 32. Let $(\Omega, \mathcal{F}, P)$ be a probability space and X a P-integrable random variable on $(\Omega, \mathcal{F}, P)$. Consider a σ-algebra $\mathcal{G} \subset \mathcal{F}$ on Ω. Then, there exists a random variable $E[X|\mathcal{G}]$, called the **conditional expected value of X given $\mathcal{G}$**, which satisfies:
(1) $E[X|\mathcal{G}]$ is $\mathcal{G}$-measurable and P-integrable;
(2) $E[X|\mathcal{G}]$ meets the functional equation

$$\int_B E[X|\mathcal{G}](\omega) dP(\omega) = \int_B X(\omega) dP(\omega) \quad \text{for each} \quad B \in \mathcal{G}.$$

We shall not prove the existence of such random variable specified in the definition above. There will be in general many such random variables $E[X|\mathcal{G}]$, any of which is called a **version** of the conditional expected value. However, any two versions are equal almost everywhere. For a given $\omega \in \Omega$, the value $E[X|\mathcal{G}](\omega)$ is intuitively interpreted as the expected value of X for an observer who knows for each $B \in \mathcal{G}$ whether or not $\omega \in B$ (ω itself in general remains unknown).

Example 26. Let $B_1, B_2, \dots$ be either a finite or a countable partition of Ω that generates the σ-algebra $\mathcal{G}$ (i.e., $\mathcal{G} = \sigma(\{B_1, B_2, \dots\})$). First, notice that since $E[X|\mathcal{G}] : \Omega \to \mathbb{R}$ is $\mathcal{G}$-measurable, it must be the case that $E[X|\mathcal{G}](\omega) = k_i$ for each $\omega \in B_i$, for each $i = 1, 2, \dots$, where each k_i is a constant. Then, by applying condition (2) of the definition of conditional expectation above when $B = B_i$, one obtains, for each $\omega \in B_i$:

$$\begin{aligned}\int_{B_i} k_i dP(\omega) &= \int_{B_i} X(\omega) dP(\omega) \\ &\Leftrightarrow E[X|\mathcal{G}](\omega) P(B_i) = \int_{B_i} X(\omega) dP(\omega) \\ &\Leftrightarrow E[X|\mathcal{G}](\omega) = \frac{1}{P(B_i)} \int_{B_i} X(\omega) dP(\omega),\end{aligned}$$

which is specified in this way for $P(B_i) > 0$.

For a random variable X on a probability space $(\Omega, \mathcal{F}, P)$, the **conditional probability** of the event $\{X \in B\}$, for $B \in \mathcal{B}_{\mathbb{R}}$, given a σ-algebra $\mathcal{G} \subset \mathcal{F}$ on Ω is a $\mathcal{G}$-measurable, P-integrable random variable $P[X \in B|\mathcal{G}] : \Omega \to \mathbb{R}$ on $(\Omega, \mathcal{F}, P)$ satisfying

$$\int_C P[X \in B|\mathcal{G}](\omega) dP(\omega) = P\big[\{X \in B\} \cap C\big], \quad \text{for each } \ C \in \mathcal{G}.$$

Intuitively, $P[X \in B|\mathcal{G}]$ gives us the probability of the event $\{X \in B\}$ when the observer knows for each $C \in \mathcal{G}$ whether or not $\omega \in C$. The existence of $P[X \in B|\mathcal{G}]$ is guaranteed if X is P-integrable. There will be in general many such random variables $P[X \in B|\mathcal{G}]$ but any two of them are equal almost everywhere.

Example 27. As in the previous example, let $C_1, C_2, \dots$ be either a finite or a countable partition of Ω that generates the σ-algebra $\mathcal{G}$ (i.e., $\mathcal{G} = \sigma(\{C_1, C_2, \dots\})$). Since $P[X \in B|\mathcal{G}]$ is $\mathcal{G}$-measurable, it must be the case that $P[X \in B|\mathcal{G}](\omega) = \alpha_i$ for each $\omega \in C_i$, for each $i = 1, 2, \dots$, where each α_i is a constant. Then, by applying the definition of conditional probability above when $C = C_i$, one obtains, for each $\omega \in C_i$:

$$\alpha_i P(C_i) = P\big[\{X \in B\} \cap C\big] \Leftrightarrow P[X \in B|\mathcal{G}](\omega) = \frac{P\big[\{X \in B\} \cap C_i\big]}{P(C_i)}$$

whenever $P(C_i) > 0$.

Next we introduce the notion of conditional probability distribution. We do so by means of the following result.

Theorem 9. *Let X be a random variable on a probability space $(\Omega, \mathcal{F}, P)$ and let $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra on Ω. Then, there exists a function $\delta : \mathcal{B}_{\mathbb{R}} \times \Omega \to \mathbb{R}$ satisfying:*
(1) for each $\omega \in \Omega$, $\delta(\cdot, \omega)$ is a probability measure on $\mathcal{B}_{\mathbb{R}}$;
(2) for each $B \in \mathcal{B}_{\mathbb{R}}$, $\delta(B, \cdot)$ is a version of $P[X \in B|\mathcal{G}]$.

The probability measure $\delta(\cdot, \omega)$ identified in the Theorem above is known as a **conditional probability distribution** of X given $\mathcal{G}$.

As for the relation between a conditional probability distribution and a conditional expected value of random variable, the following result provides the required insight.

Theorem 10. *Let X be a random variable on a probability space $(\Omega, \mathcal{F}, P)$ and, for $\omega \in \Omega$, let $\delta(\cdot, \omega)$ be a conditional probability distribution of X given a σ-algebra $\mathcal{G} \subset \mathcal{F}$ on Ω. Then $\int_{\mathbb{R}} x\delta(dx, \omega)$ is a version of $E[X|\mathcal{G}](\omega)$ for $\omega \in \Omega$.*

Now we can use this theory to obtain an analogous framework for the density of a random vector. By doing so we are able to answer the question raised at the beginning of the section. Consider a random vector (X_1, X_2) on some probability space with joint density f and respective marginal densities f_1 and f_2. Then, the **conditional density** of X given $Y = y$ is the density defined by

$$f(x|y) = \frac{f(x, y)}{f_2(y)}.$$

Now, using the Theorem above, it can be shown that the conditional expected value of X given $Y = y$ can be written as

$$E[X|Y = y] =_{\text{a.e.}} \int_{\mathbb{R}} x f(x|y) dx.$$

Example 28. Let X and Y be two absolutely continuous random variables with joint density

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

We are asked to compute the conditional density and conditional expected value of Y given $X = x$.

First, the marginal density of X is given by

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

if $0 \leq x \leq 1$ and $f_1(x) = 0$ otherwise. Therefore, for $0 \leq x \leq 1$,

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

Then, we can compute

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

To compute the integral above, note that

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

So, using that algebraical identity, we have

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

5.5. Moment generating function

Even though the definition of moment gives us a closed expression, from a practical viewpoint, sometimes it may be complicated (if not impossible) to compute the required integral. Therefore, we outline another method for deriving the moments for most distributions.

Definition 33. Let X be a random variable (on a probability space $(\Omega, \mathcal{F}, P)$) with distribution function F. The **moment generating function** of X is a real-valued function $\Phi_X : \mathbb{R} \to \mathbb{R}$ defined by

$$\Phi_X(t) := E\left[e^{tX}\right] = \int_{\mathbb{R}} e^{tx}dF(x)$$

for each $t \in \mathbb{R}$ for which $\Phi_X(t)$ is finite.

We now make use of the theory of Taylor expansions. In particular, let us first invoke the expansion result

$$e^s = 1 + s + \frac{s^2}{2!} + \frac{s^3}{3!} + \cdots = \sum_{i=0}^{\infty}\frac{s^i}{i!}, \qquad s \in \mathbb{R}.$$

Now, suppose that $\Phi_X(t)$ is well defined throughout the interval $(-\bar{t}, \bar{t})$ for some $\bar{t} > 0$. Then, since $e^{|tx|} \leq e^{tx} + e^{-tx}$ and $\int_{\mathbb{R}}[e^{tx} + e^{-tx}]dF(x)$ is finite for $|t| < \bar{t}$, it follows that

$$\int_{\mathbb{R}} e^{|tx|}dF(x) = \sum_{i=0}^{\infty} \int_{\mathbb{R}} \frac{|tx|^i}{i!}dF(x) < \infty.$$

Therefore, one obtains

$$\begin{aligned}\Phi_X(t) &= \int_{\mathbb{R}} e^{tx}dF(x) = \sum_{i=0}^{\infty} \frac{t^i}{t!} \int_{\mathbb{R}} x^i dF(x) \\ &= 1 + t\int_{\mathbb{R}} xdF(x) + \frac{t^2}{2!}\int_{\mathbb{R}} x^2 dF(x) + \frac{t^3}{3!}\int_{\mathbb{R}} x^3 dF(x) + \ldots \quad \text{for} \quad |t| < \bar{t}.\end{aligned}$$

Finally, using the theory of Taylor expansions, it follows that if the rth derivate of $\Phi_X(t)$, $\Phi_X^{(r)}(t)$, exists in some neighborhood of $t = 0$, then

$$\Phi_X^{(r)}(0) = \int_{\mathbb{R}} x^r dF(x) = m_r(X).$$

The definition of moment generating function can be extended readily to random vectors.

Definition 34. Let $X = (X_1, \ldots, X_n)$ be a random vector (on a probability space $(\Omega, \mathcal{F}, P)$) with joint distribution function F. The **moment generating function** of X is a vector-valued function $\Phi_X : \mathbb{R}^n \to \mathbb{R}$ defined by

$$\Phi_X(t_1, \ldots, t_n) := \int_{\mathbb{R}^n} \cdots \int e^{(t_1x_1 + \cdots + t_nx_n)}dF(x_1, \ldots, x_n)$$

for each $(t_1, \ldots, t_n) \in \mathbb{R}^n$ for which $\Phi_X(t_1, \ldots, t_n)$ is finite.

Then, using the theory of Taylor expansions, as in the case of a random variable, one obtains

$$\begin{aligned}\frac{\partial^r \Phi_X(0, \ldots, 0)}{\partial t_k^r} &= \int_{\mathbb{R}^n} \cdots \int x_k^r dF(x_1, \ldots, x_n) \\ &= \int_{\mathbb{R}} x_k^r \int_{\mathbb{R}^{n-1}} \cdots \int dF(x_1, \ldots, x_n) \\ &= \int_{\mathbb{R}} x_k^r dF_k(x_k) = m_r(X_k).\end{aligned}$$

The moment generating function of a random variable or of a random vector has several important applications in the study of the underlying probability distributions. In particular, the moment generating function of a random vector can be used to the study whether a set of random variables are independent or not.

Theorem 11. *Let (X, Y) be a random vector with moment generating function $\Phi_{(X,Y)}(t_1, t_2)$. Then, the random variables X and Y are independent if and only if*

$$\Phi_{(X,Y)}(t_1, t_2) = \Phi_X(t_1)\Phi_Y(t_2)$$

for each $t_1, t_2 \in \mathbb{R}$.

Also, the moment generating function of a random variable can be used to characterize the distribution of the random variable itself. Using the following result, we can identify a distribution just by identifying its moment generating function.

Theorem 12. *The moment generating function uniquely determines a probability distribution and, conversely, if the moment generating function of a distribution exists, then it is unique.*

The uniqueness property of the theorem above can be used to find the probability distribution of a transformation $Y = g(X_1, \ldots, X_n)$.

Example 29. Let X be an absolutely continuous random variable with density function

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

and consider the transformation $Y = X^2$. Then, we have

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

To compute the integral above, let us propose the following change of variables

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

Then, we obtain

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

Thus, by making use of the identity $\int_{-\infty}^{+\infty} e^{-z^2} dz = \sqrt{\pi}$, known as Gaussian integral identity, we have

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

Indeed, this particular moment generating function corresponds to a continuous random variable with density function

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

Recall that we already obtained this result in Example 23.

Example 30. Let $X_1, X_2, \ldots, X_k$ be a set of discrete random variables with common (discrete) density function

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

where $n \geq 1$ is an integer and $p \in (0, 1)$. In the next chapter we will show 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 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 that the random variable X has density function

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

Problems

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

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

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

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

where

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

Show that

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

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

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

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

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 $E[X|Y = y]$.

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

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

7. 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 $\Phi''_X(t) = E[x(x-1)t^{x-2}]$.

8. Let X and Y be two random variables (on some probability space $(\Omega, \mathcal{F}, P)$) such that

$$P[|X - Y| \leq a] = 1$$

for some constant $a \in \mathbb{R}$. Show that if Y is P-integrable, then X is P-integrable too and

$$|E[X] - E[Y]| \leq a.$$

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

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

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

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

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

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

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

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

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

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

Some special distributions

6.1. Some discrete distributions

We begin with the binomial distribution.

6.1.1. *The binomial distribution*

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

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

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

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

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

Then,

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

and

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

It follows that

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

and

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

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

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

One can easily compute

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

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

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

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

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

and

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

Theorem 13. *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 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 result has been demonstrated in Example 30.

6.1.2. *The negative binomial distribution*

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

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

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

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

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

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

6.1.3. *The multinomial distribution*

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

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

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

6.1.4. *The Poisson distribution*

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

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

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

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

One can check that

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

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

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

and

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

Theorem 14. *Let $X_i \sim P(r_i)$, $i = 1, \ldots, k$, be independent random variables. Then,*

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

The following results relate the Poisson with the binomial distribution.

Theorem 15. *Let $X \sim P(r_x)$ and $Y \sim P(r_y)$ be independent random variables. Then the conditional distribution of X given $X + Y$ is binomial. In particular, $(X|X + Y = n) \sim b(n, \frac{r_x}{r_x + r_y})$ (that is, for a sequence of n Bernoulli trials). Conversely, let X and Y are independent nonnegative integer-valued random variables with strictly positive densities. If $(X|X + Y = n) \sim b(n, p)$, then $X \sim P(\theta p/(1 - p))$ and $Y \sim P(\theta)$ for an arbitrary $\theta > 0$.*

Theorem 16. *If $X \sim P(r)$ and $(Y|X = x) \sim b(x, p)$, then $Y \sim P(rp)$.*

6.2. Some continuous distributions

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

6.2.1. *The uniform distribution*

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

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

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

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

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

Example 31. 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

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

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

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

for $0 \le y < 1$.

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

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

It is a well known result that the integral

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

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

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

Thus, if α is a positive integer, then

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

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

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

Therefore, we obtain

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

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

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

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

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

The gamma distribution is often used to model waiting times.

The distribution function associated to a gamma distribution is

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

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

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

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

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

we obtain

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

Therefore, for the gamma distribution, we obtain

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

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 an absolutely continuous random variable X with density

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

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

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

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

Theorem 18. *Let $X_i \sim \Gamma(\alpha_i, \beta)$, $i = 1, \ldots, k$, be independent random variables. Then, $Y_k = \sum_{i=1}^k X_i \sim \Gamma(\sum_{i=1}^k \alpha_i, \beta)$.*

Theorem 19. *Let $X \sim U[0,1]$. Then, $Y = -2\ln X \sim \chi^2(2)$.*

Theorem 20. *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 21. *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of independent random variables such that $X_n \sim exp(\beta)$ for each $n = 1, 2, \ldots$. Let $Y_n = \sum_{i=1}^{n} X_i$ for $n = 1, 2, \ldots$ and let Z be the random variable corresponding to the number of $Y_n \in [0,t]$ for $t > 0$. Then $Z \sim P(t/\beta)$.*

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

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

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

6.2.3. *The normal distribution*

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

Definition 35. A random variable X is said to have the **normal distribution** if its density function is given by

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

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

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

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

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

Theorem 23. *A random variable X has a $N(\mu, \sigma^2)$ distribution if and only if the random variable $(X-\mu)/\sigma$ has a $N(0,1)$ distribution.*

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

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

Second, we compute the integral above as

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

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

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

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

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

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

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

Theorem 24. *Let X be a standard normal random variable. Then,*

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

Theorem 25. *If X and Y are independent normally distributed random variables, then $X + Y$ and $X - Y$ are independent.*

Theorem 26. *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*

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

Theorem 27. *If $X \sim N(\mu, \sigma^2)$, then $(X - \mu)^2/\sigma^2 \sim \chi^2(1)$.*

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

6.2.4. *The multivariate normal distribution*

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

Definition 36. 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.[7]

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.

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

Theorem 28. *Let $X \sim N(\mu, \Sigma)$ such that X, μ, and Σ can be partitioned as*

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

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

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

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

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

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

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

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

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

Example 32. Suppose that X has a normal distribution with mean 0 and variance 1. Let W be a random variable which takes the values either 1 or -1, each with probability $1/2$, and assume W is independent of X. Now, let $Y = WX$. Then, it can be checked that

(i) X and Y are uncorrelated,

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

(iii) X and Y are not independent.

To see that X and Y are uncorrelated, notice that

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

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

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

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

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

Theorem 30. *Let $X \sim N(\mu, \Sigma)$, and let $A \in \mathbb{R}^m \times \mathbb{R}^n$ and $b \in \mathbb{R}^m$. Then,*

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

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

Theorem 31. *Let $X \sim N(\mu, \Sigma)$ such that X, μ, and Σ can be partitioned as*

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

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

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

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

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

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

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

6.2.5. *The t and the F distributions*

Definition 37. A random variable X is said to have the ***t* distribution** if its density function is given by

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

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 32. *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 33. *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*

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

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

Definition 38. 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 34. *Let $X \sim \chi^2(\alpha)$ and $Y \sim \chi^2(\beta)$ be independent random variables. Then*

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

Problems

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

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

Obtain the density function of X.

2. 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*:

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

3. 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 $X \geq 3$.

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

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

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

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

Compute $P[X \leq 4]$.

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

7. Prove Theorem 17.

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

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

10. Prove Theorem 23.

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

12. Mr. Banach carries matchbox in each of his left and right pockets. When he wants a match, he selects the left pocket with probability p and the right pocket with probability $1 - p$. Suppose that initially each box contains $k > 0$ matches. Compute the probability that Mr. Banach discovers a box empty while the other contains $0 < r \leq k$ matches.

13. The number of female insects in a given region is distributed according to a Poisson with mean λ while the number of eggs laid by each insect (male or female) is distributed according to a Poisson with mean μ. Find the probability distribution of the number of eggs of in the region.

14. Let $X_i \sim N(0,1)$, $i = 1, \ldots, 4$, 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 $y \in \mathbb{R}$.

15. 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

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

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.

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

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

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

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

20. Let (X, Y) have joint density function:

$$f(x,y) = \frac{1}{6\pi\sqrt{7}}\exp\left\{-\frac{8}{7}\left(\frac{x^2}{16} - \frac{31x}{32} + \frac{xy}{8} + \frac{y^2}{9} - \frac{4y}{3} + \frac{71}{16}\right)\right\} \quad \text{for} \quad x, y \in \mathbb{R}.$$

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

Convergence of probability distributions

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

7.1. Convergence in distribution

Convergence in distribution is the weakest mode of convergence that we shall analyze.

Definition 39. 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 $F_n \xrightarrow{w} F$.

Example 33. Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (henceforth, 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

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

Then, given $y \ge 0$,

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

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.

Example 34. 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, $\lim_{n\to\infty} E[X^r] \neq E[X]$.

7.2. Convergence in probability

Here we formalize a way of saying that a sequence of random variables approaches another random variable. Note, however, that the definition below says nothing about convergence of the random variables in the sense in which is understood in real analysis; it tells us something about the convergence of a sequence of probabilities.

Definition 40. 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 $\varepsilon > 0$,

$$\lim_{n\to\infty} P\big[\,|X_n - X| > \varepsilon\big] = 0.$$

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

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

$$\lim_{n\to\infty} P\big[\,|X_n - X| \leq \varepsilon\big] = 1.$$

Example 35. 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 36. 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 $n = 1, 2, \ldots$. Let us 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.

Theorem 35. *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 36. *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 37. *Suppose $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{L} X$.*

7.3. Weak law of large numbers

Theorem 38 (WLLN). *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of i.i.d. random variables with $E[X_n] = \mu$ and $\mathrm{Var}[X_n] = \sigma^2 < \infty$, and let $\overline{X}_n = \sum_{i=1}^{n} X_i/n$. Then, $\overline{X}_n \xrightarrow{P} \mu$.*

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

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

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 can be weaken so as to assume that the random variables $X_1, X_2, \ldots$ are independent and each of them has finite mean μ. Hence, the Strong Law is a first moment result while the Weak Law requires the existence of the second moment.

Example 37. Consider a sequence $\{X_n\}_{n=1}^{\infty}$ of i.i.d. random variables with $E[X_n] = \mu$ and $\mathrm{Var}[X_n] = \sigma^2 < \infty$. 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 $\overline{X}_n \xrightarrow{P} \mu$. We ask ourselves 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 $\mathrm{Var}[S_n^2] < \infty$ for each $n = 1, 2, \ldots$. We obtain

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

We using the WLLN, we know 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 each given $k = 1, 2, \ldots$. Then, combining the results in the Theorems above, we obtain

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

7.4. Central limit theorem

We start this section by introducing the notion of random sample.

Definition 41. 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 the results in Proposition 2 and Theorem 20, one can show easily 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 approximate of $\overline{X}_n$ to estimate μ.

Theorem 39 (Lindeberg-Lévy Central Limit Theorem). *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables with $E[X_n] = \mu$ and $0 < \mathrm{Var}[X_n] = \sigma^2 < \infty$, 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*

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

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

Example 38. 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

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

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

Problems

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 $X_n \xrightarrow{L} X$.

2. 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 $X_n \xrightarrow{P} \mu$.

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

4. 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.

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.

We formalize our problem by considering 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 $\mathcal{X}$ 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 $T : \mathcal{X} \to \Theta$.

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

8.1. Consistent estimation

Definition 42. 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 **consistent** for $\theta \in \Theta$ if

$$T(X_1, X_2, \ldots, X_n) \xrightarrow{P} \theta.$$

Intuitively, consistency requires that be likely that the estimator approaches the true value of the parameter as the size of the random sample increases. In other words, if a estimator T is consistent for a parameter θ we may interpret it as "T being close to θ on average" as n increases.

Example 39. Let $\{X_1, X_2, \ldots, X_n\}$ be a random sample from a binomial distribution $b(1,p)$. Then, $E[X_k] = p$ for each $k = 1, 2, \ldots, n$. From the WLLN, we know that

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

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,

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

for each $k = 1, 2, \ldots, n$, 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.

8.2. Sufficient estimation

The desiderata associated to the sufficiency criterion can be summarized by saying that, when proposing a estimator, we wish that the only information obtained about the unknown parameter is that provided by the sample itself. Thus, we find desirable 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 43. 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 40. 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 $t = \sum_{i=1}^{n} x_i$

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

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

Often 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 40 (**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 41. 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

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

so that, by taking $h(x_1, \ldots, x_n) = 1$ 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 42. Let $\{X_1, X_2, \ldots, X_n\}$ 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

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

Then, using the factorization theorem above, it follows that

$$T(X_1, X_2, \ldots, X_n) = \Big(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\Big)$$

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

8.3. Unbiased estimation

Definition 44. 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 **unbiased** for $\theta \in \Theta$ if

$$E_\theta\big[T(X_1, X_2, \ldots, X_n)\big] = \theta.$$

We 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

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

Second, we have

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

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

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

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

8.4. Maximum likelihood estimation

In the previous sections we have introduced several desirable properties that can be used to search for appropriate estimators. Here, we introduce another method which has a 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 1 (Regularity Conditions). *(i) For $\theta, \theta' \in \Theta$, we have*

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

(ii) 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 41. *Given Assumption 1, if θ_0 is the true value of the unknown parameter θ, then*

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

Proof. Notice that, by taking logs, the inequality $L(X_1,\ldots,X_n;\theta_0) \leq L(X_1,\ldots,X_n;\theta)$ can be rewritten as

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

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

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

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)

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

as desired. ∎

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

Definition 45. 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

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

Example 43. 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,

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

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

Example 44. 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 for $\widehat{\theta} < \max\{x_i : i = 1, \ldots, n\}$ one obtains $L(x_1, \ldots, x_n; \widehat{\theta}) = 0$.

Example 45. 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,

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

8.5. Rao-Cramér lower bound and efficient estimation

Here we introduce an important inequality which provides us with 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 1.

Assumption 2 (**Additional Regularity Conditions**). *(i) The point θ_0 is an interior point in Θ;*
(ii) f_θ is twice differentiable with respect to θ;
(iii) the integral $\int f_\theta(x_i)dx_i$ can be differentiated twice (under the integral sign) with respect to θ.

Theorem 42 (**Rao-Cramér Lower Bound**). *Under Assumptions 1 and 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 $E\big[T(X_1, X_2, \ldots, X_n)\big] = \tau(\theta)$, then*

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

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

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

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

Definition 46. 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.

Problems

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 $\lim_{n\to\infty} \text{Var}[T] = 0$, then T is consistent for θ.

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

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

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

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.

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

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

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

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

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

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.

Hypothesis testing

9.1. Neyman-Pearson theory of hypothesis testing

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

$$H_0 : \mu \leq -1, \quad \sigma^2 > 2;$$
$$H_1 : \mu > -1, \quad \sigma^2 \leq 2,$$

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 $\mathcal{X}$, 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 $T : \mathcal{X} \to \Theta$ 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 proceed to the basic theory underlying hypothesis testing.

Definition 47. 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

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

The function (as a function of θ) $E_\theta [\varphi(X_1, \ldots, X_n)]$ is known as the **power function of the test φ** and the least upper bound $\sup_{\theta \in \Theta_0} E_\theta [\varphi(X_1, \ldots, X_n)]$ 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 $E_\theta [\varphi(X_1, \ldots, X_n)] \leq \alpha$ for $\theta \in \Theta_0$ says that if H_0 were true, then the test φ rejects it with probability

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

In other words, the definition of test requires that the probability of the type I error exceeds not 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 48. 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

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

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

To obtain an intuitive interpretation, suppose that both hypotheses are simple so that $(\{\theta_0\}, \{\theta_1\}, \alpha)$ is our hypotheses testing problem. Then, note first that

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

Note that the expected value $E_{\theta_1}[\varphi(X_1, \ldots, X_n)]$ 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

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

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 $P_{\theta_1}[(X_1, \ldots, X_n) \in A^c]$ under the restriction

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

Example 46. Let $\{X_1, X_2, \ldots, X_n\}$ 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

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

so that both H_0 and H_1 are simple hypotheses. 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

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

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 43 (**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

$$\frac{L(x_1,\ldots,x_n;\theta_0)}{L(x_1,\ldots,x_n;\theta_1)} \leq k, \qquad \textit{for each} \quad (x_1,\ldots,x_n) \in A,$$
$$\frac{L(x_1,\ldots,x_n;\theta_0)}{L(x_1,\ldots,x_n;\theta_1)} \geq k, \qquad \textit{for each} \quad (x_1,\ldots,x_n) \in A^c,$$
$$\alpha = \int_A \cdots \int L(x_1,\ldots,x_n;\theta_0)\,dx_1\cdots dx_n.$$

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:

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

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

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

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 47. As in the 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

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

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

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

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 $s = 0, 1$,

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

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

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

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

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

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

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

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

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

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.

9.2. 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 49. 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 matters, $0 \leq \lambda \leq 1$ 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 44. *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 48. 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

$$P_{p_0}[X > k'] \leq \alpha \qquad \text{and} \qquad P_{p_0}[X > k' - 1] > \alpha.$$

Example 49. 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

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

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

$$\sup_{\theta \in \Theta_0} L(x_1, \ldots, x_n; \theta) = \frac{1}{(\widehat{\sigma}_0\sqrt{2\pi})^n} \exp\left[-\frac{\sum_{i=1}^n (x_i - \mu_0)^2}{2\widehat{\sigma}_0^2}\right],$$

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

$$\sup_{\theta\in\Theta} L(x_1,\ldots,x_n;\theta) = \frac{1}{(2\pi/n)^{n/2}\left[\sum_{i=1}^n (x_i-\overline{x}_n)^2\right]^{n/2}} e^{-n/2}.$$

Therefore,

$$\begin{aligned}\lambda(x_1,\ldots,x_n) &= \left[\frac{\sum_{i=1}^n (x_i-\overline{x}_n)^2}{\sum_{i=1}^n (x_i-\mu_0)^2}\right]^{n/2} = \left[\frac{\sum_{i=1}^n (x_i-\overline{x}_n)^2}{\sum_{i=1}^n (x_i-\overline{x}_n)^2 + n(\overline{x}_n-\mu_0)^2}\right]^{n/2} \\ &= \left[\frac{1}{1+\left[n(\overline{x}_n-\mu_0)^2\big/\sum_{i=1}^n (x_i-\overline{x}_n)^2\right]}\right]^{n/2},\end{aligned}$$

which happens to be a decreasing function in $(\overline{x}_n-\mu_0)^2\big/\sum_{i=1}^n (x_i-\overline{x}_n)^2$. Thus,

$$\lambda(x_1,\ldots,x_n) < k \Leftrightarrow \left|\frac{(\overline{x}_n-\mu_0)/\sqrt{n-1}}{\sqrt{\sum_{i=1}^n (x_i-\overline{x}_n)^2/(n-1)}}\right| > k' \Leftrightarrow \left|\frac{\sqrt{n}(\overline{x}_n-\mu_0)}{s_n}\right| > k'',$$

where $s_n^2 = [1/(n-1)]\sum_{i=1}^n (x_i-\overline{x}_n)^2$ 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 31 (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''.

Problems

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.

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 $H_1;\ \sigma_1^2\neq\sigma_2^2$.