Chapter 1

Probability Theory

Mathematical statistics relies on probability theory, which in turn is based on measure theory. The present chapter provides some principal concepts and notational conventions of probability theory, and some important results that are essential tools used in this book. A more complete account of probability theory can be found in many standard textbooks, for example, Billingsley (1986) and Chung (1974). The reader is assumed to be familiar with set operations and set functions (mappings) in advanced calculus.

1.1 Probability Spaces and Random Elements

In an elementary probability course, one defines a *random experiment* to be an experiment for which the outcome of the experiment cannot be predicted with certainty, and the probability of A (a collection of possible outcomes) to be the fraction of times that the outcome of the random experiment results in A in a large number of trials of the random experiment. A rigorous and logically consistent definition of probability was given by A. N. Kolmogorov in his measure-theoretic fundamental development of probability theory in 1933.

1.1.1 σ-fields and measures

Let Ω be a set of elements of interest. For example, Ω can be a set of numbers, a subinterval of the real line, or all possible outcomes of a random experiment. In probability theory, Ω is often called the outcome space, whereas in statistical theory, Ω is called the *sample space*. This is because in probability and statistics, Ω is usually the set of all possible outcomes of a random experiment under study.

A *measure* is a natural mathematical extension of the length, area, or volume of subsets in one-, two-, or three-dimensional Euclidean space. In a given sample space Ω, a measure is a set function defined for certain subsets of Ω. It will be necessary for this collection of subsets to satisfy certain properties, which are given in the following definition.

Definition 1.1. Let $\mathcal{F}$ be a collection of subsets of a sample space Ω. $\mathcal{F}$ is called a σ-field (or σ-algebra) if and only if it has the following properties.
(i) The empty set $\emptyset \in \mathcal{F}$.
(ii) If $A \in \mathcal{F}$, then the complement $A^c \in \mathcal{F}$.
(iii) If $A_i \in \mathcal{F}$, $i = 1, 2, ...$, then their union $\cup A_i \in \mathcal{F}$. ∎

A pair $(\Omega, \mathcal{F})$ consisting of a set Ω and a σ-field $\mathcal{F}$ of subsets of Ω is called a *measurable space*. The elements of $\mathcal{F}$ are called measurable sets in measure theory or *events* in probability and statistics.

Since $\emptyset^c = \Omega$, it follows from (i) and (ii) in Definition 1.1 that $\Omega \in \mathcal{F}$ if $\mathcal{F}$ is a σ-field on Ω. Also, it follows from (ii) and (iii) that if $A_i \in \mathcal{F}$, $i = 1, 2, ...$, and $\mathcal{F}$ is a σ-field, then the intersection $\cap A_i \in \mathcal{F}$. This can be shown using DeMorgan's law: $(\cap A_i)^c = \cup A_i^c$.

For any given Ω, there are two trivial σ-fields. The first one is the collection containing exactly two elements, $\emptyset$ and Ω. This is the smallest possible σ-field on Ω. The second one is the collection of all subsets of Ω, which is called the power set and is the largest σ-field on Ω.

Let us now consider some nontrivial σ-fields. Let A be a nonempty proper subset of Ω ($A \subset \Omega$, $A \neq \Omega$). Then (verify)

$$\{\emptyset, A, A^c, \Omega\} \tag{1.1}$$

is a σ-field. In fact, this is the smallest σ-field containing A in the sense that if $\mathcal{F}$ is any σ-field containing A, then the σ-field in (1.1) is a subcollection of $\mathcal{F}$. In general, the smallest σ-field containing $\mathcal{C}$, a collection of subsets of Ω, is denoted by $\sigma(\mathcal{C})$ and is called the σ-field generated by $\mathcal{C}$. Hence, the σ-field in (1.1) is $\sigma(\{A\})$. Note that $\sigma(\{A, A^c\})$, $\sigma(\{A, \Omega\})$, and $\sigma(\{A, \emptyset\})$ are all the same as $\sigma(\{A\})$. Of course, if $\mathcal{C}$ itself is a σ-field, then $\sigma(\mathcal{C}) = \mathcal{C}$.

On the real line $\mathcal{R}$, there is a special σ-field that will be used almost exclusively. Let $\mathcal{C}$ be the collection of all finite open intervals on $\mathcal{R}$. Then $\mathcal{B} = \sigma(\mathcal{C})$ is called the *Borel σ-field*. The elements of $\mathcal{B}$ are called *Borel sets*. The Borel σ-field $\mathcal{B}^k$ on the k-dimensional Euclidean space $\mathcal{R}^k$ can be similarly defined. It can be shown that all intervals (finite or infinite), open sets, and closed sets are Borel sets. To illustrate, we now show that on the real line, $\mathcal{B} = \sigma(\mathcal{O})$, where $\mathcal{O}$ is the collection of all open sets. Typically, one needs to show that $\sigma(\mathcal{C}) \subset \sigma(\mathcal{O})$ and $\sigma(\mathcal{O}) \subset \sigma(\mathcal{C})$. Since an open interval is an open set, $\mathcal{C} \subset \mathcal{O}$ and, hence, $\sigma(\mathcal{C}) \subset \sigma(\mathcal{O})$ (see Exercise 3 in §1.6). Let U be an open set. Then U can be expressed as a union

of a sequence of finite open intervals (see Royden (1968, p.39)). Hence, $U \in \sigma(\mathcal{C})$ (Definition 1.1(iii)) and $\mathcal{O} \subset \sigma(\mathcal{C})$. By the definition of $\sigma(\mathcal{O})$, $\sigma(\mathcal{O}) \subset \sigma(\mathcal{C})$. This completes the proof.

Let $C \subset \mathcal{R}^k$ be a Borel set and let $\mathcal{B}_C = \{C \cap B : B \in \mathcal{B}^k\}$. Then $(C, \mathcal{B}_C)$ is a measurable space and $\mathcal{B}_C$ is called the Borel σ-field on C.

Now we can introduce the notion of a measure.

Definition 1.2. Let $(\Omega, \mathcal{F})$ be a measurable space. A set function ν defined on $\mathcal{F}$ is called a *measure* if and only if it has the following properties.
(i) $0 \leq \nu(A) \leq \infty$ for any $A \in \mathcal{F}$.
(ii) $\nu(\emptyset) = 0$.
(iii) If $A_i \in \mathcal{F}$, $i = 1, 2, ...,$ and A_i's are disjoint, i.e., $A_i \cap A_j = \emptyset$ for any $i \neq j$, then

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

The triple $(\Omega, \mathcal{F}, \nu)$ is called a *measure space*. If $\nu(\Omega) = 1$, then ν is called a *probability measure* and we usually denote it by P instead of ν, in which case $(\Omega, \mathcal{F}, P)$ is called a *probability space*.

Although measure is an extension of length, area, or volume, sometimes it can be quite abstract. For example, the following set function is a measure:

$$\nu(A) = \begin{cases} \infty & A \in \mathcal{F}, A \neq \emptyset \\ 0 & A = \emptyset. \end{cases} \tag{1.2}$$

Since a measure can take ∞ as its value, we must know how to do arithmetic with ∞. In this book, it suffices to know that (1) for any $x \in \mathcal{R}$, $\infty + x = \infty$, $x\,\infty = \infty$ if $x > 0$, $x\,\infty = -\infty$ if $x < 0$, and $0\,\infty = 0$; (2) $\infty + \infty = \infty$; and (3) $\infty\,\infty = \infty$. However, $\infty - \infty$ or ∞/∞ is not defined.

The following examples provide two very important measures in probability and statistics.

Example 1.1 (Counting measure). Let Ω be a sample space, $\mathcal{F}$ the collection of all subsets, and $\nu(A)$ the number of elements in $A \in \mathcal{F}$ ($\nu(A) = \infty$ if A contains infinitely many elements). Then ν is a measure on $\mathcal{F}$ and is called the *counting measure*. $\blacksquare$

Example 1.2 (Lebesgue measure). There is a unique measure m on $(\mathcal{R}, \mathcal{B})$ that satisfies

$$m([a, b]) = b - a \tag{1.3}$$

for every finite interval $[a, b]$, $-\infty < a \leq b < \infty$. This is called the *Lebesgue measure*. If we restrict m to the measurable space $([0, 1], \mathcal{B}_{[0,1]})$, then m is a probability measure. $\blacksquare$

If Ω is *countable* in the sense that there is a one-to-one correspondence between Ω and the set of all integers, then one can usually consider the trivial σ-field that contains all subsets of Ω and a measure that assigns a value to every subset of Ω. When Ω is uncountable (e.g., $\Omega = \mathcal{R}$ or $[0,1]$), it is not possible to define a reasonable measure for every subset of Ω; for example, it is not possible to find a measure on all subsets of $\mathcal{R}$ and still satisfy property (1.3). This is why it is necessary to introduce σ-fields that are smaller than the power set.

The following result provides some basic properties of measures. Whenever we consider $\nu(A)$, it is implicitly assumed that $A \in \mathcal{F}$.

Proposition 1.1. Let $(\Omega, \mathcal{F}, \nu)$ be a measure space.
(i) (Monotonicity). If $A \subset B$, then $\nu(A) \leq \nu(B)$.
(ii) (Subadditivity). For any sequence $A_1, A_2, ...,$

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

(iii) (Continuity). If $A_1 \subset A_2 \subset A_3 \subset \cdots$ (or $A_1 \supset A_2 \supset A_3 \supset \cdots$ and $\nu(A_1) < \infty$), then

$$\nu\left(\lim_{n\to\infty} A_n\right) = \lim_{n\to\infty} \nu\left(A_n\right),$$

where

$$\lim_{n\to\infty} A_n = \bigcup_{i=1}^{\infty} A_i \quad \left(\text{ or } = \bigcap_{i=1}^{\infty} A_i\right).$$

Proof. We prove (i) only. The proofs of (ii) and (iii) are left as exercises. Since $A \subset B$, $B = A \cup (A^c \cap B)$ and A and $A^c \cap B$ are disjoint. By Definition 1.2(iii), $\nu(B) = \nu(A) + \nu(A^c \cap B)$, which is no smaller than $\nu(A)$ since $\nu(A^c \cap B) \geq 0$ by Definition 1.2(i). ∎

There is a one-to-one correspondence between the set of all probability measures on $(\mathcal{R}, \mathcal{B})$ and a set of functions on $\mathcal{R}$. Let P be a probability measure. The *cumulative distribution function* (c.d.f.) of P is defined to be

$$F(x) = P\left((-\infty, x]\right), \quad x \in \mathcal{R}. \tag{1.4}$$

Proposition 1.2. (i) Let F be a c.d.f. on $\mathcal{R}$. Then
(a) $F(-\infty) = \lim_{x\to-\infty} F(x) = 0$;
(b) $F(\infty) = \lim_{x\to\infty} F(x) = 1$;
(c) F is nondecreasing, i.e., $F(x) \leq F(y)$ if $x \leq y$;
(d) F is right continuous, i.e., $\lim_{y\to x, y>x} F(y) = F(x)$.
(ii) Suppose that a real-valued function F on $\mathcal{R}$ satisfies (a)-(d) in part (i). Then F is the c.d.f. of a unique probability measure on $(\mathcal{R}, \mathcal{B})$. ∎

The *Cartesian product* of any k sets $A_1, ..., A_k$ (which may be subsets of different sample spaces) is defined as the set of all $(a_1, ..., a_k)$, $a_i \in A_i$, and is denoted by $A_1 \times \cdots \times A_k$. Let $(\Omega_i, \mathcal{F}_i, \nu_i)$, $i = 1, ..., k$, be k measure spaces. We now introduce a convenient way of constructing a σ-field and a measure on the product space $\Omega_1 \times \cdots \times \Omega_k$.

First, note that $\mathcal{F}_1 \times \cdots \times \mathcal{F}_k$ is not necessarily a σ-field. We define the σ-field $\sigma(\mathcal{F}_1 \times \cdots \times \mathcal{F}_k)$ as the *product* σ-field on $\Omega_1 \times \cdots \times \Omega_k$. As an example, consider $(\Omega_i, \mathcal{F}_i) = (\mathcal{R}, \mathcal{B})$ for all i. Then the product space is $\mathcal{R}^k$ and it can be shown that the product σ-field is the same as the Borel σ-field on $\mathcal{R}^k$, which is the σ-field generated by $\mathcal{O}$, all open sets in $\mathcal{R}^k$.

In Example 1.2, the usual length of an interval $[a, b] \subset \mathcal{R}$ is the same as the Lebesgue measure of $[a, b]$. Consider a rectangle $[a_1, b_1] \times [a_2, b_2] \subset \mathcal{R}^2$. The usual area of $[a_1, b_1] \times [a_2, b_2]$ is

$$(b_1 - a_1)(b_2 - a_2) = m([a_1, b_1])m([a_2, b_2]), \tag{1.5}$$

i.e., the product of the Lebesgue measures of two intervals $[a_1, b_1]$ and $[a_2, b_2]$. Note that $[a_1, b_1] \times [a_2, b_2]$ is a measurable set by the definition of the product σ-field. Is $m([a_1, b_1])m([a_2, b_2])$ the same as the value of a measure defined on the product σ-field? To answer this, we need the following technical definition. A measure ν on $(\Omega, \mathcal{F})$ is said to be σ-*finite* if there exists a sequence $\{A_1, A_2, ...\}$ such that $\cup A_i = \Omega$ and $\nu(A_i) < \infty$ for all i. Any finite measure (such as a probability measure) is clearly σ-finite. The Lebesgue measure in Example 1.2 is σ-finite, since $\mathcal{R} = \cup A_n$ with $A_n = (-n, n)$, $n = 1, 2, ...$ The counting measure in Example 1.1 is σ-finite if and only if Ω is countable. The measure defined by (1.2), however, is not σ-finite.

Proposition 1.3 (Product measure theorem). Let $(\Omega_i, \mathcal{F}_i, \nu_i)$, $i = 1, ..., k$, be measure spaces and ν_i be σ-finite measures. Then there exists a unique σ-finite measure on the product σ-field $\sigma(\mathcal{F}_1 \times \cdots \times \mathcal{F}_k)$, called the *product measure* and denoted by $\nu_1 \times \cdots \times \nu_k$, such that

$$\nu_1 \times \cdots \times \nu_k(A_1 \times \cdots \times A_k) = \nu_1(A_1) \cdots \nu_k(A_k)$$

for all $A_i \in \mathcal{F}_i$, $i = 1, ..., k$. ∎

Thus, in $\mathcal{R}^2$ there is a unique measure, the product measure $m \times m$, for which $m \times m([a_1, b_1] \times [a_2, b_2])$ is equal to the value given by (1.5). This measure is called the Lebesgue measure on $(\mathcal{R}^2, \mathcal{B}^2)$. Similarly, we can define the Lebesgue measure on $(\mathcal{R}^3, \mathcal{B}^3)$, which exactly equals the usual volume for subsets of the form $[a_1, b_1] \times [a_2, b_2] \times [a_3, b_3]$.

In general, the product measure space generated by $(\Omega_i, \mathcal{F}_i, \nu_i)$, $i = 1, ..., k$, is denoted by $\prod_{i=1}^{k} (\Omega_i, \mathcal{F}_i, \nu_i)$.

The concept of c.d.f. can be extended to $\mathcal{R}^k$. Let P be a probability measure on $(\mathcal{R}^k, \mathcal{B}^k)$. The c.d.f. (or *joint* c.d.f.) of P is defined by

$$F(x_1, ..., x_k) = P\left((-\infty, x_1] \times \cdots \times (-\infty, x_k]\right), \quad x_i \in \mathcal{R}. \tag{1.6}$$

Note that P is not necessarily a product measure. Again, there is a one-to-one correspondence between probability measures and joint c.d.f.'s on $\mathcal{R}^k$. Some properties of a joint c.d.f. are given in Exercise 10 in §1.6.

1.1.2 Measurable functions and distributions

Since Ω can be quite arbitrary, it is often convenient to consider a function (mapping) f from Ω to a simpler space Λ (often $\Lambda = \mathcal{R}^k$, the k-dimensional Euclidean space). Let $B \subset \Lambda$. Then the *inverse image* of B under f is

$$f^{-1}(B) = \{f \in B\} = \{\omega \in \Omega : f(\omega) \in B\}.$$

The inverse function f^{-1} need not exist for $f^{-1}(B)$ to be defined. The reader is asked to verify the following properties:

(a) $f^{-1}(B^c) = (f^{-1}(B))^c$ for any $B \subset \Lambda$;
(b) $f^{-1}(\cup B_i) = \cup f^{-1}(B_i)$ for any $B_i \subset \Lambda$, $i = 1, 2, ...$

Let $\mathcal{C}$ be a collection of subsets of Λ. We define

$$f^{-1}(\mathcal{C}) = \{f^{-1}(C) : C \in \mathcal{C}\}.$$

Definition 1.3. Let $(\Omega, \mathcal{F})$ and $(\Lambda, \mathcal{G})$ be measurable spaces and f a function from Ω to Λ. The function f is called a *measurable function* from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$ if and only if $f^{-1}(\mathcal{G}) \subset \mathcal{F}$. ∎

If $\Lambda = \mathcal{R}$ and $\mathcal{G} = \mathcal{B}$ (Borel σ-field), then f is said to be *Borel measurable* or is called a *Borel function*.

In probability theory, a measurable function is called a *random element* and denoted by one of X, Y, Z,.... If X is measurable from $(\Omega, \mathcal{F})$ to $(\mathcal{R}, \mathcal{B})$, then it is called a *random variable*; if X is measurable from $(\Omega, \mathcal{F})$ to $(\mathcal{R}^k, \mathcal{B}^k)$, then it is called a *random k-vector* (as a notational convention in this book, any vector is considered to be a row vector). If $X_1, ..., X_k$ are random variables defined on a common probability space, then the vector $(X_1, ..., X_k)$ is a random k-vector.

If f is measurable from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$, then $f^{-1}(\mathcal{G})$ is a sub-σ-field of $\mathcal{F}$ (verify). It is called the σ-field generated by f and is denoted by $\sigma(f)$.

Now we consider some examples of measurable functions. If $\mathcal{F}$ is the collection of all subsets of Ω, then any function f is measurable. Let $A \subset \Omega$. The *indicator function* for A is defined as

$$I_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A. \end{cases}$$

For any $B \subset \mathcal{R}$,

$$I_A^{-1}(B) = \begin{cases} \emptyset & 0 \notin B, 1 \notin B \\ A & 0 \notin B, 1 \in B \\ A^c & 0 \in B, 1 \notin B \\ \Omega & 0 \in B, 1 \in B. \end{cases}$$

Then $\sigma(I_A)$ is the σ-field given in (1.1). If A is a measurable set, then I_A is a Borel function.

Note that $\sigma(I_A)$ is a much smaller σ-field than the original σ-field $\mathcal{F}$. This is another reason why we introduce the concept of measurable functions and random variables, in addition to the reason that it is easy to deal with numbers. Often the σ-field $\mathcal{F}$ (such as the power set) contains too many subsets and we are only interested in some of them. One can then define a random variable X with $\sigma(X)$ containing subsets that are of interest. In general, $\sigma(X)$ is between the trivial σ-field $\{\emptyset, \Omega\}$ and $\mathcal{F}$, and contains more subsets if X is more complicated. For the simplest function I_A, we have shown that $\sigma(I_A)$ contains only four elements.

The class of *simple functions* is obtained by taking linear combinations of indicators of measurable sets, i.e.,

$$\varphi(\omega) = \sum_{i=1}^{k} a_i I_{A_i}(\omega), \tag{1.7}$$

where $A_1, ..., A_k$ are measurable sets on Ω and $a_1, ..., a_k$ are real numbers. One can show directly that such a function is a Borel function, but it follows immediately from Proposition 1.4. Let $A_1, ..., A_k$ be a partition of Ω, i.e., A_i's are disjoint and $A_1 \cup \cdots \cup A_k = \Omega$. Then the simple function φ given by (1.7) with distinct a_i's exactly characterizes this partition and $\sigma(\varphi) = \sigma(\{A_1, ..., A_k\})$.

Proposition 1.4. Let $(\Omega, \mathcal{F})$ be a measurable space.
(i) If f and g are Borel, then so are fg and $af + bg$, where a and b are real numbers; also, f/g is Borel provided $g(\omega) \neq 0$ for any $\omega \in \Omega$.
(ii) f is Borel if and only if $f^{-1}(a, \infty) \in \mathcal{F}$ for all $a \in \mathcal{R}$.
(iii) If $f_1, f_2, ...$ are Borel, then so are $\sup_n f_n$, $\inf_n f_n$, $\limsup_n f_n$, and $\liminf_n f_n$. Furthermore, the set

$$A = \left\{\omega \in \Omega : \lim_{n\to\infty} f_n(\omega) \text{ exists}\right\}$$

is an event and the function

$$h(\omega) = \begin{cases} \lim_{n\to\infty} f_n(\omega) & \omega \in A \\ f_1(\omega) & \omega \notin A \end{cases}$$

is Borel.
(iv) Suppose that f is measurable from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$ and g is measurable from $(\Lambda, \mathcal{G})$ to $(\Delta, \mathcal{H})$. Then the composite function $g \circ f$ is measurable from $(\Omega, \mathcal{F})$ to $(\Delta, \mathcal{H})$.
(v) Let Ω be a Borel set in $\mathcal{R}^p$. If f is a continuous function from Ω to $\mathcal{R}^q$, then f is measurable. ∎

Proposition 1.4 indicates that there are many Borel functions. In fact, it is hard to find a non-Borel function.

Let $(\Omega, \mathcal{F}, \nu)$ be a measure space and f be a measurable function from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$. The *induced measure* by f, denoted by $\nu \circ f^{-1}$, is a measure on $\mathcal{G}$ defined as

$$\nu \circ f^{-1}(B) = \nu(f \in B) = \nu\left(f^{-1}(B)\right), \quad B \in \mathcal{G}. \tag{1.8}$$

It is usually easier to deal with $\nu \circ f^{-1}$ than to deal with ν since $(\Lambda, \mathcal{G})$ is usually simpler than $(\Omega, \mathcal{F})$. Furthermore, subsets not in $\sigma(f)$ are not involved in the definition of $\nu \circ f^{-1}$. As we discussed earlier, in some cases we are only interested in subsets in $\sigma(f)$.

If $\nu = P$ is a probability measure and X is a random variable or a random vector, then $P \circ X^{-1}$ is called the *law* or the *distribution* of X and is denoted by P_X. The c.d.f. of P_X defined by (1.4) or (1.6) is also called the c.d.f. or joint c.d.f. of X and is denoted by F_X. On the other hand, for any c.d.f. or joint c.d.f. F, there exists at least one random variable or vector (usually there are many) defined on some probability space for which $F_X = F$. The following are some examples of random variables and their c.d.f.'s. More examples can be found in §1.3.1.

Example 1.3 (Discrete c.d.f.'s). Let $a_1 < a_2 < \cdots$ be a sequence of real numbers and let p_n, $n = 1, 2, ...$, be a sequence of positive numbers such that $\sum_{n=1}^{\infty} p_n = 1$. Define

$$F(x) = \begin{cases} \sum_{i=1}^{n} p_i & a_n \le x < a_{n+1}, \quad n = 1, 2, ... \\ 0 & -\infty < x < a_1. \end{cases} \tag{1.9}$$

Then F is a *stepwise* c.d.f. It has a jump of size p_n at each a_n and is flat between a_n and a_{n+1}, $n = 1, 2, ...$ Such a c.d.f. is called a *discrete* c.d.f. and the corresponding random variable is called a *discrete random variable*. We can easily obtain a random variable having F in (1.9) as its c.d.f. For example, let $\Omega = \{a_1, a_2, ...\}$, $\mathcal{F}$ be the collection of all subsets of Ω,

$$P(A) = \sum_{i: a_i \in A} p_i, \quad A \in \mathcal{F}, \tag{1.10}$$

and $X(\omega) = \omega$. One can show that P is a probability measure and the c.d.f. of X is F in (1.9). ∎

Example 1.4 (Continuous c.d.f.'s). Opposite to the class of discrete c.d.f.'s is the class of continuous c.d.f.'s. Without the concepts of integration and differentiation introduced in the next section, we can only provide a few examples of continuous c.d.f.'s. One such example is the *uniform* c.d.f. on the interval $[a, b]$ defined as

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

Another example is the *exponential* c.d.f. defined as

$$F(x) = \begin{cases} 0 & -\infty < x < 0 \\ 1 - e^{-x/\theta} & 0 \le x < \infty, \end{cases}$$

where θ is a fixed positive constant. Note that both uniform and exponential c.d.f.'s are continuous functions. ∎

1.2 Integration and Differentiation

Differentiation and integration are two of the main components of calculus. This is also true in measure theory or probability theory, except that integration is introduced first whereas in calculus, differentiation is introduced first.

1.2.1 Integration

An important concept needed in probability and statistics is the *integration* of Borel functions with respect to (w.r.t.) a measure ν, which is a type of "average". The definition proceeds in several steps. First, we define the integral of a nonnegative simple function, i.e., a simple function φ given by (1.7) with $a_i \ge 0$, $i = 1, ..., k$.

Definition 1.4(a). The integral of a nonnegative simple function φ given by (1.7) w.r.t. ν is defined as

$$\int \varphi d\nu = \sum_{i=1}^{k} a_i \nu(A_i). \quad ∎ \tag{1.11}$$

The right-hand side of (1.11) is a weighted average of a_i's with $\nu(A_i)$'s as weights. Since $a\infty = \infty$ if $a > 0$ and $a\infty = 0$ if $a = 0$, the right-hand side of (1.11) is always well defined, although $\int \varphi d\nu = \infty$ is possible. Note

that different a_i's and A_i's may produce the same function φ; for example, with $\Omega = \mathcal{R}$,

$$2I_{(0,1)}(x) + I_{[1,2]}(x) = I_{(0,2]}(x) + I_{(0,1)}(x).$$

However, one can show that different representations of φ in (1.7) produce the same value for $\int \varphi d\nu$ so that the integral of a nonnegative simple function is well defined.

Next, we consider nonnegative Borel function f.

Definition 1.4(b). Let f be a nonnegative Borel function and let $\mathcal{S}_f$ be the collection of all nonnegative simple functions of the form (1.7) satisfying $\varphi(\omega) \leq f(\omega)$ for any $\omega \in \Omega$. The integral of f w.r.t. ν is defined as

$$\int f d\nu = \sup \left\{ \int \varphi d\nu : \ \varphi \in \mathcal{S}_f \right\}. \quad \blacksquare$$

Hence, for any Borel function $f \geq 0$, there exists a sequence of simple functions $\varphi_1, \varphi_2, ...$ such that $0 \leq \varphi_i \leq f$ for all i and $\lim_{n\to\infty} \int \varphi_n d\nu = \int f d\nu$.

Finally, for a Borel function f, we first define the positive part of f by

$$f_+(\omega) = \max\{f(\omega), 0\}$$

and the negative part of f by

$$f_-(\omega) = \max\{-f(\omega), 0\}.$$

Note that f_+ and f_- are nonnegative Borel functions, $f(\omega) = f_+(\omega) - f_-(\omega)$, and $|f(\omega)| = f_+(\omega) + f_-(\omega)$.

Definition 1.4(c). Let f be a Borel function. We say $\int f d\nu$ exists if and only if at least one of $\int f_+ d\nu$ and $\int f_- d\nu$ is finite, in which case

$$\int f d\nu = \int f_+ d\nu - \int f_- d\nu. \tag{1.12}$$

Let A be a measurable set and I_A be its indicator function. The integral of f over A is defined as

$$\int_A f d\nu = \int I_A f d\nu. \quad \blacksquare$$

Note that the left-hand side of (1.12) is always well defined, although it can be ∞ or $-\infty$. When both $\int f_+ d\nu$ and $\int f_- d\nu$ are finite, we say that f is

integrable (for a nonnegative Borel function f, f is integrable if $\int f d\nu < \infty$). Note that the existence of $\int f d\nu$ is different from the integrability of f.

The integral of f may be denoted differently whenever there is a need to indicate the variable(s) to be integrated and the integration domain; for example, $\int_\Omega f d\nu$, $\int f(\omega) d\nu$, $\int f(\omega) d\nu(\omega)$, or $\int f(\omega)\nu(d\omega)$, and so on. In probability and statistics, $\int X dP$ is usually written as EX or $E(X)$ and called the *expectation* or *expected value* of X. If F is the c.d.f. of P on $(\mathcal{R}^k, \mathcal{B}^k)$, $\int f(x) dP$ is also denoted by $\int f(x) dF(x)$ or $\int f dF$.

Example 1.5. Let Ω be a countable set, $\mathcal{F}$ be all subsets of Ω, and ν be the counting measure given in Example 1.1. For any Borel function f, the integral of f w.r.t. ν is

$$\int f d\nu = \sum_{\omega \in \Omega} f(\omega). \tag{1.13}$$

This is obvious if f is a simple function. The proof for general f is left as an exercise. ∎

Example 1.6. If $\Omega = \mathcal{R}$ and ν is the Lebesgue measure, then the integral of f over an interval $[a, b]$ agrees with the Riemann integral in calculus when the latter is well defined, and is usually written as $\int_{[a,b]} f(x) dx = \int_a^b f(x) dx$. However, there are functions for which the Lebesgue integrals are defined but not the Riemann integrals. ∎

We now introduce some properties of integrals. The proof of the following result is left to the reader.

Proposition 1.5 (Linearity of integrals). Let $(\Omega, \mathcal{F}, \nu)$ be a measure space and f and g be Borel functions.
(i) If $\int f d\nu$ exists and $a \in \mathcal{R}$, then $\int (af) d\nu$ exists and is equal to $a \int f d\nu$.
(ii) If both $\int f d\nu$ and $\int g d\nu$ exist and $\int f d\nu + \int g d\nu$ is well defined, then $\int (f + g) d\nu$ exists and is equal to $\int f d\nu + \int g d\nu$. ∎

If a statement holds for all ω in $\Omega - \mathcal{N}$ with $\nu(\mathcal{N}) = 0$, then the statement is said to hold a.e. (almost every where) ν (or simply a.e. if the measure ν is clear from the context). If ν is a probability measure, then a.e. may be replaced by a.s. (almost surely).

Proposition 1.6. Let $(\Omega, \mathcal{F}, \nu)$ be a measure space and f and g be Borel.
(i) If $f \le g$ a.e., then $\int f d\nu \le \int g d\nu$, provided that the integrals exist.
(ii) If $f \ge 0$ a.e. and $\int f d\nu = 0$, then $f = 0$ a.e. ∎

Some direct consequences of Proposition 1.6(i) are: $|\int f d\nu| \le \int |f| d\nu$; if $f \ge 0$ a.e., then $\int f d\nu \ge 0$; and if $f = g$ a.e., then $\int f d\nu = \int g d\nu$.

We now prove part (ii) of Proposition 1.6 as an illustration. The proof for part (i) is left to the reader. Let $A = \{f > 0\}$ and $A_n = \{f \geq n^{-1}\}$, $n = 1, 2, \ldots$. Then $A_n \subset A$ for any n and $\lim_{n\to\infty} A_n = A$ (why?). By Proposition 1.1(iii), $\lim_{n\to\infty} \nu(A_n) = \nu(A)$. Using Proposition 1.5 and part (i) of Proposition 1.6, we obtain that

$$n^{-1}\nu(A_n) = \int n^{-1} I_{A_n} d\nu \leq \int f I_{A_n} d\nu \leq \int f d\nu = 0$$

for any n. Hence $\nu(A) = 0$ and $f = 0$ a.e.

It is sometimes required to know whether the following interchange of two operations is valid:

$$\int \lim_{n\to\infty} f_n d\nu = \lim_{n\to\infty} \int f_n d\nu, \tag{1.14}$$

where $f_1, f_2, \ldots$ is a sequence of Borel functions. Note that we only require $\lim_{n\to\infty} f_n$ exists a.e. Also, the limit of a sequence of Borel functions is Borel (Proposition 1.4). The following example shows that (1.14) is not always true.

Example 1.7. Consider $(\mathcal{R}, \mathcal{B})$ and the Lebesgue measure. Define $f_n(x) = nI_{[0,n^{-1}]}(x)$, $n = 1, 2, \ldots$. Then $\lim_{n\to\infty} f_n(x) = 0$ for all x but $x = 0$. Since the Lebesgue measure of a single point set is 0 (see Example 1.2), $\lim_{n\to\infty} f_n(x) = 0$ a.e. and $\int \lim_{n\to\infty} f_n(x)dx = 0$. On the other hand, $\int f_n(x)dx = 1$ for any n and, hence, $\lim_{n\to\infty} \int f_n(x)dx = 1$. ∎

The following result gives some sufficient conditions under which (1.14) holds.

Theorem 1.1. Let $f_1, f_2, \ldots$ be a sequence of Borel functions.
(i) (Dominated convergence theorem). If $\lim_{n\to\infty} f_n = f$ a.e. and there exists an integrable function g such that $|f_n| \leq g$ a.e., then (1.14) holds.
(ii) (Fatou's lemma). If $f_n \geq 0$, then

$$\int \liminf_n f_n d\nu \leq \liminf_n \int f_n d\nu.$$

(iii) (Monotone convergence theorem). If $0 \leq f_1 \leq f_2 \leq \cdots$ and $\lim_{n\to\infty} f_n = f$ a.e., then (1.14) holds. ∎

The proof is omitted. However, it can be seen that if f in (iii) is integrable, then part (iii) is a consequence of part (i). The following is an application of Theorem 1.1.

Example 1.8 (Interchange of differentiation and integration). Let $(\Omega, \mathcal{F}, \nu)$ be a measure space and for any fixed θ, $f(\omega, \theta)$ be a Borel function on

Ω. Suppose that $\partial f(\omega,\theta)/\partial\theta$ exists a.e. for $\theta \in (a,b) \subset \mathcal{R}$ and that $|\partial f(\omega,\theta)/\partial\theta| \le g(\omega)$ a.e., where g is an integrable function on Ω. Then, for each $\theta \in (a,b)$, $\partial f(\omega,\theta)/\partial\theta$ is integrable and

$$\frac{d}{d\theta}\int f(\omega,\theta)d\nu = \int \frac{\partial f(\omega,\theta)}{\partial\theta}d\nu. \quad \blacksquare$$

Theorem 1.2 (Change of variables). Let f be measurable from $(\Omega,\mathcal{F},\nu)$ to $(\Lambda,\mathcal{G})$ and g be Borel on $(\Lambda,\mathcal{G})$. Then

$$\int_\Omega g\circ f d\nu = \int_\Lambda g d(\nu\circ f^{-1}), \tag{1.15}$$

i.e., if either integral exists, then so does the other, and the two are the same. $\blacksquare$

The proof is again omitted. Note that integration domains are indicated on both sides of (1.15). This result extends the change of variable formula for Riemann integrals, i.e., $\int g(y)dy = \int g(f(x))f'(x)dx$, $y = f(x)$.

Result (1.15) is very important in probability and statistics. Let X be a random variable on a probability space $(\Omega,\mathcal{F},P)$. If $EX = \int_\Omega X dP$ exists, then usually it is much simpler to compute $EX = \int_{\mathcal{R}} x dP_X$, where $P_X = P\circ X^{-1}$ is the law of X. Let Y be a random vector from Ω to $\mathcal{R}^k$ and g be Borel from $\mathcal{R}^k$ to $\mathcal{R}$. According to (1.15), $Eg(Y)$ can be computed as $\int_{\mathcal{R}^k} g(y)dP_Y$ or $\int_{\mathcal{R}} x dP_{g(Y)}$, depending on which of P_Y and $P_{g(Y)}$ is easier to handle. As a more specific example, consider $k=2$, $Y=(X_1,X_2)$, and $g(Y)=X_1+X_2$. Using Proposition 1.5(ii), $E(X_1+X_2)=EX_1+EX_2$ and, hence,

$$E(X_1+X_2) = \int_{\mathcal{R}} x dP_{X_1} + \int_{\mathcal{R}} x dP_{X_2}.$$

Then we need to handle two integrals involving P_{X_1} and P_{X_2}. On the other hand,

$$E(X_1+X_2) = \int_{\mathcal{R}} x dP_{X_1+X_2},$$

which involves one integral w.r.t. $P_{X_1+X_2}$. Unless we have some knowledge about the joint c.d.f. of (X_1,X_2), it is not easy to handle $P_{X_1+X_2}$.

The following theorem states how to evaluate an integral w.r.t. a product measure via iterated integration.

Theorem 1.3 (Fubini's theorem). Let ν_i be a σ-finite measure on $(\Omega_i,\mathcal{F}_i)$, $i=1,2$, and let f be a Borel function on $\prod_{i=1}^2(\Omega_i,\mathcal{F}_i)$ whose integral w.r.t. $\nu_1\times\nu_2$ exists. Then

$$g(\omega_2) = \int_{\Omega_1} f(\omega_1,\omega_2)d\nu_1$$

exists a.e. ν_2 and defines a Borel function on Ω_2 whose integral w.r.t. ν_2 exists, and

$$\int_{\Omega_1\times\Omega_2} f(\omega_1,\omega_2)d\nu_1\times\nu_2 = \int_{\Omega_2}\left[\int_{\Omega_1} f(\omega_1,\omega_2)d\nu_1\right]d\nu_2. \quad \blacksquare$$

This result can be naturally extended to the integral w.r.t. the product measure on $\prod_{i=1}^{k}(\Omega_i,\mathcal{F}_i)$ for any finite positive integer k.

Example 1.9. Let $\Omega_1=\Omega_2=\{0,1,2,...\}$, and $\nu_1=\nu_2$ be the counting measure (Example 1.1). A function f on $\Omega_1\times\Omega_2$ defines a double sequence. If $\int f d\nu_1\times\nu_2$ exists, then

$$\int f d\nu_1\times\nu_2 = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} f(i,j) = \sum_{j=0}^{\infty}\sum_{i=0}^{\infty} f(i,j) \tag{1.16}$$

(by Theorem 1.3 and Example 1.5). Thus, a double series can be summed in either order, if it is well defined. $\blacksquare$

1.2.2 Radon-Nikodym derivative

Let $(\Omega,\mathcal{F},\nu)$ be a measure space and f be a nonnegative Borel function. One can show that the set function

$$\lambda(A)=\int_A f d\nu, \quad A\in\mathcal{F} \tag{1.17}$$

is a measure on $(\Omega,\mathcal{F})$. Note that

$$\nu(A)=0 \quad \text{implies} \quad \lambda(A)=0. \tag{1.18}$$

If (1.18) holds for two measures λ and ν defined on the same measurable space, then we say λ is *absolutely continuous* w.r.t. ν, and write $\lambda\ll\nu$.

Formula (1.17) gives us not only a way of constructing measures, but also a method of computing measures of measurable sets. Let ν be a well-known measure (such as the Lebesgue measure or the counting measure) and λ a relatively unknown measure. If we can find a function f such that (1.17) holds, then computing $\lambda(A)$ can be done through integration. A necessary condition for (1.17) is clearly $\lambda\ll\nu$. The following result shows that $\lambda\ll\nu$ is also almost sufficient for (1.17).

Theorem 1.4 (Radon-Nikodym theorem). Let ν and λ be two measures on $(\Omega,\mathcal{F})$ and ν be σ-finite. If $\lambda\ll\nu$, then there exists a nonnegative Borel

function f on Ω such that (1.17) holds. Furthermore, f is unique a.e. ν, i.e., if $\lambda(A) = \int_A g d\nu$ for any $A \in \mathcal{F}$, then $f = g$ a.e. ν. ∎

The proof of this theorem is beyond our scope. If (1.17) holds, then the function f is called the Radon-Nikodym *derivative* or *density* of λ w.r.t. ν, and is denoted by $d\lambda/d\nu$.

A useful consequence of Theorem 1.4 is that if f is Borel on $(\Omega, \mathcal{F})$ and $\int_A f d\nu = 0$ for any $A \in \mathcal{F}$, then $f = 0$ a.e.

If $\int f d\nu = 1$ for an $f \geq 0$ a.e. ν, then λ given by (1.17) is a probability measure and f is called its *probability density function* (p.d.f.) w.r.t. ν. For any probability measure P on $\mathcal{R}^k$ corresponding to a c.d.f. F or a random vector X, if P has a p.d.f. f w.r.t. a measure ν, then f is also called the p.d.f. of F or X w.r.t. ν.

Example 1.10 (p.d.f. of a discrete c.d.f.). Consider the discrete c.d.f. F in (1.9) of Example 1.3 with its probability measure given by (1.10). Let $\Omega = \{a_1, a_2, ...\}$ and ν be the counting measure on the power set of Ω. By Example 1.5,

$$P(A) = \int_A f d\nu = \sum_{a_i \in A} f(a_i), \quad A \subset \Omega, \tag{1.19}$$

where $f(a_i) = p_i$, $i = 1, 2, ...$ That is, f is the p.d.f. of P or F w.r.t. ν. Hence any discrete c.d.f. has a p.d.f. w.r.t. counting measure. A p.d.f. w.r.t. counting measure is called a *discrete* p.d.f. ∎

Example 1.11. Let F be a c.d.f. Assume that F is differentiable in the usual sense in calculus. Let f be the derivative of F. From calculus,

$$F(x) = \int_{-\infty}^{x} f(y) dy, \quad x \in \mathcal{R}. \tag{1.20}$$

Let P be the probability measure corresponding to F. It can be shown that $P(A) = \int_A f dm$ for any $A \in \mathcal{B}$, where m is the Lebesgue measure on $\mathcal{R}$. Hence, f is the p.d.f. of P or F w.r.t. Lebesgue measure. In this case, the Radon-Nikodym derivative is the same as the usual derivative of F in calculus.

A continuous c.d.f. may not have a p.d.f. w.r.t. Lebesgue measure. A necessary and sufficient condition for a c.d.f. F having a p.d.f. w.r.t. Lebesgue measure is that F is *absolute continuous* in the sense that for any $\epsilon > 0$, there exists $\delta > 0$ such that for each finite collection of disjoint bounded open intervals (a_i, b_i), $\sum(b_i - a_i) < \delta$ implies $\sum[F(b_i) - F(a_i)] < \epsilon$. Absolute continuity is weaker than differentiability, but is stronger than continuity. Thus, any discontinuous c.d.f. (such as a discrete c.d.f.) is not absolute continuous. Note that every c.d.f. is differentiable a.e. Lebesgue

measure (Chung, 1974, Chapter 1). Hence, if f is the p.d.f. of F w.r.t. Lebesgue measure, then $f =$ the usual derivative of F a.e. Lebesgue measure and (1.20) holds. In such a case probabilities can be computed through integration. It can be shown that the uniform and exponential c.d.f.'s in Example 1.4 are absolute continuous and their p.d.f.'s are, respectively,

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

and

$$f(x) = \begin{cases} 0 & -\infty < x < 0 \\ \theta^{-1} e^{-x/\theta} & 0 \le x < \infty. \end{cases}$$

A p.d.f. w.r.t. Lebesgue measure is called a Lebesgue p.d.f. ∎

More examples of p.d.f.'s are given in §1.3.1.

The following result provides some basic properties of Radon-Nikodym derivatives.

Proposition 1.7 (Calculus with Radon-Nikodym derivatives). Let ν be a σ-finite measure on a measure space $(\Omega, \mathcal{F})$. All other measures discussed in (i)-(iii) are defined on $(\Omega, \mathcal{F})$.
(i) If λ is a measure, $\lambda \ll \nu$, and $f \ge 0$, then

$$\int f d\lambda = \int f \frac{d\lambda}{d\nu} d\nu.$$

(Notice how the $d\nu$'s "cancel" on the right-hand side.)
(ii) If λ_i, $i = 1, 2$, are measures and $\lambda_i \ll \nu$, then $\lambda_1 + \lambda_2 \ll \nu$ and

$$\frac{d(\lambda_1 + \lambda_2)}{d\nu} = \frac{d\lambda_1}{d\nu} + \frac{d\lambda_2}{d\nu} \quad \text{a.e. } \nu.$$

(iii) (Chain rule). If τ is a measure, λ is a σ-finite measure, and $\tau \ll \lambda \ll \nu$, then

$$\frac{d\tau}{d\nu} = \frac{d\tau}{d\lambda} \frac{d\lambda}{d\nu} \quad \text{a.e. } \nu.$$

In particular, if $\lambda \ll \nu$ and $\nu \ll \lambda$ (in which case λ and ν are *equivalent*), then

$$\frac{d\lambda}{d\nu} = \left(\frac{d\nu}{d\lambda} \right)^{-1} \quad \text{a.e. } \nu \text{ or } \lambda.$$

(iv) Let $(\Omega_i, \mathcal{F}_i, \nu_i)$ be a measure space and ν_i be σ-finite, $i = 1, 2$. Let λ_i be a measure on $(\Omega_i, \mathcal{F}_i)$ and $\lambda_i \ll \nu_i$, $i = 1, 2$. Then $\lambda_1 \times \lambda_2 \ll \nu_1 \times \nu_2$ and

$$\frac{d(\lambda_1 \times \lambda_2)}{d(\nu_1 \times \nu_2)}(\omega_1, \omega_2) = \frac{d\lambda_1}{d\nu_1}(\omega_1) \frac{d\lambda_2}{d\nu_2}(\omega_2) \quad \text{a.e. } \nu_1 \times \nu_2. \quad ∎$$

1.3 Distributions and Their Characteristics

We now discuss some distributions useful in statistics, and their moments and generating functions.

1.3.1 Useful probability densities

It is often more convenient to work with p.d.f.'s than to work with c.d.f.'s. We now introduce some p.d.f.'s useful in statistics.

Most discrete p.d.f.'s are w.r.t. counting measure on the space of all nonnegative integers. Table 1.1 lists all discrete p.d.f.'s in elementary probability textbooks. For any discrete p.d.f. f, its c.d.f. $F(x)$ can be obtained using (1.19) with $A = (\infty, x]$. Values of $F(x)$ can be obtained from statistical tables or software.

Two Lebesgue p.d.f.'s are introduced in Example 1.11. Some other useful Lebesgue p.d.f.'s are listed in Table 1.2. Note that the exponential p.d.f. in Example 1.11 is a special case of that in Table 1.2 with $a = 0$. For any Lebesgue p.d.f., (1.20) gives its c.d.f. A few c.d.f.'s have explicit forms, whereas many others do not and they have to be evaluated numerically or computed using tables or software.

There are p.d.f.'s that are neither discrete nor Lebesgue.

Example 1.12. Let X be a random variable on $(\Omega, \mathcal{F}, P)$ whose c.d.f. F_X has a Lebesgue p.d.f. f_X and $F_X(c) < 1$, where c is a fixed constant. Let $Y = \min(X, c)$, i.e., Y is the smaller of X and c. Note that $Y^{-1}((-\infty, x]) = \Omega$ if $x \geq c$ and $Y^{-1}((-\infty, x]) = X^{-1}((\infty, x])$ if $x < c$. Hence Y is a random variable and the c.d.f. of Y is

$$F_Y(x) = \begin{cases} 1 & x \geq c \\ F_X(x) & x < c. \end{cases}$$

This c.d.f. is discontinuous at c, since $F(c) < 1$. Thus, it does not have a Lebesgue p.d.f. It is not discrete either. Does P_Y, the probability measure corresponding to F_Y, has a p.d.f. w.r.t. some measure? Define a probability measure on $(\mathcal{R}, \mathcal{B})$, called *point mass* at c, by

$$\delta_c(A) = \begin{cases} 1 & c \in A \\ 0 & c \notin A, \end{cases} \qquad A \in \mathcal{B} \tag{1.21}$$

(which is a special case of the discrete uniform distribution in Table 1.1). Then $P_Y \ll m + \delta_c$, where m is the Lebesgue measure, and the p.d.f. of P_Y is

$$\frac{dP_Y}{d(m+\delta_c)}(x) = \begin{cases} 0 & x > c \\ 1 - F_X(c) & x = c \\ f_X(x) & x < c. \end{cases} \qquad \blacksquare \tag{1.22}$$

Table 1.1. Discrete Distributions on $\mathcal{R}$

Uniform	p.d.f.	$1/m,\ x = a_1, ..., a_m$
	m.g.f.	$\sum_{j=1}^{m} e^{a_j t}/m,\ \ t \in \mathcal{R}$
$DU(a_1, ..., a_m)$	Expectation	$\sum_{j=1}^{m} a_j/m$
	Variance	$\sum_{j=1}^{m}(a_j - \bar{a})^2/m,\ \bar{a} = \sum_{j=1}^{m} a_j/m$
	Parameter	$a_i \in \mathcal{R},\ m = 1, 2, ...$
Binomial	p.d.f.	$\binom{n}{x} p^x (1-p)^{n-x},\ \ x = 0, 1, ..., n$
	m.g.f.	$(pe^t + 1 - p)^n,\ \ t \in \mathcal{R}$
$Bi(p, n)$	Expectation	np
	Variance	$np(1-p)$
	Parameter	$p \in [0, 1],\ \ n = 1, 2, ...$
Poisson	p.d.f.	$\theta^x e^{-\theta}/x!,\ \ x = 0, 1, 2, ...$
	m.g.f.	$e^{\theta(e^t - 1)},\ \ t \in \mathcal{R}$
$P(\theta)$	Expectation	θ
	Variance	θ
	Parameter	$\theta > 0$
Geometric	p.d.f.	$(1-p)^{x-1}p,\ \ x = 1, 2, ...$
	m.g.f.	$pe^t/[1 - (1-p)e^t],\ \ t < -\log(1-p)$
$G(p)$	Expectation	$1/p$
	Variance	$(1-p)/p^2$
	Parameter	$p \in [0, 1]$
Hyper-	p.d.f.	$\binom{n}{x}\binom{m}{r-x} / \binom{N}{r}$
geometric		$x = 0, 1, ..., \min(r, n),\ \ r - x \le m$
	m.g.f.	No explicit form
$HG(r, n, m)$	Expectation	rn/N
	Variance	$rnm(N-r)/[N^2(N-1)]$
	Parameter	$r, n, m = 1, 2, ...,\ N = n + m$
Negative	p.d.f.	$\binom{x-1}{r-1} p^r (1-p)^{x-r},\ \ x = r, r+1, ...$
binomial	m.g.f.	$p^r e^{rt}/[1 - (1-p)e^t]^r,\ \ t < -\log(1-p)$
	Expectation	r/p
$NB(p, r)$	Variance	$r(1-p)/p^2$
	Parameter	$p \in [0, 1],\ \ r = 1, 2, ...$
Log-	p.d.f.	$-(\log p)^{-1} x^{-1} (1-p)^x,\ \ x = 1, 2, ...$
distribution	m.g.f.	$\log[1 - (1-p)e^t]/\log p,\ \ t \in \mathcal{R}$
	Expectation	$-(1-p)/(p \log p)$
$L(p)$	Variance	$-(1-p)[1 + (1-p)/\log p]/(p^2 \log p)$
	Parameter	$p \in (0, 1)$

All p.d.f.'s are w.r.t. counting measure.

The random variable Y in Example 1.12 is a transformation of the random variable X. Transformations of random variables or vectors are frequently used in probability and statistics. For a random variable or vector X, $f(X)$ is a random variable or vector as long as f is measurable (Proposition 1.4). How do we find the c.d.f. (or p.d.f.) of $f(X)$ when the c.d.f. (or p.d.f.) of X is known? In many cases, the most effective method is direct computation. Example 1.12 is one example. The following is another one.

Example 1.13. Let X be a random variable with c.d.f. F_X and Lebesgue p.d.f. f_X, and let $Y = X^2$. Note that $Y^{-1}((-\infty, x]) = \emptyset$ if $x < 0$ and $Y^{-1}((-\infty, x]) = Y^{-1}([0, x]) = X^{-1}([-\sqrt{x}, \sqrt{x}\,])$ if $x \geq 0$. Hence

$$\begin{aligned} F_Y(x) &= P \circ Y^{-1}((-\infty, x]) \\ &= P \circ X^{-1}([-\sqrt{x}, \sqrt{x}\,]) \\ &= F_X(\sqrt{x}) - F_X(-\sqrt{x}) \end{aligned}$$

if $x \geq 0$ and $F_Y(x) = 0$ if $x < 0$. Clearly, the Lebesgue p.d.f. of F_Y is

$$f_Y(x) = \frac{1}{2\sqrt{x}}[f_X(\sqrt{x}) + f_X(-\sqrt{x})]I_{(0,\infty)}(x). \tag{1.23}$$

In particular, if

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \tag{1.24}$$

which is the Lebesgue p.d.f. for the normal distribution $N(0, 1)$ (Table 1.2), then

$$f_Y(x) = \frac{1}{\sqrt{2\pi x}} e^{-x/2} I_{(0,\infty)}(x),$$

which is the Lebesgue p.d.f. for the chi-square distribution χ^2_1 (Table 1.2). This is actually an important result in statistics. ∎

In some cases one may apply the following general result.

Proposition 1.8. Let X be a random k-vector with a Lebesgue p.d.f. f_X and let $Y = g(X)$, where g is a Borel function from $(\mathcal{R}^k, \mathcal{B}^k)$ to $(\mathcal{R}^k, \mathcal{B}^k)$. Let $A_1, ..., A_m$ be disjoint subsets of $\mathcal{R}^k$ such that $\mathcal{R}^k - (A_1 \cup \cdots \cup A_m)$ has Lebesgue measure 0 and g on A_j is one-to-one with a nonvanishing Jacobian, i.e., $\text{Det}(\partial g(x)/\partial x) \neq 0$ on A_j, $j = 1, ..., m$, where $\text{Det}(M)$ is the determinant of a square matrix M. Then Y has the following Lebesgue p.d.f.:

$$f_Y(x) = \sum_{j=1}^{m} \left|\text{Det}\left(\partial h_j(x)/\partial x\right)\right| f_X\left(h_j(x)\right),$$

where h_j is the inverse function of g on A_j, $j = 1, ..., m$. ∎

Table 1.2. Distributions on $\mathcal{R}$ with Lebesgue p.d.f.'s

Uniform	p.d.f.	$(b-a)^{-1}I_{(a,b)}(x)$
	m.g.f.	$(e^{bt}-e^{at})/(b-a),\ \ t \in \mathcal{R}$
$U(a,b)$	Expectation	$(a+b)/2$
	Variance	$(b-a)^2/12$
	Parameter	$a,\ b \in \mathcal{R},\ \ a<b$
Normal	p.d.f.	$\frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2}$
	m.g.f.	$e^{\mu t-\sigma^2 t^2/2},\ \ t \in \mathcal{R}$
$N(\mu,\sigma^2)$	Expectation	μ
	Variance	σ^2
	Parameter	$\mu \in \mathcal{R},\ \ \sigma>0$
Exponential	p.d.f.	$\theta^{-1}e^{-(x-a)/\theta}I_{(a,\infty)}(x)$
	m.g.f.	$e^{at}(1-\theta t)^{-1},\ \ t<\theta^{-1}$
$E(a,\theta)$	Expectation	$\theta+a$
	Variance	θ^2
	Parameter	$\theta>0,\ \ a \in \mathcal{R}$
Chi-square	p.d.f.	$\frac{1}{\Gamma(k/2)2^{k/2}}x^{k/2-1}e^{-x/2}I_{(0,\infty)}(x)$
	m.g.f.	$(1-2t)^{-k/2},\ \ t<1/2$
χ^2_k	Expectation	k
	Variance	$2k$
	Parameter	$k=1,2,...$
Gamma	p.d.f.	$\frac{1}{\Gamma(\alpha)\gamma^\alpha}x^{\alpha-1}e^{-x/\gamma}I_{(0,\infty)}(x)$
	m.g.f.	$(1-\gamma t)^{-\alpha},\ \ t<\gamma^{-1}$
$\Gamma(\alpha,\gamma)$	Expectation	$\alpha\gamma$
	Variance	$\alpha\gamma^2$
	Parameter	$\gamma>0,\ \alpha>0$
Beta	p.d.f.	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}I_{(0,1)}(x)$
	m.g.f.	No explicit form
$B(\alpha,\beta)$	Expectation	$\alpha/(\alpha+\beta)$
	Variance	$\alpha\beta/[(\alpha+\beta+1)(\alpha+\beta)^2]$
	Parameter	$\alpha>0,\ \ \beta>0$
Cauchy	p.d.f.	$\frac{1}{\pi\sigma}\left[1+\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-1}$
	m.g.f.	Does not exist
$C(\mu,\sigma)$	Expectation	Does not exist
	Variance	Does not exist
	ch.f.	$e^{\sqrt{-1}\mu t-\sigma\lvert t\rvert}$
	Parameter	$\mu \in \mathcal{R},\ \ \sigma>0$

Table 1.2. (continued)

t-distribution	p.d.f.	$\frac{\Gamma[(n+1)/2]}{\sqrt{n\pi}\Gamma(n/2)}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2}$
	m.g.f.	No explicit form
t_n	Expectation	$0, \ \ (n>1)$
	Variance	$n/(n-2), \ \ (n>2)$
	Parameter	$n=1,2,...$
F-distribution	p.d.f.	$\frac{n^{n/2}m^{m/2}\Gamma[(n+m)/2]x^{n/2-1}}{\Gamma(n/2)\Gamma(m/2)(m+nx)^{(n+m)/2}}I_{(0,\infty)}(x)$
	m.g.f.	No explicit form
$F_{n,m}$	Expectation	$m/(m-2), \ \ (m>2)$
	Variance	$2m^2(n+m-2)/[n(m-2)^2(m-4)],$ $(m>4)$
	Parameter	$n=1,2,..., \ \ m=1,2,...$
Log-normal	p.d.f.	$\frac{1}{\sqrt{2\pi}\sigma}x^{-1}e^{-(\log x-\mu)^2/2\sigma^2}I_{(0,\infty)}(x)$
	m.g.f.	Does not exist
$LN(\mu,\sigma^2)$	Expectation	$e^{\mu+\sigma^2/2}$
	Variance	$e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$
	Parameter	$\mu\in\mathcal{R}, \ \ \sigma>0$
Weibull	p.d.f.	$\frac{\alpha}{\theta}x^{\alpha-1}e^{-x^\alpha/\theta}I_{(0,\infty)}(x)$
	m.g.f.	No explicit form
$W(\alpha,\theta)$	Expectation	$\theta^{1/\alpha}\Gamma(\alpha^{-1}+1)$
	Variance	$\theta^{2/\alpha}\left[\Gamma(2\alpha^{-1}+1)-\Gamma(\alpha^{-1}+1)\right]^2$
	Parameter	$\theta>0, \ \ \alpha>0$
Double	p.d.f.	$\frac{1}{2\theta}e^{-\lvert x-\mu\rvert/\theta}$
Exponential	m.g.f.	$e^{\mu t}/(1+\theta^2t^2), \ \ t\in\mathcal{R}$
	Expectation	μ
$DE(\mu,\theta)$	Variance	$2\theta^2$
	Parameter	$\mu\in\mathcal{R}, \ \ \theta>0$
Pareto	p.d.f.	$\theta a^\theta x^{-(\theta+1)}I_{(a,\infty)}(x)$
	m.g.f.	No explicit form
$Pa(a,\theta)$	Expectation	$\theta a/(\theta-1), \ \ (\theta>1)$
	Variance	$\theta a^2/[(\theta-1)^2(\theta-2)], \ \ (\theta>2)$
	Parameter	$\theta>0, \ \ a>0$
Logistic	p.d.f.	$\sigma^{-1}e^{-(x-\mu)/\sigma}/[1+e^{-(x-\mu)/\sigma}]^2$
	m.g.f.	$e^{\mu t}\Gamma(1+\sigma t)\Gamma(1-\sigma t), \ \lvert t\rvert<\sigma$
$LG(\mu,\sigma)$	Expectation	μ
	Variance	$\sigma^2\pi^2/3$
	Parameter	$\mu\in\mathcal{R}, \ \sigma>0$

One may apply Proposition 1.8 to obtain result (1.23) in Example 1.13, using $A_1 = (-\infty, 0)$, $A_2 = (0, \infty)$, and $g(x) = x^2$. Note that $h_1(x) = -\sqrt{x}$, $h_2(x) = \sqrt{x}$, and $|dh_j(x)/dx| = 1/(2\sqrt{x})$.

A p.d.f. corresponding to a joint c.d.f. is called a joint p.d.f. The following is an important joint Lebesgue p.d.f. in statistics:

$$f(x) = (2\pi)^{-k/2}[\text{Det}(\Sigma)]^{-1/2} e^{-(x-\mu)\Sigma^{-1}(x-\mu)^\tau/2}, \quad x \in \mathcal{R}^k, \tag{1.25}$$

where $\mu \in \mathcal{R}^k$ is a vector of parameters, Σ is a positive definite $k \times k$ matrix of parameters, and A^τ denotes the transpose of a vector or matrix A. The p.d.f. in (1.25) and its c.d.f. are called the k-dimensional multivariate normal p.d.f. and c.d.f. and both are denoted by $N_k(\mu, \Sigma)$. The normal distribution $N(\mu, \sigma^2)$ in Table 1.2 is a special case of $N_k(\mu, \Sigma)$ with $k = 1$. The p.d.f. in (1.24) is $N(0, 1)$ and is called the *standard* normal p.d.f. Sometimes random vectors having the $N_k(\mu, \Sigma)$ distribution are also denoted by $N_k(\mu, \Sigma)$ for convenience. Useful properties of multivariate normal distributions can be found in Exercise 51.

Let X be a random k-vector having a c.d.f. F_X. Then the ith component of X is a random variable X_i having the following c.d.f.:

$$F_{X_i}(x) = \lim_{x_j \to \infty, j=1,...,i-1,i+1,...,k} F_X(x_1, ..., x_{i-1}, x, x_{i+1}, ..., x_k),$$

which is called the *marginal* c.d.f. of X_i. That is, the k marginal c.d.f.'s are determined by the joint c.d.f. If F_X has a Lebesgue p.d.f. f_X, then X_i has the following Lebesgue p.d.f.:

$$f_{X_i}(x) = \int \cdots \int f_X(x_1, ..., x_{i-1}, x, x_{i+1}, ..., x_k) dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_k.$$

In general, a joint c.d.f. cannot be determined by k marginal c.d.f.'s. There is one special but important case in which the joint c.d.f. of a random k-vector is determined by its k marginal c.d.f.'s, i.e.,

$$F_X(x_1, ..., x_k) = F_{X_1}(x_1) \cdots F_{X_k}(x_k), \quad (x_1, ..., x_k) \in \mathcal{R}^k. \tag{1.26}$$

If (1.26) holds, then random variables $X_1, ..., X_k$ are said to be *independent*. The meaning of independence is further discussed in §1.4.2. If each X_i has a Lebesgue p.d.f. f_{X_i}, then $X_1, ..., X_k$ are independent if and only if the joint p.d.f. of X satisfies

$$f_X(x_1, ..., x_k) = f_{X_1}(x_1) \cdots f_{X_k}(x_k), \quad (x_1, ..., x_k) \in \mathcal{R}^k. \tag{1.27}$$

Example 1.14. Let $X = (X_1, X_2)$ be a random 2-vector having a joint Lebesgue p.d.f. f_X. Consider first the transformation $g(x) = (x_1, x_1 + x_2)$. Using Proposition 1.8, one can show that the joint p.d.f. of $g(X)$ is

$$f_{g(X)}(x_1, y) = f_X(x_1, y - x_1),$$

where $y = x_1 + x_2$ (note that the Jacobian equals 1). The marginal p.d.f. of $Y = X_1 + X_2$ is then

$$f_Y(y) = \int f_X(x_1, y - x_1)dx_1.$$

In particular, if X_1 and X_2 are independent, then

$$f_Y(y) = \int f_{X_1}(x_1)f_{X_2}(y - x_1)dx_1. \tag{1.28}$$

Next, consider the transformation $h(x_1, x_2) = (x_1/x_2, x_2)$, assuming that $X_2 \neq 0$ a.s. Using Proposition 1.8, one can show that the joint p.d.f. of $h(X)$ is

$$f_{h(X)}(z, x_2) = |x_2|f_X(zx_2, x_2),$$

where $z = x_1/x_2$. The marginal p.d.f. of $Z = X_1/X_2$ is

$$f_Z(z) = \int |x_2|f_X(zx_2, x_2)dx_2.$$

In particular, if X_1 and X_2 are independent, then

$$f_Z(z) = \int |x_2|f_{X_1}(zx_2)f_{X_2}(x_2)dx_2. \quad \blacksquare \tag{1.29}$$

A number of results can be derived from (1.28) and (1.29). For example, if X_1 and X_2 are independent and both have the standard normal p.d.f. given by (1.24), then, by (1.29), the Lebesgue p.d.f. of $Z = X_1/X_2$ is

$$\begin{aligned} f_Z(z) &= \frac{1}{2\pi}\int |x_2|e^{-(1+z^2)x_2^2/2}dx_2 \\ &= \frac{1}{\pi}\int_0^\infty e^{-(1+z^2)x}dx \\ &= \frac{1}{\pi(1+z^2)}, \end{aligned}$$

which is the p.d.f. of the Cauchy distribution $C(0, 1)$ in Table 1.2. Another application of formula (1.29) leads to the following important result in statistics.

Example 1.15 (t-distribution and F-distribution). Let X_1 and X_2 be independent random variables having the chi-square distributions $\chi^2_{n_1}$ and $\chi^2_{n_2}$ (Table 1.2), respectively. By (1.29), the p.d.f. of $Z = X_1/X_2$ is

$$\begin{aligned} f_Z(z) &= \frac{z^{n_1/2-1}I_{(0,\infty)}(z)}{2^{(n_1+n_2)/2}\Gamma(n_1/2)\Gamma(n_2/2)}\int_0^\infty x_2^{(n_1+n_2)/2-1}e^{-(1+z)x_2/2}dx_2 \\ &= \frac{\Gamma[(n_1+n_2)/2]}{\Gamma(n_1/2)\Gamma(n_2/2)}\frac{z^{n_1/2-1}}{(1+z)^{(n_1+n_2)/2}}I_{(0,\infty)}(z), \end{aligned}$$

where the last equality follows from the fact that

$$\frac{1}{2^{(n_1+n_2)/2}\Gamma[(n_1+n_2)/2]}x_2^{(n_1+n_2)/2-1}e^{-x_2/2}I_{(0,\infty)}(x_2)$$

is the p.d.f. of the chi-square distribution $\chi^2_{n_1+n_2}$. Using Proposition 1.8, one can show that the p.d.f. of $Y = (X_1/n_1)/(X_2/n_2) = (n_2/n_1)Z$ is the p.d.f. of the F-distribution F_{n_1,n_2} given in Table 1.2.

Let U_1 be a random variable having the standard normal distribution $N(0,1)$ and U_2 a random variable having the chi-square distribution χ^2_n. Using the same argument, one can show that if U_1 and U_2 are independent, then the distribution of $T = U_1/\sqrt{U_2/n}$ is the t-distribution t_n given in Table 1.2. This result can also be derived using the result given in this example as follows. Let $X_1 = U_1^2$ and $X_2 = U_2$. Then X_1 and X_2 are independent (which can be shown directly, but follows immediately from Proposition 1.13 in §1.4.2). By Example 1.13, the distribution of X_1 is χ^2_1. Then $Y = X_1/(X_2/n)$ has the F-distribution $F_{1,n}$ and its Lebesgue p.d.f. is

$$\frac{n^{n/2}\Gamma[(n+1)/2]x^{-1/2}}{\sqrt{n\pi}\Gamma(n/2)(n+x)^{(n+1)/2}}I_{(0,\infty)}(x).$$

Note that

$$T = \begin{cases} \sqrt{Y} & U_1 \geq 0 \\ -\sqrt{Y} & U_1 < 0. \end{cases}$$

The result follows from Proposition 1.8 and the fact that

$$P \circ T^{-1}((-\infty, -t]) = P \circ T^{-1}([t, \infty)), \quad t > 0. \quad \blacksquare \tag{1.30}$$

If a random variable T satisfies (1.30), then T and its c.d.f. and p.d.f. (if it exists) are said to be *symmetric* about 0. If T has a Lebesgue p.d.f. f_T, then T is symmetric about 0 if and only if $f_T(x) = f_T(-x)$ for any $x > 0$. T and its c.d.f. and p.d.f. are said to be symmetric about a (or symmetric for simplicity) if and only if $T - a$ is symmetric about 0 for a fixed $a \in \mathcal{R}$. The c.d.f.'s of t-distributions are symmetric about 0 and the normal, Cauchy, and double exponential c.d.f.'s are symmetric.

The chi-square, t-, and F-distributions in the previous examples are special cases of the following *noncentral* chi-square, t-, and F-distributions, which are useful in some statistical problems.

Let $X_1, ..., X_n$ be independent random variables and $X_i = N(\mu_i, \sigma^2)$, $i = 1, ..., n$. The distribution of the random variable $Y = (X_1^2 + \cdots + X_n^2)/\sigma^2$ is called the *noncentral chi-square* distribution and denoted by $\chi^2_n(\delta)$, where $\delta = (\mu_1^2 + \cdots + \mu_n^2)/\sigma^2$ is the noncentrality parameter. It can be shown

(exercise) that Y has the following Lebesgue p.d.f.:

$$e^{-\delta/2}\sum_{j=0}^{\infty}\frac{\delta^j}{2^j j!}f_{2j+n}(x), \tag{1.31}$$

where $f_k(x)$ is the Lebesgue p.d.f. of the chi-square distribution χ^2_k. It is easy to see that the chi-square distribution χ^2_k in Table 1.2 is a special case of the noncentral chi-square distribution $\chi^2_k(\delta)$ with $\delta = 0$ and, therefore, is called a *central* chi-square distribution.

The result for the t-distribution in Example 1.15 can be extended to the case where U_1 has a nonzero expectation (U_2 still has the χ^2_n distribution and is independent of U_1). The distribution of $T = U_1/\sqrt{U_2/n}$ is called the *noncentral* t-distribution and denoted by $t_n(\delta)$, where $\delta = \mu$ is the noncentrality parameter. Using the same argument as that in Example 1.15, one can show (exercise) that T has the following Lebesgue p.d.f.:

$$\frac{1}{2^{(n+1)/2}\Gamma(n/2)\sqrt{\pi n}}\int_0^\infty y^{(n-1)/2}e^{-[(x\sqrt{y/n}-\delta)^2+y]/2}dy. \tag{1.32}$$

The t-distribution t_n in Example 1.15 is called a *central* t-distribution, since it is a special case of the noncentral t-distribution $t_n(\delta)$ with $\delta = 0$.

Similarly, the result for the F-distribution in Example 1.15 can be extended to the case where X_1 has the noncentral chi-square distribution $\chi^2_{n_1}(\delta)$, X_2 has the central chi-square distribution $\chi^2_{n_2}$, and X_1 and X_2 are independent. The distribution of $Y = (X_1/n_1)/(X_2/n_2)$ is called the *noncentral* F-distribution and denoted by $F_{n_1,n_2}(\delta)$, where δ is the noncentrality parameter. It can be shown (exercise) that Y has the following Lebesgue p.d.f.:

$$\frac{e^{-\delta/2}n_1^{n_1/2}n_2^{n_2/2}}{\Gamma(n_2/2)}\sum_{j=0}^{\infty}\frac{(\delta n_1 x/2)^j\Gamma((n_1+n_2)/2+j)x^{n_1/2-1}}{j!\Gamma(n_1/2+j)(n_1x+n_2)^{(n_1+n_2)/2+j}}I_{(0,\infty)}(x). \tag{1.33}$$

The F-distribution F_{n_1,n_2} in Example 1.15 is called a *central* F-distribution, since it is a special case of the noncentral F-distribution $F_{n_1,n_2}(\delta)$ with $\delta = 0$.

1.3.2 Moments and generating functions

We have defined the expectation of a random variable in §1.2.1. It is an important characteristic of a random variable. In this section we introduce other important *moments* and two *generating functions* of a random vector.

Let X be a random variable. If EX^k is finite, where k is a positive integer, then EX^k is called the kth *moment* of X (or the distribution of

X). If $E|X|^a < \infty$ for some real number a, then $E|X|^a$ is called the ath *absolute moment* of X (or the distribution of X). If $\mu = EX$ and $E(X-\mu)^k$ are finite for a positive integer k, then $E(X-\mu)^k$ is called the kth *central moment* of X (or the distribution of X).

The expectation and the second central moment (if they exist) are two important characteristics of a random variable (or its distribution) in statistics. They are listed in Tables 1.1 and 1.2 for those useful distributions. The expectation, also called the *mean* in statistics, is a measure of the central location of the distribution of a random variable. The second central moment, also called the *variance* in statistics, is a measure of dispersion or spread of a random variable. The variance of a random variable X is denoted by $\mathrm{Var}(X)$. The variance is always nonnegative. If the variance of X is 0, then X is equal to its mean a.s. (Proposition 1.6). The squared root of the variance is called the *standard deviation*, another important characteristic of a random variable in statistics.

The concept of mean and variance can be extended to random vectors. The expectation of a random matrix M with (i,j)th element M_{ij} is defined to be the matrix whose (i,j)th element is EM_{ij}. Thus, for a random k-vector $X = (X_1, ..., X_k)$, its mean is $EX = (EX_1, ..., EX_k)$; the extension of variance is the *variance-covariance matrix* of X defined as

$$\mathrm{Var}(X) = E(X - EX)^\tau (X - EX),$$

which is a $k \times k$ symmetric matrix whose diagonal elements are variances of X_i's. The (i,j)th element of $\mathrm{Var}(X)$, $i \neq j$, is $E(X_i - EX_i)(X_j - EX_j)$, which is called the *covariance* of X_i and X_j and is denoted by $\mathrm{Cov}(X_i, X_j)$.

Let $c = (c_1, ..., c_k) \in \mathcal{R}^k$ and $X = (X_1, ..., X_k)$ be a random k-vector. Then $Y = cX^\tau = c_1X_1 + \cdots + c_kX_k$ is a random variable, and

$$EY = c_1EX_1 + \cdots + c_kEX_k = cEX^\tau$$

and

$$\begin{aligned}\mathrm{Var}(Y) &= E(cX^\tau - cEX^\tau)^2 \\ &= E[c(X - EX)^\tau (X - EX)c^\tau] \\ &= c[E(X - EX)^\tau (X - EX)]c^\tau \\ &= c\mathrm{Var}(X)c^\tau,\end{aligned}$$

assuming that all expectations exist. Since $\mathrm{Var}(Y) \geq 0$ for any $c \in \mathcal{R}^k$, the matrix $\mathrm{Var}(X)$ is nonnegative definite. Consequently,

$$[\mathrm{Cov}(X_i, X_j)]^2 \leq \mathrm{Var}(X_i)\mathrm{Var}(X_j), \quad i \neq j. \tag{1.34}$$

An important quantity in statistics is the *correlation coefficient* defined to be $\rho_{X_i,X_j} = \mathrm{Cov}(X_i, X_j)/\sqrt{\mathrm{Var}(X_i)\mathrm{Var}(X_j)}$, which is, by inequality (1.34),

always between -1 and 1. It is a measure of relationship between X_i and X_j; if ρ_{X_i,X_j} is positive (or negative), then X_i and X_j tend to be positively (or negatively) related; if $\rho_{X_i,X_j} = \pm 1$, then $P(X_i = c_1 \pm c_2 X_j) = 1$ with some constants c_1 and $c_2 > 0$; if $\rho_{X_i,X_j} = 0$ (i.e., $\mathrm{Cov}(X_i, X_j) = 0$), then X_i and X_j are said to be *uncorrelated*. One can show that if X_i and X_j are independent, then they are uncorrelated. But the converse is not necessarily true. Examples can be found in Exercises 48-49.

The following result indicates that if the rank of $\mathrm{Var}(X)$ is $r < k$, then X is in a subspace of $\mathcal{R}^k$ with dimension r. For any $k \times k$ symmetric matrix M, define $R_M = \{y \in \mathcal{R}^k : y = xM \text{ with some } x \in \mathcal{R}^k\}$.

Proposition 1.9. Let X be a random k-vector with a finite $\mathrm{Var}(X)$. Then we have the following conclusions.
(i) $P(X - EX \in R_{\mathrm{Var}(X)}) = 1$.
(ii) If $P_X \ll$ Lebesgue measure on $\mathcal{R}^k$, then the rank of $\mathrm{Var}(X)$ is k. ∎

Example 1.16. Let X be a random k-vector having the $N_k(\mu, \Sigma)$ distribution. It can be shown (exercise) that $EX = \mu$ and $\mathrm{Var}(X) = \Sigma$. Thus, μ and Σ in (1.25) are the mean vector and the variance-covariance matrix of X. If Σ is a diagonal matrix (i.e., all components of X are uncorrelated), then by (1.27), the components of X are independent. This shows an important property of random variables having normal distributions: they are independent if and only if they are uncorrelated. ∎

Moments are important characteristics of a distribution, but they do not determine a distribution in the sense that two different distributions may have the same moments of all orders. Functions that determine a distribution are introduced in the following definition.

Definition 1.5. Let X be a random k-vector.
(i) The *moment generating function* (m.g.f.) of X (or P_X) is defined as

$$\psi_X(t) = Ee^{tX^\tau}, \quad t \in \mathcal{R}^k.$$

(ii) The *characteristic function* (ch.f.) of X (or P_X) is defined as

$$\phi_X(t) = Ee^{\sqrt{-1}tX^\tau} = E[\cos(tX^\tau)] + \sqrt{-1}\,E[\sin(tX^\tau)], \quad t \in \mathcal{R}^k. \quad ∎$$

The ch.f. is complex-valued and always well defined. The m.g.f. is nonnegative but may be ∞ everywhere except at $t = 0$. If the m.g.f. is finite at some $t \neq 0$, then $\phi_X(t)$ can be obtained by replacing t in $\psi_X(t)$ by $\sqrt{-1}t$. Tables 1.1 and 1.2 contain the m.g.f. (or ch.f. when the m.g.f. is ∞ everywhere except at 0) for distributions useful in statistics. Some useful properties of the m.g.f. and ch.f. are given in the following result.

Proposition 1.10. Let X be a random k-vector with m.g.f. $\psi_X(t)$ and ch.f. $\phi_X(t)$.
(i) (Relation to moments). If EX is finite, then

$$\left.\frac{\partial \phi_X(t)}{\partial t}\right|_{t=0} = \sqrt{-1}EX.$$

If $\mathrm{Var}(X)$ is finite, then

$$\left.\frac{\partial^2 \phi_X(t)}{\partial t \partial t^\tau}\right|_{t=0} = -E(X^\tau X).$$

If $k = 1$ and EX^p is finite for a positive integer p, then

$$\left.\frac{d^p \phi_X(t)}{dt^p}\right|_{t=0} = (-1)^{p/2} EX^p.$$

If $\psi_X(t) < \infty$ for $t \in N_\epsilon = \{t \in \mathcal{R}^k : tt^\tau \le \epsilon\}$, then the components of X have finite moments of all orders,

$$\left.\frac{\partial \psi_X(t)}{\partial t}\right|_{t=0} = EX,$$

$$\left.\frac{\partial^2 \psi_X(t)}{\partial t \partial t^\tau}\right|_{t=0} = E(X^\tau X),$$

and, when $k = 1$ and p is a positive integer,

$$\left.\frac{d^p \psi_X(t)}{dt^p}\right|_{t=0} = EX^p.$$

(ii) (Uniqueness). If Y is a random k-vector and $\phi_X(t) = \phi_Y(t)$ for all $t \in \mathcal{R}^k$, then $P_X = P_Y$. If there is an $\epsilon > 0$ such that $\psi_X(t) = \psi_Y(t) < \infty$ for all $t \in N_\epsilon = \{t \in \mathcal{R}^k : tt^\tau \le \epsilon\}$, then $P_X = P_Y$.
(iii) (Sums of independent random vectors). Let Y be a random k-vector independent of X. Then

$$\psi_{X+Y}(t) = \psi_X(t)\psi_Y(t), \quad t \in \mathcal{R}^k,$$

and the same result holds when ψ is replaced by ϕ.
(iv) (Linear transformations). Let $Y = XC^\tau + c$, where C is an $m \times k$ matrix and $c \in \mathcal{R}^m$. Then

$$\psi_Y(u) = e^{uc^\tau}\psi_X(uC), \quad u \in \mathcal{R}^m,$$

and

$$\phi_Y(u) = e^{\sqrt{-1}uc^\tau}\phi_X(uC), \quad u \in \mathcal{R}^m. \quad \blacksquare$$

Proposition 1.10(ii)-(iii) provides a useful tool to obtain distributions of sums of independent random vectors. The following example is an illustration.

Example 1.17. Let X_i, $i = 1, ..., k$, be independent random variables and X_i have the gamma distribution $\Gamma(\alpha_i, \gamma)$ (Table 1.2), $i = 1, ..., k$. From Table 1.2, X_i has the m.g.f. $\psi_{X_i}(t) = (1 - \gamma t)^{-\alpha_i}$, $t < \gamma^{-1}$, $i = 1, ..., k$. By Proposition 1.10(iii), the m.g.f. of $Y = X_1 + \cdots + X_k$ is equal to $\psi_Y(t) = (1 - \gamma t)^{-(\alpha_1 + \cdots + \alpha_k)}$, $t < \gamma^{-1}$. From Table 1.2, the gamma distribution $\Gamma(\alpha_1 + \cdots + \alpha_k, \gamma)$ has the m.g.f. $\psi_Y(t)$ and, hence, is the distribution of Y (by Proposition 1.10(ii)). ∎

Using Proposition 1.10 and a result from linear algebra, we can prove the following result useful in *analysis of variance* (Scheffé, 1959; Searle, 1971).

Theorem 1.5. (i) Suppose that $Y_1, ..., Y_k$ are independent random variables and that Y_i has the noncentral chi-square distribution $\chi^2_{n_i}(\delta_i)$, $i = 1, ..., k$. Then $Y = Y_1 + \cdots + Y_k$ has the noncentral chi-square distribution $\chi^2_{n_1 + \cdots + n_k}(\delta_1 + \cdots + \delta_k)$.
(ii) (Cochran's theorem). Suppose that $X = N_n(\mu, I_n)$ and

$$XX^\tau = XA_1X^\tau + \cdots + XA_kX^\tau, \tag{1.35}$$

where A_i is a nonnegative definite $n \times n$ matrix with rank n_i, $i = 1, ..., k$. Then a necessary and sufficient condition that XA_iX^τ has the noncentral chi-square distribution $\chi^2_{n_i}(\delta_i)$, $i = 1, ..., k$, and XA_iX^τ's are independent is $n = n_1 + \cdots + n_k$, in which case $\delta_i = \mu A_i \mu^\tau$ and $\delta_1 + \cdots + \delta_k = \mu\mu^\tau$.
Proof. (i) The ch.f. of Y_i is $(1 - 2\sqrt{-1}t)^{n_i/2} e^{\sqrt{-1}\delta_i t/(1 - 2\sqrt{-1}t)}$ (Exercise 53). Then, the result follows from Proposition 1.10(ii)-(iii).
(ii) The necessity follows from (i) and the fact that XX^τ has the noncentral chi-square distribution $\chi^2_n(\mu\mu^\tau)$ (by definition). We now prove the sufficiency.

Assume that $n = n_1 + \cdots + n_k$. We use the following fact from linear algebra: there exists an $n \times n$ matrix C such that $Y = XC$ and

$$XA_iX^\tau = \sum_{j = n_1 + \cdots + n_{i-1} + 1}^{n_1 + \cdots + n_{i-1} + n_i} Y_j^2, \tag{1.36}$$

where Y_j is the jth component of Y. From (1.35) and (1.36), $XX^\tau = YY^\tau$, i.e., $CC^\tau = I_n$. Thus, $Y = N_n(\mu C, I_n)$ (Exercise 51); the independence of XA_iX^τ follows from (1.36); and the fact that XA_iX^τ has the noncentral chi-square distribution follows directly from the definition of the noncentral chi-square distribution and (1.36). ∎

1.4 Conditional Expectations

In elementary probability the conditional probability of an event B given an event A is defined as $P(B|A) = P(A \cap B)/P(A)$, provided that $P(A) > 0$. In probability and statistics, however, we sometimes need a notion of "conditional probability" even for A's with $P(A) = 0$; for example, $A = \{Y = c\}$, where Y is a random variable and $c \in \mathcal{R}$. General definitions of conditional probability, expectation, and distribution are introduced in this section, and they are shown to agree with those defined in elementary probability in special cases.

1.4.1 Conditional expectations

Definition 1.6. Let X be an integrable random variable on $(\Omega, \mathcal{F}, P)$.
(i) Let $\mathcal{A}$ be a sub-σ-field of $\mathcal{F}$. The *conditional expectation* of X given $\mathcal{A}$, denoted by $E(X|\mathcal{A})$, is the a.s.-unique random variable satisfying the following two conditions:

(a) $E(X|\mathcal{A})$ is measurable from $(\Omega, \mathcal{A})$ to $(\mathcal{R}, \mathcal{B})$;
(b) $\int_A E(X|\mathcal{A})dP = \int_A X dP$ for any $A \in \mathcal{A}$.

(ii) Let $B \in \mathcal{F}$. The *conditional probability* of B given $\mathcal{A}$ is defined to be $P(B|\mathcal{A}) = E(I_B|\mathcal{A})$.
(iii) Let Y be measurable from $(\Omega, \mathcal{F}, P)$ to $(\Lambda, \mathcal{G})$. The conditional expectation of X given Y is defined to be $E(X|Y) = E[X|\sigma(Y)]$. ∎

Essentially, the σ-field $\sigma(Y)$ contains "the information in Y". Hence, $E(X|Y)$ is the "expectation" of X given the information provided by $\sigma(Y)$. The following useful result shows that there is a Borel function h defined on the range of Y such that $E(X|Y) = h \circ Y$.

Theorem 1.6. Let Y be measurable from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$ and Z a function from $(\Omega, \mathcal{F})$ to $\mathcal{R}^k$. Then Z is measurable from $(\Omega, \sigma(Y))$ to $(\mathcal{R}^k, \mathcal{B}^k)$ if and only if there is a measurable function h from $(\Lambda, \mathcal{G})$ to $(\mathcal{R}^k, \mathcal{B}^k)$ such that $Z = h \circ Y$. ∎

The function h in $E(X|Y) = h \circ Y$ is a Borel function on $(\Lambda, \mathcal{G})$. Let $y \in \Lambda$. Then we define

$$E(X|Y = y) = h(y)$$

to be the conditional expectation of X given $Y = y$. Note that $h(y)$ is a function on Λ, whereas $h \circ Y$ is a function on Ω.

Example 1.18. Let X be an integrable random variable on $(\Omega, \mathcal{F}, P)$, $A_1, A_2, ...$ be disjoint events on $(\Omega, \mathcal{F}, P)$ such that $\cup A_i = \Omega$ and $P(A_i) > 0$

for all i, and let $a_1, a_2, ...$ be distinct real numbers. Define $Y = a_1 I_{A_1} + a_2 I_{A_2} + \cdots$. We now show that

$$E(X|Y) = \sum_{i=1}^{\infty} \frac{\int_{A_i} X dP}{P(A_i)} I_{A_i}. \tag{1.37}$$

We need to verify (a) and (b) in Definition 1.6 with $\mathcal{A} = \sigma(Y)$. Since $\sigma(Y) = \sigma(\{A_1, A_2, ...\})$, it is clear that the function on the right-hand side of (1.37) is measurable on $(\Omega, \sigma(Y))$. For any $B \in \mathcal{B}$, $Y^{-1}(B) = \cup_{i:a_i \in B} A_i$. Using properties of integrals, we obtain that

$$\begin{aligned}\int_{Y^{-1}(B)} X dP &= \sum_{i:a_i \in B} \int_{A_i} X dP \\ &= \sum_{i=1}^{\infty} \frac{\int_{A_i} X dP}{P(A_i)} P\left(A_i \cap Y^{-1}(B)\right) \\ &= \int_{Y^{-1}(B)} \left[\sum_{i=1}^{\infty} \frac{\int_{A_i} X dP}{P(A_i)} I_{A_i}\right] dP.\end{aligned}$$

This verifies (b) and thus (1.37) holds.

Let $A \in \mathcal{F}$ and $X = I_A$. Then

$$P(A|Y) = E(X|Y) = \sum_{i=1}^{\infty} \frac{P(A \cap A_i)}{P(A_i)} I_{A_i}.$$

Note that $\{Y = a_i\} = \{\omega \in \Omega : Y(\omega) = a_i\} = A_i$. If $\omega \in A_i$,

$$P(A|Y)(\omega) = \frac{P(A \cap A_i)}{P(A_i)} = P(A|A_i) = P(A|\{Y = a_i\}).$$

Hence the definition of conditional probability in Definition 1.6 agrees with that in elementary probability. More generally, let X be a discrete random variable whose range is $\{c_1, c_2, ...\}$, where c_i's are distinct real numbers. Let $C_i = X^{-1}(\{c_i\})$, $i = 1, 2, ...$ Then, by (1.37),

$$E(X|Y) = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} c_j P(C_j|A_i) I_{A_i}.$$

If $\omega \in A_i$, then

$$E(X|Y)(\omega) = \sum_{j=1}^{\infty} c_j P(C_j|A_i) = \sum_{j=1}^{\infty} c_j P(C_j|\{Y = a_i\}),$$

which agrees with $E(X|Y = a_i)$ defined in elementary probability.

Let h be a Borel function on $\mathcal{R}$ satisfying $h(a_i) = \int_{A_i} X dP/P(A_i)$. Then, by (1.37), $E(X|Y) = h \circ Y$ and $E(X|Y = y) = h(y)$. ∎

The next result generalizes the result in Example 1.18 to conditional expectations of random variables having p.d.f's.

Proposition 1.11. Let X be a random n-vector and Y a random m-vector. Suppose that (X, Y) has a joint p.d.f. $f(x, y)$ w.r.t. $\nu \times \lambda$, where ν and λ are σ-finite measures on $(\mathcal{R}^n, \mathcal{B}^n)$ and $(\mathcal{R}^m, \mathcal{B}^m)$, respectively. Let $g(x, y)$ be a Borel function on $\mathcal{R}^{n+m}$ for which $E|g(X, Y)| < \infty$. Then

$$E[g(X,Y)|Y] = \frac{\int g(x,Y)f(x,Y)d\nu(x)}{\int f(x,Y)d\nu(x)} \quad \text{a.s.} \tag{1.38}$$

Proof. Denote the right-hand side of (1.38) by $h(Y)$. By Fubini's theorem, h is Borel. Then, by Theorem 1.6, $h(Y)$ is Borel on $(\Omega, \sigma(Y))$. Also, by Fubini's theorem, $f_Y(y) = \int f(x, y)d\nu(x)$ is the p.d.f. of Y w.r.t. λ. For $B \in \mathcal{B}^m$,

$$\begin{aligned}
\int_{Y^{-1}(B)} h(Y)dP &= \int_B h(y)dP_Y \\
&= \int_B \frac{\int g(x,y)f(x,y)d\nu(x)}{\int f(x,y)d\nu(x)} f_Y(y)d\lambda(y) \\
&= \int_{\mathcal{R}^n \times B} g(x,y)f(x,y)d\nu \times \lambda \\
&= \int_{\mathcal{R}^n \times B} g(X,Y)dP_{(X,Y)} \\
&= \int_{Y^{-1}(B)} g(X,Y)dP,
\end{aligned}$$

where the first and the last equalities follow from Theorem 1.2; the second and the next to last equalities follow from the definition of h and p.d.f.'s; and the third equality follows from Theorem 1.3 (Fubini's theorem). ∎

For a random vector (X, Y) with a joint p.d.f. $f(x, y)$ w.r.t. $\nu \times \lambda$, define the *conditional* p.d.f. of X given $Y = y$ to be

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)}, \tag{1.39}$$

where $f_Y(y) = \int f(x, y)d\nu(x)$ is the marginal p.d.f. of Y w.r.t. ν. One can easily check that for each fixed y with $f_Y(y) > 0$, $f_{X|Y}(x|y)$ in (1.39) is a p.d.f. w.r.t. ν. Then equation (1.38) can be rewritten as

$$E[g(X,Y)|Y] = \int g(x,Y)f_{X|Y}(x|Y)d\nu(x).$$

Again, this agrees with the conditional expectation defined in elementary probability (i.e., the conditional expectation of $g(X,Y)$ given Y is equal to the expectation of $g(X,Y)$ w.r.t. the conditional p.d.f. of X given Y).

Now we list some useful properties of conditional expectations. The proof is left to the reader.

Proposition 1.12. Let X and Y be integrable random variables on $(\Omega, \mathcal{F}, P)$ and $\mathcal{A}$ be a sub-σ-field of $\mathcal{F}$.
(i) If $X = c$ a.s., $c \in \mathcal{R}$, then $E(X|\mathcal{A}) = c$ a.s.
(ii) If $X \le Y$ a.s., then $E(X|\mathcal{A}) \le E(Y|\mathcal{A})$ a.s.
(iii) If a and b are real numbers, then $E(aX+bY|\mathcal{A}) = aE(X|\mathcal{A})+bE(Y|\mathcal{A})$ a.s.
(iv) $E[E(X|\mathcal{A})] = EX$.
(v) $E[E(X|\mathcal{A})|\mathcal{A}_0] = E(X|\mathcal{A}_0) = E[E(X|\mathcal{A}_0)|\mathcal{A}]$ a.s., where $\mathcal{A}_0$ is a sub-σ-field of $\mathcal{A}$.
(vi) If $\sigma(Y) \subset \mathcal{A}$ and $E|XY| < \infty$, then $E(XY|\mathcal{A}) = YE(X|\mathcal{A})$ a.s.
(vii) If $E|g(X,Y)| < \infty$, then $E[g(X,Y)|Y = y] = E[g(X,y)|Y = y]$ a.s.
(viii) If $EX^2 < \infty$, then $[E(X|Y)]^2 \le E(X^2|Y)$ a.s. ∎

As an application, we consider the following example.

Example 1.19. Let X be a random variable on $(\Omega, \mathcal{F}, P)$ with $EX^2 < \infty$ and Y a measurable function from $(\Omega, \mathcal{F}, P)$ to $(\Lambda, \mathcal{G})$. One may wish to predict the value of X based on an observed value of Y. Let $g(Y)$ be a predictor, i.e., $g \in \aleph = \{\text{all Borel functions } g \text{ with } E[g(Y)]^2 < \infty\}$. Each predictor is assessed by the "mean squared prediction error" $E[X - g(Y)]^2$. We now show that $E(X|Y)$ is the best predictor of X in the sense that

$$E[X - E(X|Y)]^2 = \min_{g \in \aleph} E[X - g(Y)]^2. \tag{1.40}$$

First, it follows from Proposition 1.12(viii) that $E(X|Y) \in \aleph$. Next, for any $g \in \aleph$,

$$\begin{aligned}
E[X - g(Y)]^2 &= E[X - E(X|Y) + E(X|Y) - g(Y)]^2 \\
&= E[X - E(X|Y)]^2 + E[E(X|Y) - g(Y)]^2 \\
&\quad + 2E[X - E(X|Y)][E(X|Y) - g(Y)] \\
&= E[X - E(X|Y)]^2 + E[E(X|Y) - g(Y)]^2 \\
&\quad + 2E\big\{E\{[X - E(X|Y)][E(X|Y) - g(Y)]|Y\}\big\} \\
&= E[X - E(X|Y)]^2 + E[E(X|Y) - g(Y)]^2 \\
&\quad + 2E\{[E(X|Y) - g(Y)]E[X - E(X|Y)|Y]\} \\
&= E[X - E(X|Y)]^2 + E[E(X|Y) - g(Y)]^2 \\
&\ge E[X - E(X|Y)]^2,
\end{aligned}$$

where the third equality follows from Proposition 1.12(iv); the fourth equality follows from Proposition 1.12(vi); and the last equality follows from Proposition 1.2(iii) and (vi). ∎

1.4.2 Independence

Definition 1.7. Let $(\Omega, \mathcal{F}, P)$ be a probability space.
(i) Events A_i, $i = 1, ..., n$, are said to be *independent* if and only if for *any* subset $\{i_1, i_2, ..., i_k\}$ of $\{1, ..., n\}$,

$$P(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) = P(A_{i_1})P(A_{i_2}) \cdots P(A_{i_k}).$$

Events in an infinite (countable or uncountable) collection $\mathcal{C}$ are said to be independent if and only if events in each finite subcollection of $\mathcal{C}$ are independent.
(ii) Collections $\mathcal{C}_i \subset \mathcal{F}$, $i \in \mathcal{I}$ (an index set that can be uncountable), are said to be independent if and only if events in any collection of the form $\{A_i \in \mathcal{C}_i : i \in \mathcal{I}\}$ are independent.
(iii) Random vectors X_i, $i \in \mathcal{I}$, are said to be independent if and only if $\sigma(X_i)$, $i \in \mathcal{I}$, are independent. ∎

It is easy to see from Definition 1.7 that if X and Y are independent random vectors, then so are $g(X)$ and $h(Y)$, where g and h are Borel functions.

The following result is useful for checking the independence of several σ-fields.

Proposition 1.13. Let $\mathcal{C}_i$, $i \in \mathcal{I}$, be independent collections of events. Suppose that each $\mathcal{C}_i$ has the property that if $A \in \mathcal{C}_i$ and $B \in \mathcal{C}_i$, then $A \cap B \in \mathcal{C}_i$. Then $\sigma(\mathcal{C}_i)$, $i \in \mathcal{I}$, are independent. ∎

An immediate application of Proposition 1.13 is to show (exercise) that random variables X_i, $i = 1, ..., n$, are independent if and only if (1.26) holds. Hence, Definition 1.7 agrees with the concept of independence discussed in §1.3.1.

Independent random variables can be obtained using product measures introduced in §1.1.2. Let P_i be a probability measure on $(\mathcal{R}, \mathcal{B})$, $i = 1, ..., k$. Then any random vector whose law is the product measure $P_1 \times \cdots \times P_k$ on the space $(\mathcal{R}^k, \mathcal{B}^k)$ has independent components. If F_i is the c.d.f. of P_i, then the joint c.d.f. of $P_1 \times \cdots \times P_k$ is $F(x_1, ..., x_k) = F_1(x_1) \cdots F_k(x_k)$. Consequently, by Fubini's theorem, we obtain that if $X_1, ..., X_n$ are independent random variables and $E|X_1 \cdots X_n| < \infty$, then

$$E(X_1 \cdots X_n) = EX_1 \cdots EX_n. \tag{1.41}$$

When $n = 2$, this implies that if $X_1, ..., X_n$ are independent, then X_i and X_j are uncorrelated, $i \neq j$.

For two events A and B with $P(A) > 0$, A and B are independent if and only if $P(B|A) = P(B)$. This means that A provides no information about B. The following result is a useful extension.

Proposition 1.14. Let X, Y_1, and Y_2 be random variables with $E|X| < \infty$. Suppose that (X, Y_1) and Y_2 are independent. Then

$$E[X|(Y_1, Y_2)] = E(X|Y_1) \text{ a.s.}$$

Proof. First, $E(X|Y_1)$ is Borel on $(\Omega, \sigma(Y_1, Y_2))$, since $\sigma(Y_1) \subset \sigma(Y_1, Y_2)$. Next, we need to show that for any Borel set $B \subset \mathcal{R}^2$,

$$\int_{(Y_1,Y_2)^{-1}(B)} X dP = \int_{(Y_1,Y_2)^{-1}(B)} E(X|Y_1) dP. \tag{1.42}$$

If $B = B_1 \times B_2$, where B_i's are Borel sets in $\mathcal{R}$, then $(Y_1, Y_2)^{-1}(B) = Y_1^{-1}(B_1) \cap Y_2^{-1}(B_2)$ and

$$\begin{aligned}
\int_{Y_1^{-1}(B_1)\cap Y_2^{-1}(B_2)} E(X|Y_1) dP &= \int I_{Y_1^{-1}(B_1)} I_{Y_2^{-1}(B_2)} E(X|Y_1) dP \\
&= \int I_{Y_1^{-1}(B_1)} E(X|Y_1) dP \int I_{Y_2^{-1}(B_2)} dP \\
&= \int I_{Y_1^{-1}(B_1)} X dP \int I_{Y_2^{-1}(B_2)} dP \\
&= \int I_{Y_1^{-1}(B_1)} I_{Y_2^{-1}(B_2)} X dP \\
&= \int_{Y_1^{-1}(B_1)\cap Y_2^{-1}(B_2)} X dP,
\end{aligned}$$

where the second and the next to last equalities follow from result (1.41) and the independence of (X, Y_1) and Y_2; and the third equality follows from the fact that $E(X|Y_1)$ is the conditional expectation of X given Y_1. This shows that (1.42) holds for $B = B_1 \times B_2$. Let $\mathcal{H}$ be the collection of subsets of $\mathcal{R}^2$ for which (1.42) holds. Then we can show that $\mathcal{H}$ is a σ-field. Since we have shown that $\mathcal{B} \times \mathcal{B} \subset \mathcal{H}$, $\mathcal{B}^2 = \sigma(\mathcal{B} \times \mathcal{B}) \subset \mathcal{H}$ and thus the result follows. ∎

Let $X_1, ..., X_k$ be random variables. If X_i and X_j are independent for every pair $i \neq j$, then $X_1, ..., X_k$ are said to be *pairwise independent*. If $X_1, ..., X_k$ are independent, then clearly they are pairwise independent. However, the converse is not true. The following is an example.

Example 1.20. Let X_1 and X_2 be independent random variables each assuming the values 1 and -1 with probability 0.5, and $X_3 = X_1X_2$. Let $A_i = \{X_i = 1\}$, $i = 1, 2, 3$. Then $P(A_i) = 0.5$ for any i and $P(A_1)P(A_2)P(A_3) = 0.125$. However, $P(A_1 \cap A_2 \cap A_3) = P(A_1 \cap A_2) = P(A_1)P(A_2) = 0.25$. Hence X_1, X_2, X_3 are not independent. We now show that X_1, X_2, X_3 are pairwise independent. It is enough to show that X_1 and X_3 are independent. Let $B_i = \{X_i = -1\}$, $i = 1, 2, 3$. Note that $A_1 \cap A_3 = A_1 \cap A_2$, $A_1 \cap B_3 = A_1 \cap B_2$, $B_1 \cap A_3 = B_1 \cap B_2$, and $B_1 \cap B_3 = B_1 \cap A_2$. Then the result follows from the fact that $P(A_i) = P(B_i) = 0.5$ for any i and X_1 and X_2 are independent. ∎

1.4.3 Conditional distributions

The conditional p.d.f. was introduced in §1.4.1 for random variables having p.d.f.'s w.r.t. some measures. We now consider *conditional distributions* in general cases where we may not have any p.d.f.

Theorem 1.7 (Existence of conditional distributions). Let X be a random n-vector on a probability space $(\Omega, \mathcal{F}, P)$ and Y be measurable from $(\Omega, \mathcal{F}, P)$ to $(\Lambda, \mathcal{G})$. Then there exists $P_{X|Y}(A|y)$ such that

(a) $P_{X|Y}(A|y) = P[X^{-1}(A)|Y = y]$ a.s. P_Y for any fixed $A \in \mathcal{B}^n$, and

(b) $P_{X|Y}(\cdot|y)$ is a probability measure on $(\mathcal{R}^n, \mathcal{B}^n)$ for any fixed $y \in \Lambda$.

Furthermore, if $E|g(X, Y)| < \infty$, then

$$E[g(X, Y)|Y = y] = \int_{\mathcal{R}^n} g(x, y) dP_{X|Y}(x|y) \text{ a.s. } P_Y. \quad ∎$$

For a fixed y, $P_{X|Y=y} = P_{X|Y}(\cdot|y)$ is called the conditional distribution of X given $Y = y$. Under the conditions in Theorem 1.7, if Y is a random m-vector and (X, Y) has a p.d.f. w.r.t. $\nu \times \lambda$ (ν and λ are σ-finite measures on $(\mathcal{R}^n, \mathcal{B}^n)$ and $(\mathcal{R}^m, \mathcal{B}^m)$, respectively), then $f_{X|Y}(x|y)$ defined in (1.39) is the p.d.f. of $P_{X|Y=y}$ w.r.t. ν for any fixed y.

Given a collection of conditional distributions $\{P_{X|Y=y} : y \in \Lambda\}$ and a distribution P_Y on Λ, we can construct a joint distribution, as the following result states.

Proposition 1.15. Let $\mathcal{B}^n$ be the Borel σ-field on $\mathcal{R}^n$ and $(\Lambda, \mathcal{G}, P_2)$ be a probability space. Suppose that P_1 is a function from $\mathcal{B}^n \times \Lambda$ to $\mathcal{R}$ and satisfies

(a) $P_1(\cdot, y)$ is a probability measure on $(\mathcal{R}^n, \mathcal{B}^n)$ for any $y \in \Lambda$, and

(b) $P_1(B, \cdot)$ is Borel for any $B \in \mathcal{B}^n$.

Then there is a unique probability measure P on $(\mathcal{R}^n \times \Lambda, \sigma(\mathcal{B}^n \times \mathcal{G}))$ such

that, for $B \in \mathcal{B}$ and $C \in \mathcal{G}$,

$$P(B \times C) = \int_C P_1(B, y) dP_2(y). \tag{1.43}$$

Furthermore, if $(\Lambda, \mathcal{G}) = (\mathcal{R}^m, \mathcal{B}^m)$, $X(x, y) = x$, and $Y(x, y) = y$ define the coordinate random vectors, then $P_Y = P_2$, $P_{X|Y=y} = P_1(\cdot, y)$, and the probability measure in (1.43) is the joint distribution of (X, Y), which has the following joint c.d.f.:

$$F(x, y) = \int_{(-\infty, y]} P_{X|Y=z}\big((-\infty, x]\big) dP_Y(z), \quad x \in \mathcal{R}^n, y \in \mathcal{R}^m, \tag{1.44}$$

where $(-\infty, a]$ denotes $(-\infty, a_1] \times \cdots \times (-\infty, a_k]$ for $a = (a_1, ..., a_k)$. ∎

Proposition 1.15 is sometimes called the "two-stage experiment theorem" for the following reason. If $Y \in \mathcal{R}^m$ is selected in stage 1 of an experiment according to its marginal distribution $P_Y = P_2$, and X is chosen afterward according to its conditional distribution $P_{X|Y=y} = P_1(\cdot, y)$, then the combined two-stage experiment produces a jointly distributed pair (X, Y) with distribution $P_{(X,Y)}$ given by (1.43). The following is an example.

Example 1.21. A market survey is conducted to study whether a new product is preferred over the product currently available in the market (old product). The survey is conducted by mail. Questionnaires are sent along with the sample products (both new and old) to N customers randomly selected from a population, where N is a positive integer. Each customer is asked to fill out the questionnaire and return it. Responses from customers are either 1 (new is better than old) or 0 (otherwise). Some customers, however, do not return the questionnaires. Let X be the number of ones in the returned questionnaires. What is the distribution of X?

If every customer returns the questionnaire, then (from elementary probability) X has the binomial distribution $Bi(p, N)$ in Table 1.1 (assuming that the population is large enough so that customers respond independently), where $p \in (0, 1)$ is the overall rate of customers who prefer the new product. Now, let Y be the number of customers who respond. Then Y is random. Suppose that customers respond independently with the same probability $\pi \in (0, 1)$. Then P_Y is the binomial distribution $Bi(\pi, N)$. Given $Y = y$ (an integer between 0 and N), $P_{X|Y=y}$ is the binomial distribution $Bi(p, y)$ if $y \geq 1$ and the point mass at 0 if $y = 0$. Using (1.44) and the fact that binomial distributions have p.d.f.'s (Table 1.1) w.r.t counting measure, we obtain that the joint c.d.f. of (X, Y) is

$$F(x, y) = \sum_{k=0}^{y} P_{X|Y=k}\big((-\infty, x]\big) \binom{N}{k} \pi^k (1 - \pi)^{N-k}$$

$$= \sum_{k=0}^{y} \sum_{j=0}^{\min(x,k)} \binom{k}{j} p^j (1-p)^{k-j} \binom{N}{k} \pi^k (1-\pi)^{N-k},$$

for $x = 0, 1, ..., y$, $y = 0, 1, ..., N$. The marginal c.d.f. $F_X(x) = F(x, \infty) = F(x, N)$. The p.d.f. of X w.r.t. counting measure is

$$\begin{aligned} f_X(x) &= \sum_{k=x}^{N} \binom{k}{x} p^x (1-p)^{k-x} \binom{N}{k} \pi^k (1-\pi)^{N-k} \\ &= \binom{N}{x} (\pi p)^x (1-\pi p)^{N-x} \sum_{k=x}^{N} \binom{N-x}{k-x} \left(\frac{\pi - \pi p}{1-\pi p}\right)^{k-x} \left(\frac{1-\pi}{1-\pi p}\right)^{N-k} \\ &= \binom{N}{x} (\pi p)^x (1-\pi p)^{N-x} \end{aligned}$$

for $x = 0, 1, ..., N$. It turns out that the marginal distribution of X is the binomial distribution $Bi(\pi p, N)$. ∎

1.5 Asymptotic Theorems

Asymptotic theory studies limiting behavior of random variables (vectors) and their distributions. It is an important tool for statistical analysis. A more complete coverage of asymptotic theory in statistical analysis can be found in Serfling (1980) and Sen and Singer (1993).

1.5.1 Convergence modes and stochastic orders

There are several convergence modes for random variables/vectors. For any k-vector $c \in \mathcal{R}^k$, $\|c\|$ denotes the usual distance between 0 and c, i.e., $\|c\|^2 = cc^\tau$.

Definition 1.8. Let $X, X_1, X_2, \ldots$ be random k-vectors defined on a probability space.
(i) We say that the sequence $\{X_n\}$ converges to X almost surely (a.s.) and write $X_n \to_{a.s.} X$ if and only if

$$P\left(\lim_{n\to\infty} \|X_n - X\| = 0\right) = 1.$$

(ii) We say that $\{X_n\}$ converges to X in probability and write $X_n \to_p X$ if and only if, for every fixed $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\|X_n - X\| > \epsilon\right) = 0.$$

(iii) We say that $\{X_n\}$ converges to X in L_p (or in pth-moment) and write $X_n \to_{L_p} X$ if and only if

$$\lim_{n\to\infty} E\|X_n - X\|^p = 0,$$

where $p > 0$ is a fixed constant.
(iv) Let F_{X_n} be the c.d.f. of X_n, $n = 1, 2,...$, and F_X be the c.d.f. of X. We say that $\{X_n\}$ converges to X in distribution (or in law) and write $X_n \to_d X$ if and only if, for each continuity point x of F_X,

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x). \quad \blacksquare$$

The a.s. convergence in Definition 1.8(i) is almost the same as the pointwise convergence of functions in calculus. The concept of convergence in probability, convergence in L_p, or a.s. convergence represents a sense in which, for n sufficiently large, X_n and X approximate each other as functions on the original probability space. The concept of convergence in distribution in Definition 1.8(iv), however, depends only on the distributions F_{X_n} and F_X and does not necessitate that X_n and X are close in any sense; in fact, Definition 1.8(iv) still makes sense even if X and X_n's are not defined on the same probability space. In Definition 1.8(iv), it is *not* required that $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ for every x. However, if $F_X(x)$ is a continuous function, then we have the following stronger result.

Proposition 1.16 (Pólya's theorem). If $X_n \to_d X$ and F_X is continuous, then

$$\lim_{n\to\infty} \sup_x |F_{X_n}(x) - F_X(x)| = 0. \quad \blacksquare$$

The following result describes the relationship among four convergence modes.

Theorem 1.8. Let $X, X_1, X_2, \ldots$ be random k-vectors.
(i) If $X_n \to_{a.s.} X$, then $X_n \to_p X$.
(ii) If $X_n \to_{L_p} X$ for a $p > 0$, then $X_n \to_p X$.
(iii) If $X_n \to_p X$, then $X_n \to_d X$.
(iv) If $X_n \to_d X$, then there are random vectors $Y, Y_1, Y_2, ...$ defined on a probability space such that $P_Y = P_X$, $P_{Y_n} = P_{X_n}$, $n = 1, 2,...$, and $Y_n \to_{a.s.} Y$.
(v) If, for every $\epsilon > 0$,

$$\sum_{n=1}^{\infty} P(\|X_n - X\| \geq \epsilon) < \infty,$$

then $X_n \to_{a.s.} X$.
(vi) If $X_n \to_p X$, then there is a subsequence $\{X_{n_k}, k = 1, 2, ...\}$ such that $X_{n_k} \to_{a.s.} X$ as $k \to \infty$.
(vii) Suppose that $X_n \to_d X$. Then, for any $r > 0$,

$$\lim_{n\to\infty} E\|X_n\|^r = E\|X\|^r < \infty$$

if and only if $\{\|X_n\|^r\}$ is *uniformly integrable* in the sense that

$$\lim_{t\to\infty} \sup_n E\left[\|X_n\|^r I_{(t,\infty)}(\|X_n\|)\right] = 0. \quad \blacksquare \tag{1.45}$$

The converse of Theorem 1.8(i), (ii), or (iii) is generally not true (see Example 1.22 and Exercise 75). Note that part (iv) of Theorem 1.8 is not a converse of part (i), but it is an important result in probability theory. Part (v) of Theorem 1.8 indicates that the converse of part (i) is true under the additional condition that $P(\|X_n - X\| \geq \epsilon)$ tends to 0 fast enough. A consequence of Theorem 1.8(vii) is that if $X_n \to_p X$ and $\{\|X_n - X\|^p\}$ is uniformly integrable, then $X_n \to_{L_p} X$; i.e., the converse of Theorem 1.8(ii) is true under the additional condition of uniform integrability. A useful sufficient condition for uniform integrability of $\{\|X_n\|^r\}$ is that

$$\sup_n E\|X_n\|^{r+\delta} < \infty \tag{1.46}$$

for a $\delta > 0$.

Example 1.22. Let $\theta_n = 1 + n^{-1}$ and X_n be a random variable having the exponential distribution $E(0, \theta_n)$ (Table 1.2), $n = 1, 2, ...$ Let X be a random variable having the exponential distribution $E(0, 1)$. For any $x > 0$,

$$F_{X_n}(x) = 1 - e^{-x/\theta_n} \to 1 - e^{-x} = F_X(x)$$

as $n \to \infty$. Since $F_{X_n}(x) \equiv 0 \equiv F_X(x)$ for $x \leq 0$, we have shown that $X_n \to_d X$.

Is it true that $X_n \to_p X$? This question cannot be answered without any further information about the random variables X and X_n. We consider two cases in which different answers can be obtained. First, suppose that $X_n \equiv \theta_n X$ (then X_n has the given c.d.f.). Note that $X_n - X = (\theta_n - 1)X = n^{-1}X$, which has the c.d.f. $(1 - e^{-nx})I_{[0,\infty)}(x)$. Hence

$$P\left(|X_n - X| \geq \epsilon\right) = e^{-n\epsilon} \to 0$$

for any $\epsilon > 0$. In fact, by Theorem 1.8(iv), $X_n \to_{a.s.} X$; since $E|X_n - X|^p = n^{-p}EX^p < \infty$ for any $p > 0$, $X_n \to_{L_p} X$ for any $p > 0$. Next, suppose

that X_n and X are independent random variables. Using result (1.28) and the fact that the p.d.f.'s for X_n and $-X$ are $\theta_n^{-1}e^{-x/\theta_n}I_{(0,\infty)}(x)$ and $e^x I_{(-\infty,0)}(x)$, respectively, we obtain that

$$P\left(|X_n - X| \le \epsilon\right) = \int_{-\epsilon}^{\epsilon}\int \theta_n^{-1}e^{-x/\theta_n}e^{y-x}I_{(0,\infty)}(x)I_{(-\infty,x)}(y)dxdy,$$

which converges to (by the dominated convergence theorem)

$$\int_{-\epsilon}^{\epsilon}\int e^{-x}e^{y-x}I_{(0,\infty)}(x)I_{(-\infty,x)}(y)dxdy = 1 - e^{-\epsilon}.$$

Thus, $P\left(|X_n - X| \ge \epsilon\right) \to e^{-\epsilon} > 0$ for any $\epsilon > 0$ and, therefore, $\{X_n\}$ does not converge to X in probability. ∎

The following result gives some useful sufficient and necessary conditions for convergence in distribution.

Theorem 1.9. Let $X, X_1, X_2, \ldots$ be random k-vectors.
(i) $X_n \to_d X$ if and only if $\lim_{n\to\infty} E[h(X_n)] = E[h(X)]$ for every bounded continuous function h from $\mathcal{R}^k$ to $\mathcal{R}$.
(ii) (Lévy-Cramér continuity theorem). Let $\phi_X, \phi_{X_1}, \phi_{X_2}, \ldots$ be the ch.f.'s of $X, X_1, X_2, \ldots$, respectively. $X_n \to_d X$ if and only if $\lim_{n\to\infty}\phi_{X_n}(t) = \phi_X(t)$ for all $t \in \mathcal{R}^k$.
(iii) (Cramér-Wold device). $X_n \to_d X$ if and only if $X_n c^\tau \to_d Xc^\tau$ for every $c \in \mathcal{R}^k$. ∎

Examples of applications of Theorem 1.9 are given as exercises in §1.6. The following result can be used to check whether $X_n \to_d X$ when X has a p.d.f. f and X_n has a p.d.f. f_n.

Proposition 1.17 (Scheffé's theorem). Let $\{f_n\}$ be a sequence of p.d.f.'s on $\mathcal{R}^k$ w.r.t. a measure ν. Suppose that $\lim_{n\to\infty} f_n(x) = f(x)$ a.e. ν and $f(x)$ is a p.d.f. w.r.t. ν. Then $\lim_{n\to\infty}\int |f_n(x) - f(x)|d\nu = 0$.
Proof. Let $g_n(x) = [f(x) - f_n(x)]I_{\{f\ge f_n\}}(x)$, $n = 1, 2, \ldots$. Then

$$\int |f_n(x) - f(x)|d\nu = 2\int g_n(x)d\nu.$$

Since $0 \le g_n(x) \le f(x)$ for all x and $g_n \to 0$ a.e. ν, the result follows from the dominated convergence theorem. ∎

As an example, consider the Lebesgue p.d.f. f_n of the t-distribution t_n (Table 1.2), $n = 1, 2, \ldots$. One can show (exercise) that $f_n \to f$, where f is the standard normal p.d.f. This is an important result in statistics.

We now introduce the notion of $O(\cdot)$, $o(\cdot)$, and stochastic $O(\cdot)$ and $o(\cdot)$. In calculus, two sequences of real numbers, $\{a_n\}$ and $\{b_n\}$, satisfy $a_n = O(b_n)$ if and only if $|a_n| \leq c|b_n|$ for all n and a constant c; and $a_n = o(b_n)$ if and only if $a_n/b_n \to 0$ as $n \to \infty$.

Definition 1.9. Let $X_1, X_2, ...$ and $Y_1, Y_2, ...$ be random variables defined on a probability space $(\Omega, \mathcal{F}, P)$.
(i) $X_n = O(Y_n)$ a.s. if and only if $X_n(\omega) = O(Y_n(\omega))$ a.s. P.
(ii) $X_n = o(Y_n)$ a.s. if and only if $X_n/Y_n \to_{a.s.} 0$.
(iii) $X_n = O_p(Y_n)$ if and only if, for any $\epsilon > 0$, there is a constant $C_\epsilon > 0$ such that

$$\sup_n P(|X_n| \geq C_\epsilon |Y_n|) < \epsilon.$$

(iv) $X_n = o_p(Y_n)$ if and only if $X_n/Y_n \to_p 0$. ∎

Note that $X_n = o_p(Y_n)$ implies $X_n = O_p(Y_n)$; $X_n = O_p(Y_n)$ and $Y_n = O_p(Z_n)$ implies $X_n = O_p(Z_n)$; but $X_n = O_p(Y_n)$ does not imply $Y_n = O_p(X_n)$. The same conclusion can be obtained if $O_p(\cdot)$ and $o_p(\cdot)$ are replaced by $O(\cdot)$ a.s. and $o(\cdot)$ a.s., respectively. Some results related to O_p are given in Exercise 88. For example, if $X_n \to_d X$ for a random variable X, then $X_n = O_p(1)$. Since $a_n = O(1)$ means that $\{a_n\}$ is bounded, $\{X_n\}$ is said to be bounded in probability if $X_n = O_p(1)$.

1.5.2 Convergence of transformations

Transformation is an important tool in statistics. For random vectors X_n converging to X in some sense, we often want to know whether $g(X_n)$ converges to $g(X)$ in the same sense. The following result provides an answer to most problems.

Theorem 1.10. Let $X, X_1, X_2, ...$ be random k-vectors defined on a probability space and g be a measurable function from $(\mathcal{R}^k, \mathcal{B}^k)$ to $(\mathcal{R}^l, \mathcal{B}^l)$. Suppose that g is continuous a.s. P_X. Then
(i) $X_n \to_{a.s.} X$ implies $g(X_n) \to_{a.s.} g(X)$.
(ii) $X_n \to_p X$ implies $g(X_n) \to_p g(X)$.
(iii) $X_n \to_d X$ implies $g(X_n) \to_d g(X)$. ∎

Example 1.23. (i) Let $X_1, X_2, ...$ be random variables. If $X_n \to_d X$, where X has the $N(0,1)$ distribution, then $X_n^2 \to_d Y$, where Y has the chi-square distribution χ_1^2 (Example 1.13).
(ii) Let (X_n, Y_n) be random 2-vectors satisfying $(X_n, Y_n) \to_d (X, Y)$, where X and Y are independent random variables having the $N(0,1)$ distribution, then $X_n/Y_n \to_d X/Y$, which has the Cauchy distribution $C(0,1)$ (§1.3.1).

(iii) Under the conditions in part (ii), $\max(X_n, Y_n) \to_d \max(X, Y)$, which has the c.d.f. $[\Phi(x)]^2$ ($\Phi(x)$ is the c.d.f. of $N(0,1)$). ∎

In Example 1.23(ii) and (iii), the condition that $(X_n, Y_n) \to_d (X, Y)$ cannot be relaxed to $X_n \to_d X$ and $Y_n \to_d Y$ (exercise); i.e., we need the convergence of the joint c.d.f. of (X_n, Y_n). This is different when $\to_d$ is replaced by $\to_p$ or $\to_{a.s.}$. The following result, which plays an important role in probability and statistics, establishes the convergence in distribution of $X_n + Y_n$ or $X_n Y_n$ when no information regarding the joint c.d.f. of (X_n, Y_n) is provided.

Theorem 1.11 (Slutsky's theorem). Let $X, X_1, X_2, ..., Y_1, Y_2, ...$ be random variables on a probability space. Suppose that $X_n \to_d X$ and $Y_n \to_p c$, where c is a fixed real number. Then
(i) $X_n + Y_n \to_d X + c$;
(ii) $Y_n X_n \to_d cX$;
(iii) $X_n / Y_n \to_d X/c$ if $c \neq 0$.

Proof. We prove (i) only. The proofs of (ii) and (iii) are left as exercises. Let $t \in \mathcal{R}$ and $\epsilon > 0$ be fixed constants. Then

$$\begin{aligned} F_{X_n+Y_n}(t) &= P(X_n + Y_n \le t) \\ &\le P(\{X_n + Y_n \le t\} \cap \{|Y_n - c| < \epsilon\}) + P(|Y_n - c| \ge \epsilon) \\ &\le P(X_n \le t - c + \epsilon) + P(|Y_n - c| \ge \epsilon) \end{aligned}$$

and, similarly,

$$F_{X_n+Y_n}(t) \ge P(X_n \le t - c - \epsilon) - P(|Y_n - c| \ge \epsilon).$$

If $t - c$, $t - c + \epsilon$, and $t - c - \epsilon$ are continuity points of F_X, then it follows from the previous two inequalities and the hypotheses of the theorem that

$$F_X(t - c - \epsilon) \le \liminf_n F_{X_n+Y_n}(t) \le \limsup_n F_{X_n+Y_n}(t) \le F_X(t - c + \epsilon).$$

Since ϵ can be arbitrary (why?),

$$\lim_{n\to\infty} F_{X_n+Y_n}(t) = F_X(t - c).$$

The result follows from $F_{X+c}(t) = F_X(t - c)$. ∎

An application of Theorem 1.11 is given in the proof of the following important result.

Theorem 1.12. Let $X_1, X_2, ...$ and Y be random k-vectors satisfying

$$a_n(X_n - c) \to_d Y, \tag{1.47}$$

where $c \in \mathcal{R}^k$ and $\{a_n\}$ is a sequence of positive numbers with $\lim_{n\to\infty} a_n = \infty$. Let g be a function from $\mathcal{R}^k$ to $\mathcal{R}$.
(i) If g is differentiable at c, then

$$a_n[g(X_n) - g(c)] \to_d \nabla g(c) Y^\tau, \tag{1.48}$$

where $\nabla g(x)$ denotes the k-vector of partial derivatives of g at x.
(ii) Suppose that g has continuous partial derivatives of order $m > 1$ in a neighborhood of c, with all the partial derivatives of order j, $1 \le j \le m-1$, vanishing at c, but with the mth-order partial derivatives not all vanishing at c. Then

$$a_n^m[g(X_n) - g(c)] \to_d \frac{1}{m!} \sum_{i_1=1}^{k} \cdots \sum_{i_m=1}^{k} \frac{\partial^m g}{\partial x_{i_1} \cdots \partial x_{i_m}}\bigg|_{x=c} Y_{i_1} \cdots Y_{i_m}, \tag{1.49}$$

where Y_j is the jth component of Y.
Proof. We prove (i) only. The proof of (ii) is similar. Let

$$Z_n = a_n[g(X_n) - g(c)] - a_n \nabla g(c)(X_n - c)^\tau.$$

If we can show that $Z_n = o_p(1)$, then by (1.47), Theorem 1.9(iii), and Theorem 1.11(i), result (1.48) holds.

The differentiability of g at c implies that for any $\epsilon > 0$, there is a $\delta_\epsilon > 0$ such that

$$|g(x) - g(c) - \nabla g(c)(x-c)^\tau| \le \epsilon \|x - c\| \tag{1.50}$$

whenever $\|x - c\| < \delta_\epsilon$. Let $\eta > 0$ be fixed. By (1.50),

$$P(|Z_n| \ge \eta) \le P(\|X_n - c\| \ge \delta_\epsilon) + P(a_n \|X_n - c\| \ge \eta/\epsilon).$$

Since $a_n \to \infty$, (1.47) and Theorem 1.11(ii) imply $X_n \to_p c$. By Theorem 1.10(iii), (1.47) implies $a_n\|X_n - c\| \to_d \|Y\|$. Without loss of generality, we can assume that η/ϵ is a continuity point of $F_{\|Y\|}$. Then

$$\begin{aligned} \limsup_n P(|Z_n| \ge \eta) &\le \lim_{n\to\infty} P(\|X_n - c\| \ge \delta_\epsilon) \\ &\quad + \lim_{n\to\infty} P(a_n\|X_n - c\| \ge \eta/\epsilon) \\ &= P(\|Y\| \ge \eta/\epsilon). \end{aligned}$$

The proof is complete since ϵ can be arbitrary. ∎

In statistics, we often need a nondegenerated limiting distribution of $a_n[g(X_n) - g(c)]$ so that probabilities involving $a_n[g(X_n) - g(c)]$ can be approximated by the c.d.f. of $\nabla g(c) Y^\tau$, if (1.48) holds. Hence, result (1.48) is not useful for this purpose if $\nabla g(c) = 0$, and in such cases result (1.49) may be applied.

A useful method in statistics, called the *delta-method*, is based on the following corollary of Theorem 1.12. Recall that $N(\mu, \sigma^2)$ denotes a random variable having the $N(\mu, \sigma^2)$ distribution.

Corollary 1.1. Assume the conditions of Theorem 1.12. If Y has the $N_k(0, \Sigma)$ distribution, then

$$a_n[g(X_n) - g(c)] \to_d N\left(0, \nabla g(c)\Sigma[\nabla g(c)]^\tau\right). \quad \blacksquare$$

Example 1.24. Let $X_1, X_2, ...$ be random variables such that

$$\sqrt{n}(X_n - c) \to_d N(0, 1).$$

Consider the function $g(x) = x^2$. If $c \neq 0$, then an application of Corollary 1.1 gives that

$$\sqrt{n}(X_n^2 - c^2) \to_d N(0, 4c^2).$$

If $c = 0$, the first-order derivative of g at 0 is 0 but the second-order derivative of $g \equiv 2$. Hence, an application of result (1.49) gives that

$$nX_n^2 \to_d [N(0, 1)]^2,$$

which has the chi-square distribution χ_1^2 (Example 1.13). The last result can also be obtained by applying Theorem 1.10(iii). $\blacksquare$

1.5.3 The law of large numbers

The law of large numbers concerns the limiting behavior of sums of independent random variables. The weak law of large numbers (WLLN) refers to convergence in probability, whereas the strong law of large numbers (SLLN) refers to a.s. convergence. Our first result gives the WLLN and SLLN for a sequence of independent and identically distributed (i.i.d.) random variables.

Theorem 1.13. Let $X_1, X_2, ...$ be i.i.d. random variables having a c.d.f. F.
(i) (The WLLN). The existence of a sequence of real numbers $\{a_n\}$ for which

$$\frac{1}{n}\sum_{i=1}^{n} X_i - a_n \to_p 0$$

if and only if

$$\lim_{x\to\infty} x[1 - F(x) + F(-x)] = 0,$$

in which case we may take $a_n = E[X_1 I_{(-n,n)}(X_1)]$.
(ii) (The SLLN). The existence of a constant c for which

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to_{a.s.} c$$

if and only if $E|X_1| < \infty$, in which case $c = EX_1$ and

$$\frac{1}{n}\sum_{i=1}^{n} c_i(X_i - EX_1) \to_{a.s.} 0$$

for any bounded sequence of real numbers $\{c_i\}$.
(iii) (The Marcinkiewicz SLLN). If $E|X_1|^\delta < \infty$ for a $\delta \in (0,1)$, then

$$\frac{1}{n^{1/\delta}}\sum_{i=1}^{n} |X_i| \to_{a.s.} 0. \quad \blacksquare$$

The proof of this theorem can be found in Billingsley (1986) or Chung (1974). The next result is for sequences of independent but not necessarily identically distributed random variables.

Theorem 1.14. Let $X_1, X_2, ...$ be independent random variables with finite expectations.
(i) (The SLLN). If

$$\sum_{i=1}^{\infty} \frac{\mathrm{Var}(X_i)}{i^2} < \infty,$$

then

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - EX_i) \to_{a.s.} 0.$$

(ii) (The WLLN). If there is a $\delta \in (0,1)$ such that

$$\lim_{n\to\infty} \frac{1}{n^{1+\delta}}\sum_{i=1}^{n} E|X_i|^{1+\delta} = 0,$$

then

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - EX_i) \to_{p} 0. \quad \blacksquare$$

The WLLN and SLLN have many applications in probability and statistics. The following is an example. Other examples can be found in later chapters.

Example 1.25. Let f and g be continuous functions on $[0,1]$ satisfying $0 \le f(x) \le Cg(x)$ for all x, where $C > 0$ is a constant. We now show that

$$\lim_{n\to\infty} \int_0^1 \int_0^1 \cdots \int_0^1 \frac{\sum_{i=1}^n f(x_i)}{\sum_{i=1}^n g(x_i)} dx_1 dx_2 \cdots dx_n = \frac{\int_0^1 f(x)dx}{\int_0^1 g(x)dx} \tag{1.51}$$

(assuming that $\int_0^1 g(x)dx \neq 0$). Let $X_1, X_2, ...$ be i.i.d. random variables having the uniform distribution on $[0,1]$. By Proposition 1.7(i), $E[f(X_1)] = \int_0^1 f(x)dx < \infty$ and $E[g(X_1)] = \int_0^1 g(x)dx < \infty$. By the SLLN (Theorem 1.13(i)),

$$\frac{1}{n}\sum_{i=1}^n f(X_i) \to_{a.s.} E[f(X_1)]$$

and the same result holds when f is replaced by g. By Theorem 1.10(i),

$$\frac{\sum_{i=1}^n f(X_i)}{\sum_{i=1}^n g(X_i)} \to_{a.s.} \frac{E[f(X_1)]}{E[g(X_1)]}. \tag{1.52}$$

Since the random variable on the left-hand side of (1.52) is bounded by C, result (1.51) follows from the dominated convergence theorem and the fact that the left-hand side of (1.51) is the expectation of the random variable on the left-hand side of (1.52). ∎

1.5.4 The central limit theorem

The WLLN and SLLN may not be useful in approximating the distributions of (normalized) sums of independent random variables. We need to use the *Central Limit Theorem* (CLT), which plays a fundamental role in statistical asymptotic theory.

Theorem 1.15 (Lindeberg's CLT). Let $\{X_{nj}, j = 1, ..., k_n\}$ be independent random variables with $0 < \sigma_n^2 = \mathrm{Var}(\sum_{j=1}^{k_n} X_{nj}) < \infty$, $n = 1, 2, ...$, and $k_n \to \infty$ as $n \to \infty$. If

$$\lim_{n\to\infty} \frac{1}{\sigma_n^2} \sum_{j=1}^{k_n} E\left[(X_{nj} - EX_{nj})^2 I_{\{|X_{nj}-EX_{nj}|>\epsilon\sigma_n\}}\right] = 0 \tag{1.53}$$

for any $\epsilon > 0$, then

$$\frac{1}{\sigma_n}\sum_{j=1}^{k_n}(X_{nj} - EX_{nj}) \to_d N(0,1). \quad ∎$$

The proof of this theorem can be found in Billingsley (1986) or Chung (1974). Condition (1.53) with an arbitrary $\epsilon > 0$ is called Lindeberg's condition. It is implied by the following condition, called Liapunov's condition, which is somewhat easier to verify:

$$\sum_{j=1}^{k_n} E|X_{nj} - EX_{nj}|^{2+\delta} = o(\sigma_n^{2+\delta}) \tag{1.54}$$

for some $\delta > 0$.

Example 1.26. Let $X_1, X_2, ...$ be independent random variables. Suppose that X_i has the binomial distribution $Bi(p_i, 1)$, $i = 1, 2,...$, and that $\sigma_n^2 = \sum_{i=1}^n \text{Var}(X_i) = \sum_{i=1}^n p_i(1-p_i) \to \infty$ as $n \to \infty$. For each i, $EX_i = p_i$ and

$$E|X_i - EX_i|^3 = (1-p_i)^3 p_i + p_i^3(1-p_i) \le 2p_i(1-p_i).$$

Hence

$$\sum_{i=1}^n E|X_i - EX_i|^3 \le 2\sigma_n^2,$$

i.e., Liapunov's condition (1.54) holds with $\delta = 1$. Thus, by Theorem 1.15,

$$\frac{1}{\sigma_n}\sum_{i=1}^n (X_i - p_i) \to_d N(0,1). \quad \blacksquare$$

The following are useful corollaries of Theorem 1.15 (and Theorem 1.9(iii)).

Corollary 1.2 (Multivariate CLT). Let $X_1, ..., X_n$ be i.i.d. random k-vectors with a finite $\Sigma = \text{Var}(X_1)$. Then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - EX_1) \to_d N_k(0, \Sigma). \quad \blacksquare$$

Corollary 1.3. Let $X_{ni} \in \mathcal{R}^{m_i}$, $i = 1, ..., k_n$, be independent random vectors with $m_i \le m$ (a fixed integer), $n = 1, 2,...$, $k_n \to \infty$ as $n \to \infty$, and $\inf_{i,n} \lambda_-[\text{Var}(X_{ni})] > 0$, where $\lambda_-[A]$ is the smallest eigenvalue of A. Let $c_{ni} \in \mathcal{R}^{m_i}$ be vectors such that

$$\lim_{n\to\infty} \left(\max_{1\le i\le k_n} \|c_{ni}\|^2 \Big/ \sum_{i=1}^{k_n} \|c_{ni}\|^2 \right) = 0.$$

(i) Suppose that $\sup_{i,n} E\|X_{ni}\|^{2+\delta} < \infty$ for some $\delta > 0$. Then

$$\sum_{i=1}^{k_n}(X_{ni} - EX_{ni})c_{ni}^\tau \Big/ \left[\sum_{i=1}^{k_n}\mathrm{Var}(X_{ni}c_{ni}^\tau)\right]^{1/2} \to_d N(0,1). \tag{1.55}$$

(ii) Suppose that X_{ni}, $i = 1, ..., k_n$, $n = 1, 2, ...$, have a common distribution and $E\|X_{11}\|^2 < \infty$. Then (1.55) holds. ∎

Applications of these corollaries can be found in later chapters. More results on the CLT can be found, for example, in Serfling (1980) and Shorack and Wellner (1986).

Let Y_n be a sequence of random variables, $\{\mu_n\}$ and $\{\sigma_n\}$ be sequences of real numbers such that $\sigma_n > 0$ for all n and $(Y_n - \mu_n)/\sigma_n \to_d N(0,1)$. Then, by Proposition 1.16,

$$\lim_{n\to\infty} \sup_x |F_{(Y_n-\mu_n)/\sigma_n}(x) - \Phi(x)| = 0, \tag{1.56}$$

where Φ is the c.d.f. of $N(0,1)$. This implies that for any sequence of real numbers $\{c_n\}$,

$$\lim_{n\to\infty} \left| P\left(Y_n \le c_n\right) - \Phi\left(\frac{c_n - \mu_n}{\sigma_n}\right) \right| = 0,$$

i.e., $P(Y_n \le c_n)$ can be approximated by $\Phi\left(\frac{c_n-\mu_n}{\sigma_n}\right)$, regardless of whether $\{c_n\}$ has a limit. Since $\Phi\left(\frac{t-\mu_n}{\sigma_n}\right)$ is the c.d.f. of $N(\mu_n, \sigma_n^2)$, Y_n is said to be *asymptotically distributed* as $N(\mu_n, \sigma_n^2)$ or simply *asymptotically normal*. For example, $\sum_{i=1}^{k_n} c_{ni}X_{ni}^\tau$ in Corollary 1.3 is asymptotically normal. This can be extended to random vectors. For example, $\sum_{i=1}^n X_i$ in Corollary 1.2 is asymptotically distributed as $N_k(EX_1, \Sigma/n)$.

1.6 Exercises

1. Let A and B be two nonempty proper subsets of a sample space Ω, $A \neq B$ and $A \cap B \neq \emptyset$. Obtain $\sigma(\{A, B\})$, the smallest σ-field containing A and B.

2. Let $\mathcal{C}$ be a collection of subsets of Ω and let $\Gamma = \{\mathcal{F} : \mathcal{F}$ is a σ-field on Ω and $\mathcal{C} \subset \mathcal{F}\}$. Show that $\Gamma \neq \emptyset$ and $\sigma(\mathcal{C}) = \cap\{\mathcal{F} : \mathcal{F} \in \Gamma\}$.

3. Show that if $\mathcal{C}_1 \subset \mathcal{C}_2$, then $\sigma(\mathcal{C}_1) \subset \sigma(\mathcal{C}_2)$.

4. Let $\mathcal{C}$ be the collection of intervals of the form $(a, b]$, where $-\infty < a < b < \infty$, and let $\mathcal{D}$ be the collection of closed sets on $\mathcal{R}$. Show that $\mathcal{B} = \sigma(\mathcal{C}) = \sigma(\mathcal{D})$, where $\mathcal{B}$ is the Borel σ-field on $\mathcal{R}$.

5. Let $(\Omega, \mathcal{F})$ be a measurable space and $C \in \mathcal{F}$. Show that $\mathcal{F}_C = \{C \cap A : A \in \mathcal{F}\}$ is a σ-field on C.

6. Prove part (ii) and part (iii) of Proposition 1.1.

7. Let ν_i, $i = 1, 2, ...,$ be measures on $(\Omega, \mathcal{F})$ and a_i, $i = 1, 2, ...,$ be positive numbers. Show that $a_1\nu_1 + a_2\nu_2 + \cdots$ is a measure on $(\Omega, \mathcal{F})$.

8. Let $A_1, A_2, ...$ be a sequence of measurable sets and P be a probability measure. Define

$$\limsup_n A_n = \bigcap_{n=1}^{\infty} \bigcup_{i=n}^{\infty} A_i \quad \text{and} \quad \liminf_n A_n = \bigcup_{n=1}^{\infty} \bigcap_{i=n}^{\infty} A_i.$$

(a) Show that

$$P\left(\liminf_n A_n\right) \leq \liminf_n P(A_n)$$

and

$$\limsup_n P(A_n) \leq P\left(\limsup_n A_n\right).$$

(b) (Borel-Cantelli's lemma). Show that if $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $P(\limsup_n A_n) = 0$.
(c) (Borel-Cantelli's lemma). Show that if $A_1, A_2, ...$ are independent (Definition 1.7) and $\sum_{n=1}^{\infty} P(A_n) = \infty$, then $P(\limsup_n A_n) = 1$.

9. Prove part (i) of Proposition 1.2.

10. Let $F(x_1, ..., x_k)$ be a c.d.f on $\mathcal{R}^k$. Show that
(a) $F(x_1, ..., x_{k-1}, x_k) \leq F(x_1, ..., x_{k-1}, x_k')$ if $x_k \leq x_k'$.
(b) $\lim_{x_i \to -\infty} F(x_1, ..., x_k) = 0$ for any $1 \leq i \leq k$.
(c) $F(x_1, ..., x_{k-1}, \infty) = \lim_{x_k \to \infty} F(x_1, ..., x_{k-1}, x_k)$ is a c.d.f on $\mathcal{R}^{k-1}$.

11. Let $(\Omega_i, \mathcal{F}_i) = (\mathcal{R}, \mathcal{B})$, $i = 1, ..., k$. Show that the product σ-field $\sigma(\mathcal{F}_1 \times \cdots \times \mathcal{F}_k)$ is the same as the Borel σ-field generated by $\mathcal{O}$, all open sets in $\mathcal{R}^k$.

12. Let ν and λ be two measures on $(\Omega, \mathcal{F})$ such that $\nu(A) = \lambda(A)$ for any $A \in \mathcal{C}$, where $\mathcal{C} \subset \mathcal{F}$ is a collection having the property that if A and B are in $\mathcal{C}$, then so is $A \cap B$. Assume that there are $A_i \in \mathcal{C}$, $i = 1, 2, ...,$ such that $\cup A_i = \Omega$ and $\nu(A_i) < \infty$ for all i. Show that $\nu(A) = \lambda(A)$ for any $A \in \sigma(\mathcal{C})$. This proves the uniqueness part of Proposition 1.3. (Hint: show that $\{A \in \sigma(\mathcal{C}) : \nu(A) = \lambda(A)\}$ is a σ-field.)

13. Show that $f^{-1}(B^c) = (f^{-1}(B))^c$ and $f^{-1}(\cup B_i) = \cup f^{-1}(B_i)$.

14. Show that $f^{-1}(\mathcal{G})$ is a σ-field, if f is a function from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$.

15. Prove parts (i)-(iv) of Proposition 1.4.

16. Show that a monotone function from $\mathcal{R}$ to $\mathcal{R}$ is Borel.

17. Let f be a nonnegative Borel function on $(\Omega, \mathcal{F})$. Show that f is the limit of a sequence of simple functions $\{\varphi_n\}$ satisfying $0 \leq \varphi_1 \leq \varphi_2 \leq \cdots \leq f$.

18. Let f be a function from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$ and $A_1, A_2, ...$ be disjoint events in $\mathcal{F}$ such that $\cup A_i = \Omega$. Let f_n be a function from $(A_n, \mathcal{F}_{A_n})$ to $(\Lambda, \mathcal{G})$ such that $f_n(\omega) = f(\omega)$ for any $\omega \in A_n$, $n = 1, 2, ...$ Show that f is measurable from $(\Omega, \mathcal{F})$ to $(\Lambda, \mathcal{G})$ if and only if f_n is measurable from $(A_n, \mathcal{F}_{A_n})$ to $(\Lambda, \mathcal{G})$ for each n.

19. Let $\prod_{i=1}^{k}(\Omega_i, \mathcal{F}_i)$ be a product measurable space and π_i be the ith *projection*, i.e., π_i is a function from $\Omega_1 \times \cdots \times \Omega_k$ to Ω_i such that $\pi_i(\omega_1, ..., \omega_k) = \omega_i$, $\omega_i \in \Omega_i$, $i = 1, ..., k$. Show that π_i's are measurable.

20. Let f be a Borel function on $\mathcal{R}^2$. Define a function g from $\mathcal{R}$ to $\mathcal{R}$ as $g(x) = f(x, y)$, where y is a fixed point in $\mathcal{R}$. Show that g is Borel. Is it true that f is Borel from $\mathcal{R}^2$ to $\mathcal{R}$ if $f(x, y)$ with any fixed y or fixed x is Borel from $\mathcal{R}$ to $\mathcal{R}$?

21. Show that the set function defined in (1.8) is a measure.

22. Prove (1.13) in Example 1.5.

23. Prove Proposition 1.5.

24. Prove Proposition 1.6(i).

25. (Chebyshev's inequality). Let X be a random variable and ϕ a strictly positive and increasing function on $(0, \infty)$ satisfying $\phi(-t) = \phi(t)$. Show that for each $t > 0$,

$$P(|X| \geq t) \leq \frac{E\phi(X)}{\phi(t)}.$$

26. Let ν_i, $i = 1, 2$, be measures on $(\Omega, \mathcal{F})$ and f be Borel. Show that

$$\int f d(\nu_1 + \nu_2) = \int f d\nu_1 + \int f d\nu_2,$$

i.e., if either side of the equality is well defined, then so is the other side, and the two sides are equal.

27. Let f be an integrable Borel function on $(\Omega, \mathcal{F}, \nu)$. Show that for each $\epsilon > 0$, there is a δ_ϵ such that $\nu(A) < \delta_\epsilon$ and $A \in \mathcal{F}$ imply $\int_A |f| d\nu < \epsilon$.

28. Consider Example 1.9. Show that (1.16) does not hold for the following function:

$$f(i,j) = \begin{cases} 1 & i = j \\ -1 & i = j-1 \\ 0 & \text{otherwise.} \end{cases}$$

Does this contradict Fubini's theorem?

29. Let f be a nonnegative Borel function on $(\Omega, \mathcal{F}, \nu)$ with a σ-finite ν. Let m be the Lebesgue measure on $(\mathcal{R}, \mathcal{B})$ and

$$A = \{(\omega, x) \in \Omega \times \mathcal{R} : 0 \le x \le f(\omega)\}.$$

Show that $A \in \sigma(\mathcal{F} \times \mathcal{B})$ and $\int_\Omega f d\nu = \nu \times m(A)$.

30. For any c.d.f. F and any $a \ge 0$, show that

$$\int [F(x+a) - F(x)]dx = a.$$

31. (Integration by parts). Let F and G be two c.d.f.'s on $\mathcal{R}$. Show that if F and G have no common points of discontinuity in the interval $[a, b]$, then

$$\int_{(a,b]} G(x)dF(x) = F(b)G(b) - F(a)G(a) - \int_{(a,b]} F(x)dG(x).$$

32. Let f be a Borel function on $\mathcal{R}^2$ such that $f(x,y) = 0$ for each $x \in \mathcal{R}$ and $y \notin C_x$, where $m(C_x) = 0$ for each x and m is the Lebesgue measure. Show that $f(x,y) = 0$ for each $y \notin C$ and $x \notin B_y$, where $m(C) = 0$ and $m(B_y) = 0$ for each $y \notin C$.

33. Show that the set function defined by (1.17) is a measure.

34. Consider Example 1.11. Show that if (1.20) holds, then $P(A) = \int_A f(x)dx$ for any Borel set A. (Hint: $\mathcal{A} = \{A : P(A) = \int_A f(x)dx\}$ is a σ-field containing all sets of the form $(-\infty, x]$.)

35. Prove Proposition 1.7.

36. Let F_i be a c.d.f. having a Lebesgue p.d.f. f_i, $i = 1, 2$. Assume that there is a $c \in \mathcal{R}$ such that $F_1(c) < F_2(c)$. Define

$$F(x) = \begin{cases} F_1(x) & -\infty < x < c \\ F_2(x) & c \le x < \infty. \end{cases}$$

Show that the probability measure P corresponding to F satisfies $P \ll m + \delta_c$, where m is the Lebesgue measure and δ_c is the point mass at c, and find the p.d.f. of F w.r.t. $m + \delta_c$.

37. Let X be a random variable having the Lebesgue p.d.f. $\frac{2x}{\pi^2} I_{(0,\pi)}(x)$. Derive the p.d.f. of $Y = \sin X$.

38. Let X be a random variable having a continuous c.d.f. F. Show that $Y = F(X)$ has the uniform distribution $U(0,1)$ (Table 1.2).

39. Let U be a random variable having the uniform distribution $U(0,1)$ and let F be a c.d.f. Show that the c.d.f. of $Y = F^{-1}(U)$ is F, where $F^{-1}(t) = \inf\{x \in \mathcal{R} : F(x) \geq t\}$.

40. Let X_i, $i = 1, 2, 3$, be independent random variables having the same Lebesgue p.d.f. $f(x) = e^{-x} I_{(0,\infty)}(x)$. Obtain the joint Lebesgue p.d.f. of (Y_1, Y_2, Y_3), where $Y_1 = X_1 + X_2 + X_3$, $Y_2 = X_1/(X_1 + X_2)$, and $Y_3 = (X_1 + X_2)/(X_1 + X_2 + X_3)$. Are the Y_i independent?

41. Let X_1 and X_2 be independent random variables having the standard normal distribution. Obtain the joint Lebesgue p.d.f. of (Y_1, Y_2), where $Y_1 = \sqrt{X_1^2 + X_2^2}$ and $Y_2 = X_1/X_2$. Are the Y_i independent?

42. Let X_1 and X_2 be independent random variables and $Y = X_1 + X_2$. Show that $F_Y(y) = \int F_{X_2}(y - x) dF_{X_1}(x)$.

43. Let X be a random variable and $a > 0$. Show that $E|X|^a < \infty$ if and only if $\sum_{n=1}^{\infty} n^{a-1} P(|X| \geq n) < \infty$.

44. Let X be a random variable with range $\{0, 1, 2, ...\}$. Show that if $EX < \infty$, then
$$EX = \sum_{n=1}^{\infty} P(X \geq n).$$

45. Let X be a random variable having a c.d.f. F_X. Show that if $X \geq 0$ a.s., then
$$EX = \int [1 - F_X(x)] dx;$$
in general, if EX exists, then
$$EX = \int_0^{\infty} [1 - F_X(x)] dx - \int_{-\infty}^{0} F_X(x) dx.$$

46. (Jensen's inequality). Let X be a random variable and f be a convex function on $\mathcal{R}$, i.e., f satisfies
$$f\left(\sum_{i=1}^{k} a_i x_i\right) \leq \sum_{i=1}^{k} a_i f(x_i), \quad x_i \in \mathcal{R}$$

for every set of positive $a_1, ..., a_k$ with $\sum_{i=1}^k a_i = 1$. Suppose that $E|X| < \infty$ and $E|f(X)| < \infty$. Show that $f(EX) \le Ef(X)$. (Hint: consider nonnegative simple functions first and use the fact that f is continuous.)

47. Show that $EX = \mu$ and $\mathrm{Var}(X) = \Sigma$ for the $N_k(\mu, \Sigma)$ distribution.

48. Let X be a random variable with $EX^2 < \infty$ and let $Y = |X|$. Suppose that X has a p.d.f. symmetric about 0. Show that X and Y are uncorrelated, but they are not independent.

49. Let (X, Y) be a random 2-vector with the following Lebesgue p.d.f.:

$$f(x,y) = \begin{cases} \pi^{-1} & x^2 + y^2 \le 1 \\ 0 & x^2 + y^2 > 1. \end{cases}$$

Show that X and Y are uncorrelated, but are not independent.

50. Let $X_1, ..., X_k$ be independent random variables and $Y = X_1 + \cdots + X_k$. Prove the following statements, using Proposition 1.10.
(a) If X_i has the binomial distribution $Bi(p, n_i)$, $i = 1, ..., k$, then Y has the binomial distribution $Bi(p, n_1 + \cdots + n_k)$.
(b) If X_i has the Poisson distribution $P(\theta_i)$, $i = 1, ..., k$, then Y has the Poisson distribution $P(\theta_1 + \cdots + \theta_k)$.
(c) If X_i has the negative binomial distribution $NB(p, r_i)$, $i = 1, ..., k$, then Y has the negative binomial distribution $NB(p, r_1 + \cdots + r_k)$.
(d) If X_i has the exponential distribution $E(0, \theta)$, $i = 1, ..., k$, then Y has the gamma distribution $\Gamma(k, \theta)$.
(e) If X_i has the Cauchy distribution $C(0, 1)$, $i = 1, ..., k$, then Y/k has the same distribution as X_1.

51. Show the following properties of the multivariate normal distribution $N_k(\mu, \Sigma)$.
(a) The m.g.f. of $N_k(\mu, \Sigma)$ is $e^{\mu t^\tau + t\Sigma t^\tau/2}$.
(b) Let X be a random k-vector having the $N_k(\mu, \Sigma)$ distribution and $Y = XA + c$, where A is a $k \times l$ matrix of rank $l \le k$ and $c \in \mathcal{R}^l$. Then Y has the $N_l(\mu A + c, A^\tau \Sigma A)$ distribution.
(c) A random k-vector X has a k-dimensional normal distribution if and only if for any $c \in \mathcal{R}^k$, Xc^τ has a univariate normal distribution.
(d) Let X be a random k-vector having the $N_k(\mu, \Sigma)$ distribution. Let A be a $k \times l$ matrix and B be a $k \times m$ matrix. Then XA and XB are independent if and only if they are uncorrelated.
(e) Let (X_1, X_2) be a random k-vector having the $N_k(\mu, \Sigma)$ distribution with

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where X_1 is a random l-vector and Σ_{11} is an $l \times l$ matrix. Then the conditional p.d.f. of X_2 given $X_1 = x_1$ is

$$N_{k-l}\left(\mu_2 + (x_1 - \mu_1)\Sigma_{11}^{-1}\Sigma_{12}, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right),$$

where $\mu_i = EX_i$, $i = 1, 2$. (Hint: consider $X_2 - \mu_2 - (X_1 - \mu_1)\Sigma_{11}^{-1}\Sigma_{12}$ and $X_1 - \mu_1$.)

52. Let ϕ_n be the ch.f. of a probability measure P_n, $n = 1, 2, ...$ Let $\{a_n\}$ be a sequence of nonnegative numbers with $\sum_{n=1}^{\infty} a_n = 1$. Show that $\sum_{n=1}^{\infty} a_n \phi_n$ is a ch.f. and find its corresponding probability measure.

53. Let Y be a random variable having the noncentral chi-square distribution $\chi_k^2(\delta)$. Show that
    (a) Y has the Lebesgue p.d.f. given by (1.31);
    (b) the ch.f. of Y is $(1 - 2\sqrt{-1}t)^{k/2} e^{\sqrt{-1}\delta t/(1-2\sqrt{-1}t)}$;
    (c) $E(Y) = k + \delta$ and $\mathrm{Var}(Y) = 2k + 4\delta$.

54. Let T be a random variable having the noncentral t-distribution $t_n(\delta)$. Show that
    (a) T has the Lebesgue p.d.f. given by (1.32);
    (b) $E(T) = \delta\Gamma((n-1)/2)\sqrt{n/2}/\Gamma(n/2)$ when $n > 1$;
    (c) $\mathrm{Var}(T) = \frac{n(1+\delta^2)}{n-2} - \frac{\delta^2 n}{2}\left[\frac{\Gamma((n-1)/2)}{\Gamma(n/2)}\right]^2$ when $n > 2$.

55. Let $\mathtt{F}$ be a random variable having the noncentral F-distribution $F_{n_1,n_2}(\delta)$. Show that
    (a) $\mathtt{F}$ has the Lebesgue p.d.f. given by (1.33);
    (b) $E(\mathtt{F}) = \frac{n_2(n_1+\delta)}{n_1(n_2-2)}$ when $n_2 > 2$;
    (c) $\mathrm{Var}(\mathtt{F}) = \frac{2n_2^2[(n_1+\delta)^2 + (n_2-2)(n_1+2\delta)]}{n_1^2(n_2-2)^2(n_2-4)}$ when $n_2 > 4$.

56. Let $X = N_n(\mu, I_n)$. Apply Cochran's theorem (Theorem 1.5) to show that if $A^2 = A$, then XAX^τ has the noncentral chi-square distribution $\chi_r^2(\delta)$, where $r =$ rank of A and $\delta = \mu A \mu^\tau$.

57. Let $X_1, ..., X_n$ be independent and $X_i = N(0, \sigma_i^2)$, $i = 1, ..., n$. Let $\tilde{X} = \sum_{i=1}^n \sigma_i^{-2} X_i / \sum_{i=1}^n \sigma_i^{-2}$ and $\tilde{S}^2 = \sum_{i=1}^n \sigma_i^{-2}(X_i - \tilde{X})^2$. Apply Cochran's theorem to show that $\tilde{X}^2$ and $\tilde{S}^2$ are independent and that $\tilde{S}^2$ has the chi-square distribution χ_{n-1}^2.

58. Let $X = N_n(\mu, I_n)$ and $A_i^2 = A_i$, $i = 1, 2$. Show that a necessary and sufficient condition that XA_1X^τ and XA_2X^τ are independent is $A_1A_2 = 0$.

59. Prove Theorem 1.6. (Hint: first consider simple functions.)

60. Prove Proposition 1.12.

61. Let X and Y be random variables on $(\Omega, \mathcal{F}, P)$ and $\mathcal{A} \subset \mathcal{F}$ be a σ-field. Suppose that X is integrable and Y is bounded. Show that $E[YE(X|\mathcal{A})] = E[XE(Y|\mathcal{A})]$.

62. Let (X, Y) be a random 2-vector having a Lebesgue p.d.f. $f(x, y)$. Suppose that $E|X| < \infty$ and $Z = X + Y$. Show that

$$E(X|Z) = \frac{\int xf(x, Z - x)dx}{\int f(x, Z - x)dx}$$

without using Proposition 1.11.

63. (Convergence theorems for conditional expectations). Let $X_1, X_2, ...$ and X be integrable random variables on $(\Omega, \mathcal{F}, P)$ and $\mathcal{A} \subset \mathcal{F}$ be a σ-field. Show that
(a) (Fatou's lemma). If $X_n \geq 0$ for any n, then

$$E\left(\liminf_n X_n|\mathcal{A}\right) \leq \liminf_n E(X_n|\mathcal{A}) \text{ a.s.}$$

(b) (Monotone convergence theorem). If $0 \leq X_1 \leq X_2 \leq \cdots$ and $\lim_{n\to\infty} X_n = X$ a.s., then

$$E(X|\mathcal{A}) = \lim_{n\to\infty} E(X_n|\mathcal{A}) \text{ a.s.}$$

(c) (Dominated convergence theorem). Suppose that there is an integrable random variable Y such that $|X_n| \leq Y$ for any n and $\lim_{n\to\infty} X_n = X$ a.s. Then the result in (b) holds.

64. Let X be a nonnegative integrable random variable on $(\Omega, \mathcal{F})$ and $\mathcal{A} \subset \mathcal{F}$ be a σ-field. Show that

$$E(X|\mathcal{A}) = \int_0^\infty P\big(X > t|\mathcal{A}\big)dt.$$

65. Let X be an integrable random variable on $(\Omega, \mathcal{F}, P)$, $\mathcal{A} \subset \mathcal{F}$ be a σ-field, and f be a convex function on $\mathcal{R}$. Show that $f(E(X|\mathcal{A})) \leq E[f(X)|\mathcal{A}]$ a.s.

66. Show that two events A and B are independent if and only if two random variables I_A and I_B are independent.

67. Show that random variables X_i, $i = 1, ..., n$, are independent according to Definition 1.7 if and only if (1.26) holds.

68. Show that a random variable X is independent of itself if and only if X is constant a.s. Can X and $f(X)$ be independent for a Borel f?

69. Let X and Y be independent random variables on a probability space. Show that if $E|X|^a < \infty$ for some $a \geq 1$ and $E|Y| < \infty$, then $E|X+Y|^a \geq E|X+EY|^a$.

70. Let (X, Y) be a random 2-vector with the following Lebesgue p.d.f.:

$$f(x,y) = \begin{cases} 8xy & 0 \leq x \leq y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Find the marginal p.d.f.'s of X and Y. Are X and Y independent?

71. Let (X, Y, Z) be a random 3-vector with the following Lebesgue p.d.f.:

$$f(x,y,z) = \begin{cases} \frac{1-\sin x \sin y \sin z}{8\pi^3} & 0 \leq x \leq 2\pi, 0 \leq y \leq 2\pi, 0 \leq z \leq 2\pi \\ 0 & \text{otherwise.} \end{cases}$$

Show that X, Y, and Z are not independent, but are pairwise independent.

72. Let P_Y be a discrete distribution on $\{0, 1, 2, ...\}$ and $P_{X|Y=y}$ be the binomial distribution $Bi(p, y)$. Let (X, Y) be the random vector having the joint c.d.f. given by (1.44). Show that
(a) if Y has the Poisson distribution $P(\theta)$, then the marginal distribution of X is the Poisson distribution $P(p\theta)$;
(b) if $Y + r$ has the negative binomial distribution $NB(\pi, r)$, then the marginal distribution of $X + r$ is the negative binomial distribution $NB(\pi/[1-(1-p)(1-\pi)], r)$.

73. Let $X, X_1, X_2, ...$ be random vectors on a probability space. Show that $X_n \to_{a.s.} X$ if and only if for every $\epsilon > 0$,

$$\lim_{m\to\infty} P(\|X_n - X\| \leq \epsilon \text{ for all } n \geq m) = 1.$$

74. Let $X_1, X_2, ...$ be a sequence of identically distributed random variables and $Y_n = n^{-1} \max_{i \leq n} |X_i|$. Show that $Y_n \to_{a.s.} 0$ and $Y_n \to_{L_1} 0$.

75. Let $X, X_1, X_2, ...$ be random variables. Find an example for each of the following cases:
(a) $X_n \to_p X$, but $\{X_n\}$ does not converge to X a.s.
(b) $X_n \to_p X$, but $\{X_n\}$ does not converge to X in L_p for any $p > 0$.
(c) $X_n \to_d X$, but $\{X_n\}$ does not converge to X in probability (do not use Example 1.22).
(d) $X_n \to_p X$, but $\{g(X_n)\}$ does not converge to $g(X)$ in probability for some function g.

76. Show that (1.46) implies (1.45).

77. Let $X, X_1, X_2, ...$ be random k-vectors satisfying $P(\|X_n\| \geq c) \leq P(\|X\| \geq c)$ for all n and $c > 0$. Show that if $E\|X\| < \infty$, then $\{\|X_n\|\}$ is uniformly integrable.

78. Let $X_1, X_2, ...$ and $Y_1, Y_2, ...$ be random variables. Show that
(a) if $\{|X_n|\}$ and $\{|Y_n|\}$ are uniformly integrable, then $\{|X_n + Y_n|\}$ is uniformly integrable;
(b) if $\{|X_n|\}$ is uniformly integrable, then $\{|n^{-1}\sum_{i=1}^n X_i|\}$ is uniformly integrable.

79. Let $X, Y, X_1, X_2, ...$ be random k-vectors satisfying $X_n \to_p X$ and $P(\|X_n\| \leq \|Y\|) = 1$ for all n. Show that if $E\|Y\|^p < \infty$, then $X_n \to_{L_p} X$.

80. Show that if $X_n \to_d X$ and $X = c$ a.s., where $c \in \mathcal{R}^k$, then $X_n \to_p X$.

81. Show that if $X_n \to_d X$, then for every $\epsilon > 0$, there exists $M_\epsilon > 0$ such that $P(\|X_n\| > M_\epsilon) < \epsilon$.

82. Let X_n, Y_n, $n = 1, 2, ...$ be random k-vectors such that

$$\lim_{n\to\infty} P\left(\|X_n - Y_n\| \geq \epsilon\right) = 0$$

for any $\epsilon > 0$. Show that if $X_n \to_d X$ for a random vector X, then $Y_n \to_d X$.

83. Let $X_1, X_2, ..., X, Y$ be random k-vectors. Show that $X_n \to_p X$ and $X_n \to_p Y$ implies that $P(X = Y) = 1$.

84. Let $X, X_1, X_2, ...$ be random k-vectors and $Y, Y_1, Y_2, ...$ be random l-vectors. Suppose that $X_n \to_d X$, $Y_n \to_d Y$, and X_n and Y_n are independent for each n. Show that (X_n, Y_n) converges in distribution to a random $(k + l)$-vector.

85. Let X_n be a random variable having the $N(\mu_n, \sigma_n^2)$ distribution, $n = 1, 2,...,$ and X be a random variable having the $N(\mu, \sigma^2)$ distribution. Show that $X_n \to_d X$ if and only if $\mu_n \to \mu$ and $\sigma_n \to \sigma$.

86. Suppose that X_n is a random variable having the binomial distribution $Bi(p_n, n)$. Show that if $np_n \to \theta > 0$, then $X_n \to_d X$, where X has the Poisson distribution $P(\theta)$.

87. Let f_n be the Lebesgue p.d.f. of the t-distribution t_n, $n = 1, 2,....$ Show that $f_n(x) \to f(x)$ for any $x \in \mathcal{R}$, where f is the Lebesgue p.d.f. of the standard normal distribution.

88. Let $X_1, X_2, ...,\ Y_1, Y_2, ...,\ Z_1, Z_2, ...$ be random variables. Prove the following statements.
(a) If $X_n \to_d X$ for a random variable X, then $X_n = O_p(1)$.
(b) If $X_n = O_p(Z_n)$ and $P(Y_n = 0) = 0$ for all n, then $X_n Y_n = O_p(Y_n Z_n)$.
(c) If $X_n = O_p(Z_n)$ and $Y_n = O_p(Z_n)$, then $X_n + Y_n = O_p(Z_n)$.
(d) If $E|X_n| = O(a_n)$ for a sequence of positive numbers $a_1, a_2, ...,$ then $X_n = O_p(a_n)$.

89. Let $X, X_1, X_2, ...$ be random variables such that $X_n \to_{a.s.} X$. Show that $\sup_n |X_n| = O_p(1)$.

90. Prove Theorem 1.10.

91. Show by example that $X_n \to_d X$ and $Y_n \to_d Y$ does not necessarily imply that $g(X_n, Y_n) \to_d g(X, Y)$, where g is a continuous function on $\mathcal{R}^2$.

92. Prove Theorem 1.11(ii)-(iii) and Theorem 1.12(ii).

93. Let $U_1, U_2, ...$ be i.i.d. random variables having the uniform distribution on $[0, 1]$ and $Y_n = \left(\prod_{i=1}^n U_i\right)^{-1/n}$. Show that $\sqrt{n}(Y_n - e) \to_d N(0, e^2)$.

94. Let X_n be a random variable having the Poisson distribution $P(n\theta)$, where $\theta > 0$ and $n = 1, 2,....$ Show that $X_n/n \to_{a.s.} \theta$. Show that the same conclusion can be drawn if X_n has the binomial distribution $Bi(\theta, n)$.

95. Let $X_1, ..., X_n$ be i.i.d. random variables with

$$P(X_1 = \pm x) = \left(\sum_{x=3}^{\infty} \frac{1}{x^2 \log x}\right)^{-1} \frac{1}{2x^2 \log x}, \quad x = 3, 4, ...$$

Show that $E|X_1| = \infty$ but $n^{-1} \sum_{i=1}^n X_i \to_p 0$, using Theorem 1.13(i).

96. Let $X_1, ..., X_n$ be i.i.d. random variables with $\text{Var}(X_1) < \infty$. Show that

$$\frac{2}{n(n+1)} \sum_{j=1}^{n} j X_j \to_p EX_1.$$

97. Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of random variables such that

$$P(X_n \le t, Y_n \ge t + \epsilon) + P(X_n \ge t + \epsilon, Y_n \le t) = o(1)$$

for any fixed $t \in \mathcal{R}$ and $\epsilon > 0$. Show that $X_n - Y_n = o_p(1)$.

98. Show that Liapunov's condition (1.54) implies Lindeberg's condition, i.e., condition (1.53) with arbitrary $\epsilon > 0$.

99. Prove Corollaries 1.2 and 1.3.

100. Let X_n be a random variable having the Poisson distribution $P(n\theta)$, where $\theta > 0$, $n = 1, 2, ...$ Show that $(X_n - n\theta)/\sqrt{n\theta} \to_d N(0, 1)$.

101. Let $X_1, ..., X_n$ be random variables and $\{\mu_n\}$, $\{\sigma_n\}$, $\{a_n\}$, and $\{b_n\}$ be sequences of real numbers with $\sigma_n \geq 0$ and $a_n \geq 0$. Suppose that X_n is asymptotically distributed as $N(\mu_n, \sigma_n^2)$. Show that $a_n X_n + b_n$ is asymptotically distributed as $N(\mu_n, \sigma_n^2)$ if and only if $a_n \to 1$ and $[\mu_n(a_n - 1) + b_n]/\sigma_n \to 0$.

102. Let $X_1, X_2, ...$ be independent random variables such that X_j has the uniform distribution on $[-j, j]$, $j = 1, 2, ...$ Show that Lindeberg's condition is satisfied and state the resulting CLT.

103. Let $X_1, X_2, ...$ be independent random variables with $P(X_j = \pm\sqrt{j}\,)$ $= 0.5$, $j = 1, 2, ...$ Can we apply Theorem 1.15 to $\{X_j\}$ by checking Liapunov's condition (1.54)?

104. Let $X_1, X_2, ...$ be independent random variables with $P(X_j = -j^a) = P(X_j = j^a) = P(X_j = 0) = 1/3$, where $a > 0$, $j = 1, 2, ...$ Can we apply Theorem 1.15 to $\{X_j\}$ by checking Liapunov's condition (1.54)?

105. Let $X_1, X_2, ...$ be independent random variables such that for $j = 1, 2, ...,$

$$P(X_j = \pm j^a) = \frac{1}{6j^{2(a-1)}} \quad \text{and} \quad P(X_j = 0) = 1 - \frac{1}{3j^{2(a-1)}},$$

where $a > 1$ is a constant. Show that Lindeberg's condition is satisfied if and only if $a < 1.5$.

106. Suppose that X_n is a random variable having the binomial distribution $Bi(\theta, n)$, where $0 < \theta < 1$, $n = 1, 2, ...$ Define

$$Y_n = \begin{cases} \log(X_n/n) & X_n \geq 1 \\ 1 & X_n = 0. \end{cases}$$

Show that $Y_n \to_{a.s.} \log\theta$ and $\sqrt{n}(Y_n - \log\theta) \to_d N\left(0, \frac{1-\theta}{\theta}\right)$.