CHAPTER ONE

Probability and Statistics

The organization of this book is such that by the time reader gets to the last chapter, all necessary terminology and methods of solutions of standard mathematical background has been covered. Thus, we start the book with the basics of probability and statistics, although we could have placed the chapter in a later location. This is because some chapters are independent of the others.

In this chapter, the basics of probability and some important properties of the theory of probability, such as discrete and continuous random variables and distributions, as well as conditional probability, are covered.

After the presentation of the basics of probability, we will discuss statistics. Note that there is still a dispute as to whether statistics is a subject on its own or a branch of mathematics. Regardless, statistics deals with gathering, analyzing, and interpreting data. Statistics is an important concept that no science can do without. Statistics is divided in two parts: *descriptive statistics* and *inferential statistics*. Descriptive statistics includes some important basic terms that are widely used in our day-to-day lives. The latter is based on *probability theory*. To discuss this part of the statistics, we include point estimation, interval estimation, and hypothesis testing.

We will discuss one more topic related to both probability and statistics, which is extremely necessary for business and industry, namely *reliability of a system*. This concept is also needed in applications such as queueing networks, which will be discussed in the last chapter.

In this chapter, we cover as much probability and statistics as we will need in this book, except some parts that are added for the sake of completeness of the subject.

1.1. BASIC DEFINITIONS AND CONCEPTS OF PROBABILITY

Nowadays, it has been established in the scientific world that since quantities needed are not quite often predictable in advance, randomness should be

Difference and Differential Equations with Applications in Queueing Theory, First Edition.
Aliakbar Montazer Haghighi and Dimitar P. Mishev.

accounted for in any realistic world phenomenon, and that is why we will consider random experiments in this book.

Determining probability, or chance, is to quantify the variability in the outcome or outcomes of a random experiment whose exact outcome or outcomes cannot be predicted by certainty. Satellite communication systems, such as radar, are built of electronic components such as transistors, integrated circuits, and diodes. However, as any engineer would testify, the components installed usually never function as the designer has anticipated. Thus, not only is the probability of failure to be considered, but the reliability of the system is also quite important, since the failure of the system may have not only economic losses but other damages as well. With probability theory, one may answer the question, "How reliable is the system?"

Definition 1.1.1. ***Basics***

(a) Any result of performing an experiment is called an *outcome* of that experiment. A set of outcomes is called an *event*.

(b) If occurrences of outcomes are not certain or completely predictable, the experiment is called a *chance* or *random experiment*.

(c) In a random experiment, sets of outcomes that cannot be broken down into smaller sets are called *elementary* (or *simple* or *fundamental*) *events*.

(d) An elementary event is, usually, just a singleton (a set with a single element, such as $\{e\}$). Hence, a combination of elementary events is just an *event*.

(e) When any element (or outcome) of an event happens, we say that the *event occurred*.

(f) The *union* (set of all elements, with no repetition) of all events for a random experiment (or the set of all possible outcomes) is called the *sample space*.

(g) In "set" terminology, an *event* is a *subset* of the sample space. Two events A_1 and A_2 are called *mutually exclusive* if their intersection is the empty set, that is, they are disjoint subsets of the sample space.

(h) Let $A_1, A_2, \ldots, A_n$ be mutually exclusive events such that $A_1 \cup A_2 \cup \ldots \cup A_n = \Omega$. The set of $\{A_1, A_2, \ldots, A_n\}$ is then called a *partition* of the sample space Ω.

(i) For an experiment, a collection or a set of all individuals, objects, or measurements of interest is called a (statistical) *population*.

For instance, to determine the average grade of the differential equation course for all mathematics major students in four-year colleges and universities in Texas, the totality of students majoring mathematics in the colleges and universities in the Texas constitutes the population for the study.

Usually, studying the population may not be practically or economically feasible because it may be quite time consuming, too costly, and/or impossible to identify all members of it. In such cases, sampling is being used.

(j) A portion, subset, or a part of the population of interest (finite or infinite number of them) is called a *sample*.

Of course, the sample must be *representative* of the entire population in order to make any prediction about the population.

(k) An element of the sample is called a *sample point*. By *quantification* of the sample we mean changing the sample points to numbers.

(l) The *range* is the difference between the smallest and the largest sample points.

(m) A sample selected such that each element or unit in the population has the same chance to be selected is called a *random sample*.

(n) The *probability of an event A*, denoted by $P(A)$, is a number between 0 and 1 (inclusive) describing likelihood of the event A to occur.

(o) An event with probability 1 is called an *almost sure event*. An event with probability 0 is called a *null* or an *impossible event*.

(p) For a sample space with n (finite) elements, if all elements or outcomes have the same chance to occur, then we assign probability $1/n$ to each member. In this case, the sample space is called *equiprobable*.

For instance, to choose a digit at random from 1 to 5, we mean that every digit of {1, 2, 3, 4, 5} has the same chance to be picked, that is, all elementary events in {1}, {2}, {3}, {4}, and {5} are equiprobable. In that case, we may associate probability 1/5 to each digit singleton.

(q) If a random experiment is repeated, then the chance of occurrence of an outcome, intuitively, will be approximated by the ratio of occurrences of the outcome to the total number of repetitions of the experiment. This ratio is called the *relative frequency*.

Axioms of Probabilities of Events

We now state properties of probability of an event A through *axioms of probability*. The Russian mathematician Kolmogorov originated these axioms in early part of the twentieth century. By an axiom, it is meant a statement that cannot be proved or disproved. Although all probabilists accept the three axioms of probability, there are axioms in mathematics that are still controversial, such as the axiom of choice, and not accepted by some prominent mathematicians.

Let Ω be the sample space, $\mathcal{B}$ the set function containing all possible events drawn from Ω, and P denote the probability of an event. The triplet $(\Omega, \mathcal{B}, P)$ is then called the *probability space*. Later, after we define a random variable, we will discuss this space more rigorously.

Axioms of Probability

Axiom A1. $0 \le P(A) \le 1$ for each event A in $\mathcal{B}$.

Axiom A2. $P(\Omega) = 1$.

Axiom A3. If A_1 and A_2 are *mutually exclusive* events in $\mathcal{B}$, then:

$$P(A_1 \cup A_2) = P(A_1) + P(A_2),$$

where mutually exclusive events are events that have no sample point in common, and the symbol $\cup$ means the union of two sets, that is, the set of all elements in both set without repetition.

Note that the axioms stated earlier are for events. Later, we will define another set of axioms of probability involving random variables.

If the occurrence of an event has influence on the occurrence of other events under consideration, then the probabilities of those events change.

Definition 1.1.2
Suppose $(\Omega, \mathcal{B}, P)$ is a probability space and B is an event (i.e., $B \in \mathcal{B}$) with positive probability, $P(B) > 0$. The *conditional probability of A given B*, denoted by $P(A|B)$, defined on $\mathcal{B}$, is then given by:

$$P(A|B) = \frac{P(AB)}{P(B)}, \text{ for any event } A \text{ in } \mathcal{B}, \text{ and for } P(B) > 0. \tag{1.1.1}$$

If $P(B) = 0$, then $P(A|B)$ is not defined. Under the condition given, we will have a new triple, that is, a new probability space $(\Omega, \mathcal{B}, P(A|B))$. This space is called the *conditional probability space induced on* $(\Omega, \mathcal{B}, P)$, *given* B.

Definition 1.1.3
For any two events A and B with conditional probability $P(B \mid A)$ or $P(A \mid B)$, we have the *multiplicative law*, which states:

$$P(AB) = P(B|A)P(A) = P(A|B)P(B). \tag{1.1.2}$$

We leave it as an exercise to show that for n events $A_1, A_2, \ldots, A_n$, we have:

$$P(A_1A_2 \ldots A_n) = P(A_1) \times P(A_2|A_1) \times P(A_3|A_1A_2) \times \cdots \times P(A_n|A_1A_2 \ldots A_{n-1}). \tag{1.1.3}$$

Definition 1.1.4
We say that events A and B are *independent* if and only if:

$$P(AB) = P(A)P(B). \tag{1.1.4}$$

It will be left as an exercise to show that if events A and B are independent and $P(B) > 0$, then:

$$P(A \mid B) = P(A). \tag{1.1.5}$$

It can be shown that if $P(B) > 0$ and (1.1.5) is true, then A and B are independent. For proof, see Haghighi et al. (2011a, p. 139).

The concept of independence can be extended to a finite number of events.

Definition 1.1.5
Events $A_1, A_2, \ldots, A_n$ are *independent* if and only if the probability of the intersection of any subset of them is equal to the product of corresponding probabilities, that is, for every subset $\{i_1, \ldots, i_k\}$ of $\{1, \ldots, n\}$ we have:

$$P\{(A_{i_1} A_{i_2} \ldots A_{i_n})\} = P(A_{i_1}) \times P(A_{i_2}) \times \cdots \times P(A_{i_k}). \tag{1.1.6}$$

As one of the very important applications of conditional probability, we state the following theorem, whose proof may be found in Haghighi et al. (2011a):

Theorem 1.1.1. ***The Law of Total Probability***
Let $A_1, A_2, \ldots, A_n$ be a partition of the sample space Ω. For any given event B, we then have:

$$P(B) = \sum_{i=1}^{n} P(A_i) P(B \mid A_i). \tag{1.1.7}$$

Theorem 1.1.1 leads us to another important application of conditional probability. Proof of this theorem may also found in Haghighi et al. (2011a).

Theorem 1.1.2. ***Bayes' Formula***
Let $A_1, A_2, \ldots, A_n$ be a partition of the sample space Ω. If an event B occurs, the probability of any event A_j given an event B is:

$$P(A_j \mid B) = \frac{P(A_j) P(B \mid A_j)}{\sum_{i=1}^{n} P(A_i) P(B \mid A_i)}, \quad j = 1, 2, \ldots, n. \tag{1.1.8}$$

Example 1.1.1
Suppose in a factory three machines A, B, and C produce the same type of products. The percent shares of these machines are 20, 50, and 30, respectively. It is observed that machines A, B, and C produce 1%, 4%, and 2% defective items, respectively. For the purpose of quality control, a produced item is chosen at random from the total items produced in a day. Two questions to answer:

1. What is the probability of the item being defective?
2. Given that the item chosen was defective, what is the probability that it was produced by machine B?

Answers

To answer the first question, we denote the event of defectiveness of the item chosen by D. By the law of total probability, we will then have:

$$\begin{aligned} P(D) &= P(\mathrm{A})P(D|\mathrm{A}) + P(\mathrm{B})P(D|B) + P(\mathrm{C})P(D|C) \\ &= 0.20 \times 0.01 + 0.50 \times 0.04 + 0.30 \times 0.20 \\ &= 0.002 + 0.020 + 0.006 = 0.028. \end{aligned}$$

Hence, the probability of the produced item chosen at random being defective is 2.8%.

To answer the second question, let the conditional probability in question be denoted by $P(\mathrm{B} \mid D)$. By Bayes' formula and answer to the first question, we then have:

$$P(\mathrm{B}|D) = \frac{P(\mathrm{B})P(D|\mathrm{B})}{P(D)} = \frac{0.50 \times 0.04}{0.028} = 0.714.$$

Thus, the probability that the defective item chosen be produced by machine C is 71.4%.

Example 1.1.2

Suppose there are three urns that contain black and white balls as follows:

$$\begin{cases} \text{Urn 1:} & 2\ \textit{blacks} \\ \text{Urn 2:} & 2\ \textit{whites} \\ \text{Urn 1:} & 1\ \textit{black} \text{ and } 1\ \textit{white}. \end{cases} \tag{1.1.9}$$

A ball is drawn randomly and it is "white." Discuss possible probabilities.

Discussion

The sample space Ω is the set of all pairs $(\cdot,\cdot)$, where the first dot represents the urn number (1, 2, or 3) and the second represents the color (black or white). Let U_1, U_2 and U_3 denote events that drawing was chosen from, respectively. Assuming that urns are identical and balls have equal chances to be chosen, we will then have:

$$P(U_1) = P(U_2) = P(U_3) = \frac{1}{3}. \tag{1.1.10}$$

Also, $U_1 = (1,\cdot)$, $U_2 = (2,\cdot)$, $U_3 = (3,\cdot)$.

Let W denote the event that a white ball was drawn, that is, $W = \{(\cdot, w)\}$. From (1.1.9), we have the following conditional probabilities:

$$P(W|U_1) = 0, \quad P(W|U_2) = 1, \quad P(U_3) = \frac{1}{2}. \tag{1.1.11}$$

From Bayes' rule, (1.1.9), (1.1.10), and (1.1.11), we have:

$$P(U_1|W) = \frac{P(W|U_1)P(U_1)}{P(W|U_1)P(U_1) + P(W|U_2)P(U_2) + P(W|U_3)P(U_3)}, \tag{1.1.12}$$

$$= 0. \tag{1.1.13}$$

Note that denominator of (1.1.12) is:

$$0 + (1)\left(\frac{1}{3}\right) + \left(\frac{1}{2}\right)\left(\frac{1}{3}\right) = \frac{1}{3} + \frac{1}{6} = \frac{1}{2}. \tag{1.1.14}$$

Using (1.1.14), we have:

$$P(U_2|W) = \frac{P(W|U_2)P(U_2)}{P(W|U_1)P(U_1) + P(W|U_2)P(U_2) + P(W|U_3)P(U_3)}$$
$$= \frac{(1)\left(\frac{1}{3}\right)}{\frac{1}{2}} = \frac{2}{3}. \tag{1.1.15}$$

Again, using (1.1.14), we have:

$$P(U_3|W) = \frac{P(W|U_3)P(U_3)}{P(W|U_1)P(U_1) + P(W|U_2)P(U_2) + P(W|U_3)P(U_3)}$$
$$= \frac{\left(\frac{1}{2}\right)\left(\frac{1}{3}\right)}{\frac{1}{2}} = \frac{1}{3}. \tag{1.1.16}$$

Now, observing from (1.1.13), (1.1.15), and (1.1.16), there is a better chance that the ball was drawn from the second urn. Hence, if we assume that the ball was drawn from the second urn, there is one white ball that remains in it. That is, we will have the three urns with 0, 1, and 1 white ball, respectively, in urns 1, 2, and 3.

1.2. DISCRETE RANDOM VARIABLES AND PROBABILITY DISTRIBUTION FUNCTIONS

As we have seen so far, elements of a sample space are not necessarily numbers. However, for convenience, we would rather have them so. This is done through

what is called a *random variable*. In other words, a *random variable* quantifies the sample space. That is, a *random variable* assigns numerical (or set) labels to the sample points. Formally, we define a random variable as follows:

Definition 1.2.1
A *random variable* is a function (or a mapping) on the sample space.

We note that a random variable is really neither a variable (as known independent variable) nor random, but as mentioned, it is just a function. Also note that sometimes the range of a random variable may not be numbers. This is simply because we defined a random variable as a mapping. Thus, it maps elements of a set into some elements of another set. Elements of either set do not have to necessarily be numbers.

There are two main types of random variables, namely, *discrete* and *continuous*. We will discuss each in detail.

Definition 1.2.2
A *discrete random variable* is a function, say X, from a countable sample space, Ω (that could very well be a numerical set), into the set of real numbers.

Example 1.2.1
Suppose we are to select two digits from 1 to 6 such that the sum of the two numbers selected equals 7. Assume that repetition is not allowed. The sample space under consideration will then be $S = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$, which is discrete. This set can also be described as $S = \{(i, j): i + j = 7, i, j = 1, 2, \ldots, 6\}$.

Now, the random variable X can be defined by $X((i, j)) = k, k = 1, 2, \ldots, 6$. That is, the range of X is the set $\{1, 2, 3, 4, 5, 6\}$ such that, for instance, $X((1, 6)) = 1$, $X((2, 5)) = 2$, $X((3, 4)) = 3$, $X((4, 3)) = 4$, $X((5, 2)) = 5$, and $X((6, 1)) = 6$. In other words, the discrete random variable X has quantified the set of ordered pairs S to a set of positive integers from 1 to 6.

Example 1.2.2
Toss a fair coin three times. Denoting heads by H and tails by T, the sample space will then contain eight triplets as $\Omega = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$. Each tossing will result in either heads or tails. Thus, we might define the random variable X to take values 1 and 0 for heads and tails, respectively, at the jth tossing. In other words,

$$X_j = \begin{cases} 1, & \text{if } j\text{th outcome is heads,} \\ 0, & \text{if } j\text{th outcome is tails.} \end{cases}$$

Hence, $P\{X_j = 0\} = 1/2$ and $P\{X_j = 1\} = 1/2$. Now from the sample space we see that the probability of the element HTH is:

$$P\{X_1 = 1, X_2 = 0, X_3 = 1\} = \frac{1}{8}. \tag{1.2.1}$$

In contrast, product of individual probabilities is:

$$P\{X_1 = 1\} \times P\{X_2 = 0\} \times P\{X_3 = 1\} = \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{8}. \tag{1.2.2}$$

From (1.2.1) and (1.2.2), we see that X_1, X_2, and X_3 are mutually independent.

Now suppose we define X and Y as the total number of heads and tails, respectively, after the third toss. The probability, then, of three heads and three tails is obviously zero, since these two events cannot occur at the same time, that is, $P\{X = 3, Y = 3\} = 0$. However, from the sample space probabilities of individual events are $P\{X = 3\} = 1/8$ and $P\{Y = 3\} = 1/8$. Thus, the product is:

$$P\{X = 3\} \times P\{Y = 3\} = \frac{1}{8} \times \frac{1}{8} = \frac{1}{64} \neq 0.$$

Hence, X and Y, in this case, are not independent.

One of the useful concepts using random variable is the *indicator function* (or *indicator random variable* that we will define in the next section.

Definition 1.2.3
Let A be an event from the sample space Ω. The random variable $I_A(\omega)$ for $\omega \in A$ defined as:

$$I_A(\omega) = \begin{cases} 1, & \text{if} \quad \omega \in A, \\ 0, & \text{if} \quad \omega \in A^c, \end{cases} \tag{1.2.3}$$

is called the indicator function (or indicator random variable).

Note that for every $\omega \in \Omega$, $I_\Omega(\omega) = 1$ and $I_\phi(\omega) = 0$.

We leave it as an exercise for the reader to show the following properties of random variables:

(a) if X and Y are two discrete random variables, then $X \pm Y$ and XY are also random variables, and
(b) if $\{Y = 0\}$ is empty, X/Y is also a random variable.

The way probabilities of a random variable are distributed across the possible values of that random variable is generally referred to as the *probability distribution* of that random variable. The following is the formal definition.

Definition 1.2.4
Let X be a discrete random variable defined on a sample space Ω and x is a typical element of the range of X. Let p_x denote the probability that the random variable X takes the value x, that is,

$$p_x = P([X = x]) \quad \text{or} \quad p_x = P(X = x), \tag{1.2.4}$$

where p_X is called the *probability mass function* (pmf) of X and also referred to as the (*discrete*) *probability density function* (pdf) of X.

Note that $\sum_x p_x = 1$, where x varies over all possible values for X.

Example 1.2.3
Suppose a machine is in either "good working condition" or "not good working condition." Let us denote "good working condition" by 1 and " not good working condition" by 0. The sample space of states of this machine will then be $\Omega = \{0, 1\}$. Using a random variable X, we define $P([X = 1])$ as the probability that the machine is in "good working condition" and $P([X = 0])$ as the probability that the machine is not in "good working condition." Now if $P([X = 0]) = 4/5$ and $P([X = 0]) = 1/5$, then we have a distribution for X.

Definition 1.2.5
Suppose X is a discrete random variable, and x is a real number from the interval $(-\infty, x]$. Let us define $F_X(x)$ as:

$$F_X(x) = P([X \leq x]) = \sum_{n=-\infty}^{x} p_n, \tag{1.2.5}$$

where p_n is defined as $P([X = n])$ or $P(X = n)$. $F_X(x)$ is then called the *cumulative distribution function* (*cdf*) *for X*.

Note that from the set of axioms of probability mentioned earlier, for all x, we have:

$$p_x \geq 0, \quad \text{and} \quad \sum_x p_x = 1. \tag{1.2.6}$$

We now discuss selected important discrete probability distribution functions. Before that, we note that a random experiment is sometimes called a *trial*.

Definition 1.2.6
A *Bernoulli trial* is a trial with exactly two possible outcomes. The two possible outcomes of a Bernoulli trial are often referred to as *success* and *failure* denoted by *s* and *f*, respectively. If a Bernoulli trial is repeated independently n times with the same probabilities of success and failure on each trial, then the process is called *Bernoulli trials*.

Notes:

(1) From Definition 1.2.6, if the probability of *s* is p, $0 \leq p \leq 1$, then, by the second axiom of probability, the probability of *f* will be $q = 1 - p$.
(2) By its definition, in a Bernoulli trial, the sample space for each trial has two sample points.

Definition 1.2.7
Now, let X be a random variable taking values 1 and 0, corresponding to success and failure, respectively, of the possible outcome of a Bernoulli trial, with p ($p > 0$) as the probability of success and q as probability of failure. We will then have:

$$P(X = k) = p^k q^{1-k}, \quad k = 0, 1. \tag{1.2.7}$$

Formula (1.2.7) is the probability distribution function (pmf) of the Bernoulli random variable X.

Note that (1.2.7) is because first of all, $p^k q^{1-k} > 0$, and second, $\sum_{k=0}^{1} p^k q^{1-k} = p + q = 1$.

Example 1.2.4
Suppose we test 6 different objects for strength, in which the probability of breakdown is 0.2. What is the probability that the third object test be successful is, that is, does not breakdown?

Answer
In this case, we have a sequence of six Bernoulli trials. Let us assume 1 for a success and 0 for a failure. We would then have a 6-tuple (001000) to symbolize our objective. Hence, the probability would be (0.2)(0.2)(0.8)(0.2)(0.2)(0.2) = 0.000256.

Now suppose we repeat a Bernoulli trial independently finitely many times. We would then be interested in the probability of given number of times that one of the two possible outcomes occurs regardless of the order of their occurrences. Therefore, we will have the following definition:

Definition 1.2.8
Suppose X_n is the random variable representing the number of successes in n independent Bernoulli trials. Denote the *pmf* of X_n by $\mathrm{B_k} = b(k; n, p)$. $\mathrm{B_k} = b(k; n, p)$ is called the *binomial distribution function* with *parameters n* and *p* of the random variable X, where the parameters n, p and the number k refer to the number of independent trials, probability of *success* in each trial, and the number of successes in n trials, respectively. In this case, X is called the *binomial random variable*. The notation $X \sim b(k; n, p)$ is used to indicate that X is a binomial random variable with parameters n and p.

We leave it as an exercise to prove that:

$$\mathrm{B_k} = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, 2, \ldots, n, \tag{1.2.8}$$

where $q = 1 - p$.

Example 1.2.5
Suppose two identical machines run together, each to choose a digit from 1 to 9 randomly five times. We want to know what the probability that a sum of 6 or 9 appears k times ($k = 0, 1, 2, 3, 4, 5$) is.

Answer
To answer the question, note that we have five independent trials. The sample space in this case for one trial has 81 sample points and can be written in a matrix form as follows:

$$\begin{pmatrix} (1,1) & (1,2) & \cdots & (1,8) & (1,9) \\ (2,1) & (2,2) & \cdots & (2,8) & (2,9) \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ (8,1) & (8,2) & \ddots & (8,8) & (8,9) \\ (9,1) & (9,2) & \cdots & (9,8) & (9,9) \end{pmatrix}.$$

There are 13 sample points, where the sum of the components is 6 or 9. They are:

$$(1,5), (2,4), (3,3), (4,2), (5,1), (1,8), (2,7), (3,6), (4,5), (5,4), (6,3), (7,2), (8,1).$$

Hence, the probability of getting a sum as 6 or 9 on one selection of both machines together (i.e., probability of a success) is $p = 13/81$. Now let X be the random variable representing the total times a sum as 6 or 9 is obtained in 5 trials. Thus, from (1.2.8), we have:

$$P([X = k]) = \binom{5}{k}\left(\frac{13}{81}\right)^k \left(\frac{68}{81}\right)^{5-k}, \quad k = 0, 1, 2, 3, 4, 5.$$

For instance, the probability that the sum as 6 or 9 does not appear at all will be $(68/81)^5 = 0.42$, that is, there is a $(100 - 42) = 58\%$ chance that we do get at least a sum as 6 or 9 during the five trials.

Based on a sequence of independent Bernoulli trials, we now define two other important discrete random variables. Consider a sequence of independent Bernoulli trials with probability of success in each trial as p, $0 \le p \le 1$. Suppose we are interested in the total number of trials required to have the rth success, r being a fixed positive integer. The answer is in the following definition:

Definition 1.2.9
Let X be a random variable with *pmf* as:

$$f(k; r, p) = \binom{r+k-1}{k} p^r q^k, \quad k = 0, 1, \ldots. \tag{1.2.9}$$

Formula (1.2.9) is then called a *negative binomial* (or *Pascal*) *probability distribution function* (or *binomial waiting time*). In particular, if $r = 1$ in (1.2.9), then we will have:

$$f(k; 1, p) = P(x = k + 1) = pq^k, \quad k = 0, 1, , \ldots. \tag{1.2.10}$$

The pmf given by (1.2.10) is called a geometric probability distribution function.

Example 1.2.6
As an example, suppose a satellite company finds that 40% of call for services received need advanced technology service. Suppose also that on a particular crazy day, all tickets written are put in a pool and requests are drawn randomly for service. Finally, suppose that on that particular day there are four advance service personnel available. We want to find the probability that the fourth request for advanced technology service is found on the sixth ticket drawn from the pool.

Answer
In this problem, we have independent trials with $p = 0.4$ as probability of success, that is, in need of advanced technology service, on any trial. Let X represent the number of the tickets on which the fourth request in question is found. Thus,

$$P(X = 4) = \binom{6}{4}(0.4)^4 (0.6)^2 = 0.09216.$$

Example 1.2.7
We now want to derive (1.2.9) differently. Suppose treatment of a cancer patient may result in "response" or "no response." Let the probability of a response be p and for a no response be $1 - p$. Hence, the simple space in this case has two outcomes, simply, "response" and "no response." We now repeatedly treat other patients with the same medicine and observe the reactions. Suppose we are looking for the probability of the number of trials required to have exactly k "responses."

Answer
Denoting the sample space by S, $S = \{\text{response, no response}\}$. Let us define the random variable X on S to denote the number of trials needed to have exactly k responses. Let A be the event, in S, of observing $k - 1$ responses in the first $x - 1$ treatments. Let B be the event of observing a response at the xth treatment. Let also C be the event of treating x patients to obtain exactly k responses. Hence, $C = A \cap B$. The probability of C is:

$$P(C) = P(A \cap B) = P(A) \cdot P(B \mid A).$$

In contrast, $P(B \mid A) = p$ and:

$$P(A) = \binom{x-1}{k-1} p^{k-1}(1-p)^{x-k}.$$

Moreover, $P(X = x) = P(C)$. Hence:

$$P(X = x) = \binom{x-1}{k-1} p^{k}(1-p)^{x-k}, \quad x = k, k+1, \ldots. \tag{1.2.11}$$

We leave it as an exercise to show that (1.2.11) is equivalent to (1.2.9).

Definition 1.2.10
Let n represent a sample (sampling without replacement) from a finite population of size N that consists of two types of items n_1 of "defective," say, and n_2 of "nondefective," say, $n_1 + n_2 = n$. Suppose we are interested in the probability of selecting x "defective" items from the sample. n_1 must be at least as large as x. Hence, x must be less than or equal to the smallest of n and n_1. Thus,

$$p_x \equiv P(X = x) = \frac{\binom{n_1}{x} \times \binom{N-n_1}{n-x}}{\binom{N}{n}}, \quad x = 0, 1, 2, \ldots, \min(n, n_1), \tag{1.2.12}$$

defines the general form of *hypergeometric pmf* of the random variable X.

Notes:

i. If sampling would have been with replacement, distribution would have been binomial.
ii. p_x is the probability of waiting time for the occurrence of exactly x "defective" outcomes. We could think of this scenario as an urn containing N white and green balls. From the urn, we select a random sample (a sample selected such that each element has the same chance to be selected) of size n, one ball at a time without replacement. The sample consists of n_1 white and n_2 green balls, $n_1 + n_2 = n$. What is the probability of having x white balls drawn in a row? This model is called an *urn model*.
iii. If we let x_i equal to 1 if a defective item is selected and 0 if a nondefective item is selected, and let x be the total number of defectives selected, then $x = \sum_{i=1}^{n} x_i$. Now, if we consider selection of a defective item as a success, for instance, then we could also interoperate (1.2.12) as:

$$p_x = \frac{(\text{number of ways for } x \text{ successes}) \times (\text{number of ways for } n-x \text{ failures})}{\text{total number of ways to select}}. \tag{1.2.13}$$

Example 1.2.8
Suppose we have 100 balls in a box and 10 of them are red. If we randomly take out 40 of them (without replacement), what is the probability that we will have at least 6 red balls?

Answer
In this example, which is a hypergeometric distribution, if we assume that all 40 balls are withdrawn at the same time, $N = 100$, $n = 40$, $n_1 = 10$, "defective," is replaced by "red," and $n_2 = 90$, and "nondefective" is replaced by "nonred". The question is to find the probability of selecting at least six red balls. To find the probabilities, we need to calculate the probabilities of 7, 8, 9, and 10 red balls and sum them all or calculate $p \equiv 1 - P\{\text{number of red balls}\} \leq 5$. To do this, we could use statistical software, called Stata, for instance. However, using (1.2.12) or (1.2.13), we have:

$$P\{\text{number of red balls}\} \leq 5 = \frac{\binom{10}{0}\times\binom{90}{40-0}}{\binom{100}{40}} + \frac{\binom{10}{1}\times\binom{90}{40-1}}{\binom{100}{40}} + \frac{\binom{10}{2}\times\binom{90}{40-2}}{\binom{100}{40}} + \frac{\binom{10}{3}\times\binom{90}{40-3}}{\binom{100}{40}} + \frac{\binom{10}{4}\times\binom{90}{40-4}}{\binom{100}{4}} + \frac{\binom{10}{5}\times\binom{90}{40-5}}{\binom{100}{40}} = 0.846.$$

Thus, $p = 1 - 0.846 = 0.154 = 15.4\%$.

We caution that if one uses Excel formula as: "=HYPGEOMDIST (6,40,10,100)," which works out to be 10%, it is not quite right.

As our final important discrete random variable, we define the Poisson probability distribution function.

Definition 1.2.11
A *Poisson random variable* is a nonnegative random variable X such that:

$$p_k = P([X = x]) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, \ldots, \tag{1.2.14}$$

where λ is a constant. Formula (1.2.14) is called a *Poisson probability distribution function* with parameter λ.

Example 1.2.9
Suppose that the number of telephone calls arriving to a switchboard of an institution every working day has a Poisson distribution with parameter 20. What is the probability that there will be:

(a) 20 calls in one day?
(b) at least 30 calls in one day?
(c) at most 30 calls in one day?

Answers
Using $\lambda = 20$ in (1.2.12) we will have:

(a) $P_{30} = P([X = 30]) = \left(\left(e^{-20}\right)\left(20^{30}\right)/30!\right) = 0.0083.$
(b) $P([X \geq 30]) = \sum_{k=30}^{\infty}\left(e^{-20}20^{k}/k!\right) = 0.0218.$
(c) $P([X \leq 30]) = 1 - P([X \geq 30]) + P([X = 30])$
$= 1 - 0.0218 + 0.0083 = 0.9865.$

Let X be a binomial random variable with distribution function B_k and $\lambda = np$ be fixed. We will then leave it as an exercise to show that:

$$B_k = \lim_{\substack{n\to\infty,\\ p\to 0}} = \frac{\lambda^k e^{-k}}{k!}, \quad k = 0, 1, 2, \ldots. \tag{1.2.15}$$

1.3. MOMENTS OF A DISCRETE RANDOM VARIABLE

We now discuss some properties of a discrete distribution.

Definition 1.3.1
Suppose X is a discrete random variable defined on a sample space Ω with pmf of p_X. The *mathematical expectation* or simply *expectation* of X, or *expected value* of X, or the *mean* of X or *the first moment* of X, denoted by $E(X)$, is then defined as follows:If Ω is finite and the range of X is $\{x_1, x_2, \ldots, x_n\}$, then:

$$E(X) = \sum_{i=1}^{n} x_i p_{X_i}, \tag{1.3.1}$$

and if Ω is infinite and the range of X is $\{x_1, x_2, \ldots, x_n, \ldots\}$, then:

$$E(X) = \sum_{i=1}^{\infty} x_i p_{X_i}, \tag{1.3.2}$$

provided that the series converges. If Ω is finite and p_{X_i}, $i = 1, 2, \ldots, n$, is constant for all i's, say $1/n$, then the right-hand side of (1.3.1) will become $x_1 + x_2$

$+ \ldots + x_n/n$. This expression is denoted by $\bar{x}$ and is called *arithmetic average* of $x_1, x_2, \ldots, x_n$, that is,

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}. \tag{1.3.3}$$

p_{X_i}, $i = 1, 2, \ldots, n$ in (1.3.1), (1.3.2), and (1.3.3) is called the *weight* for the values of the random variable X. Hence, in (1.3.1) and (1.3.2), the weights vary and $E(X)$ is called the *weighted average*, while in (1.3.3) the weights are the same and $\bar{x}$ is called the *arithmetic average* or simply the *average*.

We next state some properties of the first moment without proof. We leave the proof as exercises.

Properties of the First Moment

1. The expected value of the indicator function $I_A(\omega)$ defined in (1.2.3) is $P(A)$, that is,

$$E(I_A) = P(A). \tag{1.3.4}$$

2. If c is a constant, then $E(c) = c$.
3. If c, c_1, and c_2 are constants and X and Y are two random variables, then:

$$E(cX) = cE(X) \quad \text{and} \quad E(c_1X + c_2Y) = c_1E(X) + c_2E(Y). \tag{1.3.5}$$

4. If $X_1, X_2, \ldots, X_n$ are n random variables, then:

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_1) + \cdots + E(X_n). \tag{1.3.6}$$

5. Let X_1 and X_2 be two independent random variables with marginal mass (density) functions p_{x_1} and p_{x_2}, respectively. If $E(X_1)$ and $E(X_2)$ exist, we will then have:

$$E(X_1X_2) = E(X_1)E(X_2). \tag{1.3.7}$$

6. For a finite number of random variables, that is, if $X_1, X_2, \ldots, X_n$ are n independent random variables, then:

$$E(X_1X_2 \ldots X_n) = E(X_1)E(X_2) \ldots E(X_n). \tag{1.3.8}$$

We now extend the concept of moments.

Definition 1.3.2
Let X be a discrete random variable and r a positive integer. $E(X^r)$ is then called the rth *moment* of X or *moment of order r* of X. In symbols:

$$E\left(X^r\right)=\sum_{k=1}^{\infty}x_k^r P\left(X=x_k\right). \tag{1.3.9}$$

Note that if $r = 1$, $E(X^r) = E(X)$, that is, the moment of first order or the first moment of X is just the expected value of X. The second moment, that is, $E(X^2)$ is also important, as we will see later.

Let us denote $E(X)$ by μ, that is, $E(X) \equiv \mu$. It is clear that if X is a random variable, so is $X - \mu$, where μ is a constant. However, since $E(X - \mu) = E(X) - E(\mu) = \mu - \mu = 0$, we can *center* X by choosing the new random variable $X - \mu$. This leads to the following definition.

Definition 1.3.3
The rth moment of the random variable $X - \mu$, denoted by $\mu_r(X)$ is defined by $E[(X - \mu)^r]$, and is called the rth *central moment* of X, that is,

$$\mu_r(X)=E(X-\mu)^r. \tag{1.3.10}$$

Note that the random variable $X - \mu$ measures the *deviation* of X from its mean. Thus, we have the next definition:

Definition 1.3.4
The *variance* of a random variable, X, denoted by $Var(X)$ or equivalently by $\sigma^2(X)$, or if there is no fear of confusion, just σ^2, is defined as the second central moment of X, that is,

$$\sigma^2(X)=E\left[(X-\mu)^2\right]. \tag{1.3.11}$$

The positive square root of the variance of a random variable X is called the *standard deviation* and is denoted by $\sigma(X)$.

It can easily be shown that if X is a random variable and μ is finite, then:

$$Var(X)=E\left(X^2\right)-\mu^2. \tag{1.3.12}$$

It can also be easily proven that if X is a random variable and c is a real number, then:

$$Var(X+c)=Var(X), \tag{1.3.13}$$

$$Var(cX)=c^2Var(X). \tag{1.3.14}$$

Example 1.3.1
Consider the Indicator function defined in (1.2.3). That is,

$$I_A(\omega)=\begin{cases}1, & \text{if} \quad \omega\in A,\\ 0, & \text{if} \quad \omega\in A^c.\end{cases}$$

The expected value of $I_A(\omega)$ is:

$$E(I_A(\omega)) = 1 \cdot P(A) + 0 \cdot [1 - P(A)] = P(A).$$

Example 1.3.2
Consider the Bernoulli random variable defined in Definition 1.2.7. Thus, the random variable X takes two values 1 and 0, for instance, for success and failure, respectively. The probability of success is assumed to be p. Thus, the expected value of X is:

$$E(X) = 1 \cdot p + 0 \cdot (1 - p) = p.$$

To find the variance, note that:

$$E(X^2) = (1^2 =) 1 \cdot p + (0^2 =) 0 \cdot (1 - p) = p.$$

Hence,

$$Var(X) = p - p^2 = p(1 - p).$$

Example 1.3.3
We want to find the mean and variance of the random variable X having binomial distribution defined in (1.2.8).

Answer
From (1.2.8), we have:

$$E(X) = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^{n} k \binom{n}{k} p^k (1-p)^{n-k}$$
$$= np \sum_{r=0}^{n-1} \binom{n-1}{r} p^r (1-p)^{n-r-1} = np.$$

We leave it as an exercise to show that the $Var(X) = np(1 - p)$.

Example 1.3.4
Consider the Poisson distribution defined by (1.2.14). We want to find the mean and variance of the random variable X having Poisson pmf as given in (1.2.14).

Answer

$$E(X) = \sum_{x=0}^{\infty} x \frac{e^{-\lambda} \lambda^x}{x!} = \lambda e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \lambda e^{-\lambda} e^{\lambda} = \lambda,$$

$$E(X^2) = \sum_{x=0}^{\infty} x^2 \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty} x\frac{\lambda^x}{(x-1)!} = e^{-\lambda}\sum_{x=0}^{\infty}(x-1+1)\frac{e^{-\lambda}}{(x-1)!}$$

$$= e^{-\lambda}\left[\sum_{x=1}^{\infty}\frac{(x-1)\lambda^x}{(x-1)!} + \sum_{x=1}^{\infty}\frac{\lambda^x}{(x-1)!}\right]$$

$$= e^{-\lambda}\left[\lambda^2\sum_{x=2}^{\infty}\frac{\lambda^{x-2}}{(x-2)!} + \lambda\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!}\right]$$

$$= e^{-\lambda}\left[\lambda^2 e^{\lambda} + \lambda e^{\lambda}\right] = \lambda^2 + \lambda,$$

$$Var(X) = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

1.4. CONTINUOUS RANDOM VARIABLES

So far we have been discussing discrete random variables, discrete distribution functions, and some of their properties. We now discuss continuous cases.

Definition 1.4.1
When the values of outcomes of a random experiment are real numbers (not necessarily integers or rational), the sample space, Ω, is a called a *continuous sample space*, that is, Ω is the entire real number set $\mathbb{R}$ or a subset of it (i.e., an interval or a union of intervals).

The set consisting of all subsets of real numbers $\mathbb{R}$ is extremely large and it will be impossible to assign probabilities to all of them. It has been shown in the theory of probability that a smaller set, say $\mathcal{B}$, may be chosen that contains all events of our interest. In this case, $\mathcal{B}$ is, loosely, referred to as the *Borel field*. We now pause to discuss Borel field more rigorously.

Definition 1.4.2
A nonempty set-function $\mathcal{F}$ is called a σ-*algebra* if it is closed under complements, and under finite or countable unions, that is,

(i) $A_1, A_2 \in \mathcal{F}$, then $A_1 \cup A_2$ and $A_1^c \in \mathcal{F}$, and
(ii) $A_i \in \mathcal{F}$, $i \geq 1$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

Note that axioms (i) and (ii) imply that, $\mathcal{F}$ should be closed under finite and countable intersections as well.

Example 1.4.1
The power set of a set X, $\mathcal{F}$, is a σ-algebra.

Definition 1.4.3
Earlier in this chapter we defined "function." The way it was defined, it was a "point function" since values were assigned to each point of a set. A *set function F* assigns values to sets or regions of the space.

Definition 1.4.4
A *measure,* μ, on a set $\mathcal{F}$ is a set function that assigns a real number to each subset of $\mathcal{F}$, (which intuitively determines the size of the set $\mathcal{F}$) such that:

i. $\mu(\phi) = 0$, where ϕ is the empty set,
ii. $\mu(A) \geq 0, \forall A \in \mathcal{F}$, that is, nonnegative, and
iii. if $\{A_i, i \in \mathbb{Z}\} \in \mathcal{F}$ is a finite or countable sequence of mutually disjoint sets in $\mathcal{F}$, then $\mu(\bigcup_{i\in\mathbb{Z}} A_i) = \sum_{i\in\mathbb{Z}} \mu(A_i)$, the countably additive axiom, where $\mathbb{Z}$ is the set of integers.

Notes:

1. What the third axiom says is that the measure of a "large" subset (the union of subsets A_i's) that can be partitioned into a finite (or countable) number of "smaller" disjoint subsets is the sum of the measures of the "smaller" subsets.
2. Generally speaking, if we were to associate a size to each subset of a given set so that we are consistent yet satisfying the other axioms of a measure, only trivial examples like the counting measure would be available. To remove this barrier, a measure would be defined only on a subcollection of all subsets, the so called *measurable subsets*, which are required to form a σ-algebra. In other words, elements of the σ-algebra are called *measurable sets*. This means that countable unions, countable intersections, and complements of measurable subsets are measurable.
3. Existence of a nonmeasurable set involves axiom of choice.
4. Main applications of measures are in the foundations of the Lebesgue integral, in Kolmogorov's axiomatization of probability theory, including ergodic theory. (Andrey Kolmogorov was a Russian mathematician who was a pioneer of probability theory.)
5. It can be proven that a measure is *monotone*, that is, if A is a subset of B, then the measure of A is less than or equal to the measure of B.

Let us now consider the following example.

Example 1.4.2
Consider the life span of a patient with cancer who is under treatment. Hence, the duration of the patient's life is a positive real number. This number is actually an outcome of our treatment (experiment). Let us denote this outcome by ω. Thus, the sample space is the set of all real numbers (of course, in reality, truncated positive real line). Now, we could include the nonpositive part of the real line to our sample space as long as probabilities assigned to them are zeros. Thus, the sample space would become just the real line. While the treatment goes on, we might ask, what is the probability that the patient dies before a preassigned time, say ω_i?

It might also be of interest to know the time interval, say $(\omega_{t_1}, \omega_{t_2})$, in which the dose level of a medicine needs to show a reaction.

To answer questions of these types, we would have to consider intervals $(-\infty, \omega_t)$ and $(\omega_{t_1}, \omega_{t_2})$ as events. Thus, we need to consider a family of sets, called Borel sets.

Definition 1.4.5
Let Ω be a sample space. A family (or a collection) $\mathcal{B}$ of subsets of Ω satisfies the following axioms:

Axiom B1. $\Omega \in \mathcal{B}$,

Axiom B2. If $A \in \mathcal{B}$, then $A^c \in \mathcal{B}$, that is, is closed under complement, and

Axiom B3. If $\{A_i, i \in \mathbb{Z}\}$ is a finite or countable family of subsets of Ω in $\mathcal{B}$, then also $\bigcup_{i \in \mathbb{Z}} A_i \in \mathcal{B}$, that is, $\mathcal{B}$ is closed under the union of at most countable many of its members, called the *class of events*. The class of events $\mathcal{B}$ satisfying Axioms B1–B3 is called a *Borel field* or a *σ-field* (reads as sigma-field, which is a σ-algebra).

Example 1.4.3
The family of events $\mathcal{B} = \{\phi, \Omega\}$ satisfies axioms B1–B3 stated in Definition 1.4.5 and hence is a Borel field. In fact, this is the smallest family that satisfies the axioms. This family is called the *trivial Borel field*.

Example 1.4.4
Let A be a nonempty set such that $\phi \subset A \subset \Omega$. Thus, $\{\phi, A, A^c, \Omega\}$ is the smallest Borel field that contains A.

Definition 1.4.6
The smallest Borel field of subsets of the real line $\mathbb{R}$ that contains all intervals $(-\infty, \omega)$ is called *Borel sets*.

Notes:

1. A subset of the real line $\mathbb{R}$ is a Borel set if and only if it belongs to the Borel field mentioned in Definition 1.4.5.
2. Since intersection of σ-algebras is again a σ-algebra, the Borel sets are intersection of all σ-algebras containing the collection of open sets in $\mathcal{T}$.
3. We may define Borel sets as follows: let $(X, \mathcal{T})$ be a topological space. The smallest σ-algebra containing the open sets in $\mathcal{T}$ is then called the collection of Borel sets0 in X, if such a collection exists.

Definition 1.4.7
Let X be a set and $\mathcal{B}$ be a Borel set. The ordered pair $(X, \mathcal{B})$ is then called a *measurable space*.

Definition 1.4.8
Let Ω be a sample space. Also let $P(\cdot)$ be a nonnegative function defined on a Borel set $\mathcal{B}$. The function $P(\cdot)$ is then called a *probability measure*, if and only if the following axioms M1–M3 are satisfied:

Axiom M1. $P(\Omega) = 1$.

Axiom M2. $P(A) \geq 0, \forall A \in \mathcal{B}$.

Axiom M3. If $\{A_i, i \in I\} \in \mathcal{B}$ is a finite or countable sequence of mutually disjoint sets (i.e., $A_i \cap A_j = \phi$, for each $i \neq j$) in $\mathcal{B}$, then $P(\bigcup_{i \in I} A_i) = \sum_{i \in I} P(A_i)$, *the countably additive axiom*.

Example 1.4.5
Let Ω be a countable set and $\mathcal{B}$ the set of all subsets of Ω. Next, let $P(A) = \sum_{\omega \in A} p(\omega)$, where $p(\omega) \geq 0$ and $\sum_{\omega \in \Omega} p(\omega) = 1$. The function $P(\cdot)$, then, is a probability measure.

Definition 1.4.9
The triplet $(\Omega, \mathcal{B}, P)$, where Ω is a sample space, $\mathcal{B}$ is a Borel set, and $P : \mathcal{B} \to [0, 1]$ is a probability measure, is called the *probability space*.

Note that a probability space is a measure space with a probability measure.

Definition 1.4.10
A measure λ on the real line $\mathbb{R}$ such that $\lambda((a, b]) = b - a, \forall a < b$, and $\lambda(\mathbb{R}) = \infty$, is called a *Lebesgue measure*.

Example 1.4.6
The Lebesgue measure of the interval $[0, 1]$ is its length, that is, 1.

Note that a particularly important example is the Lebesgue measure on a Euclidean space, which assigns the conventional length, area, and volume of Euclidean geometry to suitable subsets of the n-dimensional Euclidean space $\mathbb{R}^n$.

We now return to the discussion of a continuous random variable.

If $A_1, A_2, A_3, \ldots$ is a sequence of mutually exclusive events represented as intervals of $\mathbb{R}$ and $P(A_i), i = 1, 2, \ldots$, is the probability of the event $A_i, i = 1, 2, \ldots$, then, by the third axiom of probability, A3, we will have:

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i). \tag{1.4.1}$$

For a random variable, X, defined on a continuous sample space, Ω, the probability associated with the sample points for which the values of X falls on the interval $[a, b]$ is denoted by $P(a \leq X \leq b)$.

Definition 1.4.11
Suppose the function $f(x)$ is defined on the set of real numbers, $\mathbb{R}$, such that $f(x) \geq 0$, for all real x, and $\int_{-\infty}^{\infty} f(x)dx = 1$. Then, $f(x)$ is called a *continuous probability density function* (*pdf*) (or just *density function*) on $\mathbb{R}$ and it is denoted by, $f_X(x)$. If X is a random variable that its probability is described by a continuous pdf as:

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx, \text{ for any interval } [a, b], \tag{1.4.2}$$

then X is called a *continuous random variable*. The *probability distribution function* of X, denoted by $F_X(x)$, is defined as:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt. \tag{1.4.3}$$

Notes:

1. There is a significant difference between discrete pdf and continuous pdf. For a discrete pdf, $f_X = P(X = x)$ is a probability, while with a continuous pdf, $f_X(x)$ is not a probability. The best we can say is that $f_X(x)dx \approx P(x \leq X \leq x + dx)$ for all infinitesimally small dx.
2. If there is no fear of confusion, we will suppress the subscript "X" from $f_X(x)$ and $F_X(x)$ and write $f(x)$ and $F(x)$, respectively.

As it can be seen from (1.4.3), distribution function can be described as the area under the graph of the density function.

Note from (1.4.2) and (1.4.3) that if $a = b = x$, then:

$$P(x \leq X \leq x) = P(X = x) = \int_x^x f(t)dt = 0. \tag{1.4.4}$$

What (1.4.4) says is that if X is a continuous random variable, then the probability of any given point is zero. That is, for a continuous random variable to have a positive probability, we have to choose an interval.

Notes:

(a) From (1.4.4), we will have:

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b). \tag{1.4.5}$$

(b) Since by the fundamental theorem of integral calculus, $f_x(x) = dF_x(x)/dx$, the density function of a continuous random variable can be obtained as the derivative of the distribution function, that is, $F'_X(x) = f_X(x)$.

Conversely, the cumulative distribution function can be recovered from the probability density function with (1.4.3).

(c) $F_X(x)$ collects all the probabilities of values of X up to and including x. Thus, it is the *cumulative distribution function* (*cdf*) of X.

(d) For $x_1 < x$, the intervals $(-\infty, x_1]$ and $(x_1, x]$ are disjoint and their union is $(-\infty, x]$. Hence, from (1.4.2) and (1.4.3), if $a \le b$, then:

$$P(a \le X \le b) = P(X \le b) - P(X \le a) = F_X(b) - F_X(a). \tag{1.4.6}$$

(e) It can easily be verified that $F_X(-\infty) = 0$ and $F_X(\infty) = 0$ (why?).

(f) As it can be seen from (1.4.3), the concept and definition of cdf applies to both discrete and continuous random variables. If the random variable is discrete, $F_X(x)$ is the sum of f_x's. However, if the random variable is continuous, then the sum becomes a limit and eventually an integral of the density function. The most obvious difference between the cdf for continuous and discrete random variables is that F_X is a continuous function if X is continuous, while it is a step function if X is discrete.

1.5. MOMENTS OF A CONTINUOUS RANDOM VARIABLE

As part of properties of continuous random variables, we now discuss continuous moments. Before doing that, we note that in an integral when the variable of integration, X, is replaced with a function, say $F(x)$, the integral becomes a Stieltjes integral that looks as $\int_a^b dF(x)$, found by Stieltjes in late nineteenth century. If $F(x)$ is a continuous function and its derivative is denoted by $f(x)$, then the *Lebesgue–Stieltjes integral* becomes $\int_a^b dF(x) = \int_a^b f(x)\,dx$. Henri Lebesgue was a French mathematician (1875–1941) and Thomas Joannes Stieltjes was a Dutch astronomer and mathematician (1856–1899).

Definition 1.5.1
Let X be a continuous random variable defined on the probability space $(\Omega, \mathcal{B}, F)$ with pdf $f_X(x)$. The *mathematical expectation* or simply *expectation* of X, or *expected value* of X, or the *mean* of X or *the first moment* of X, denoted by $E(X)$, is then defined as the Lebesgue–Stieltjes integral:

$$E(X) = \int_\Omega X dF = \int_{-\infty}^{\infty} x f(x)\,dx, \tag{1.5.1}$$

provided that the integral exists. Formula (1.5.1), for a case for an arbitrary (continuous measurable) function of X, say $g(X)$, where X is a bounded random variable with continuous pdf $f_X(x)$, will be:

$$E(g(X)) = \int_\Omega g(X)\,dF = \int_{-\infty}^{\infty} g(x) f(x)\,dx, \tag{1.5.2}$$

provided that the integral converges absolutely.

Definition 1.5.2
The kth *moments* of the continuous random variable X with pdf $f_X(x)$, denoted by $E[X^k]$, $k = 1, 2, \ldots$, is defined as:

$$E\left(X^k\right) = \int_{-\infty}^{\infty} x^k f(x)\,dx, \tag{1.5.3}$$

provided that the integral exists, that is,

$$E\left(|X|^k\right) = \int_{-\infty}^{\infty} |x|^k\, f(x)\,dx < +\infty. \tag{1.5.4}$$

In other words, from (1.5.3) and (1.5.4), the kth of X exists if and only if the kth *absolute moment* of X, $E(|X|^k)$, is finite.

Notes:

(1) It can be shown that if the kth, $k = 1, 2, \ldots$ moment of a random variable (discrete or continuous) exists, then do all the lower order moments.
(2) Among standard known continuous distributions, *Cauchy probability distribution* with *density function*:

$$f_X(x) = \frac{1}{\pi\left(1+x^2\right)}, \tag{1.5.5}$$

is the only one that its kth, $k = 1, 2, \ldots$, moments do not exist for even values of k and exist for odd values of k in the sense that the Cauchy principal values of the integral exist and are equal to zero.

Several examples will be given in the next section.

1.6. CONTINUOUS PROBABILITY DISTRIBUTION FUNCTIONS

As in the discrete case, we now list a selected number of continuous probability distributions that we may be using in this book.

Definition 1.6.1
A continuous random variable X that has the probability density function:

$$f_X(x) = \begin{cases} \dfrac{1}{b-a}, & \text{if } a \le b, \\ 0, & \text{elsewhere.} \end{cases} \tag{1.6.1}$$

has a *uniform distribution* over an interval $[a, b]$.

It is left for the reader to show that (1.6.1) defines a probability density function and that the *uniform distribution function of X* is given by:

$$F_X(x) = \begin{cases} 0, & \text{if } x < a, \\ \dfrac{x-a}{b-a}, & \text{if } a \le x \le b, \\ 1, & \text{if } x > b. \end{cases} \tag{1.6.2}$$

Note that since the graphs of the uniform density and distribution functions have rectangular shapes, they are sometimes referred to as the *rectangular density functions* and *rectangular distribution functions*, respectively.

Example 1.6.1
Suppose X is distributed uniformly over [0, 10]. We want to find $P(3 < X \le 7)$. Thus:

$$P(3 < X \le 7) = \frac{1}{10}\int_3^7 dx = \frac{1}{10}(7-3) = 0.4.$$

Example 1.6.2
Suppose a counter registers events according to a Poisson distribution with rate 4. The counter begins at 8:00 A.M. and registers 1 event in 30 minutes. What is the probability that the event occurred by 8:20 A.M.?

Answer
We will restate the problem symbolically and then substitute the values of the parameters to answer the question. We have a Poisson distribution with rate λ and registration rate of 1 per τ minutes. We are to find the pdf of the occurrence at time t. Hence, we let $N(t)$ be the number of events registered from start to time t. We also let T_1 be the time of occurrence of the first event. Next, using properties of the conditional probability and the Poisson distribution, the cdf is:

$$\begin{aligned} F(t) &= P\{T_1 \le t | N(\tau) = 1, 0 \le t \le \tau\} \\ &= P\{N(t) = 1 | N(\tau) = 1\} \\ &= \frac{P\{N(t) = 1 \text{ and } N(\tau) = 1\}}{P\{N(\tau) = 1\}} \\ &= \frac{P\{N(\tau) = 1 | N(0) = 1\} P\{N(t) = 1\}}{P\{N(\tau) = 1\}} \\ &= \frac{P\{N(\tau - t) = 1 | N(0) = 1\} P\{N(t) = 1\}}{P\{N(\tau) = 1\}} \\ &= \frac{P\{N(\tau - t) = 0 | N(0) = 0\} P\{N(t) = 1\}}{P\{N(\tau) = 1\}} \\ &= \frac{e^{-(\tau - t)\lambda} \lambda t e^{-\lambda t}}{\lambda \tau e^{-\lambda \tau}} \\ &= \frac{t}{\tau}, \quad 0 < t < \tau. \end{aligned}$$

Therefore, T_1 is a uniform random variable in $(0, \tau)$. Hence, the first count happened before 8:20 A.M. with probability (20/30) = 66.67%.

Definition 1.6.2
A continuous random variable X with pdf

$$f_X(t) = \begin{cases} \mu e^{-\mu t}, & t \geq 0, \\ 0, & \text{elsewhere}, \end{cases} \tag{1.6.3}$$

and cdf

$$F_X(t) = \begin{cases} 1 - e^{-\mu t}, & t \geq 0, \\ 0, & \text{elsewhere}, \end{cases} \tag{1.6.4}$$

is called a *negative exponential* (or *exponential*) *random variable*. Relation (1.6.3) and relation (1.6.4) are called *exponential density function* and *exponential distribution function*, respectively. μ is called the parameter for the pdf and cdf. See Figure 1.6.1 and Figure 1.6.2.

We note that the expected value of exponential distribution is the reciprocal of its parameter. This is because from (1.6.3) we have:

$$E(X) = \int_0^\infty \mu t e^{-\mu t} dt = \frac{1}{\mu}.$$

See also Figure 1.6.1 and Figure 1.6.2.

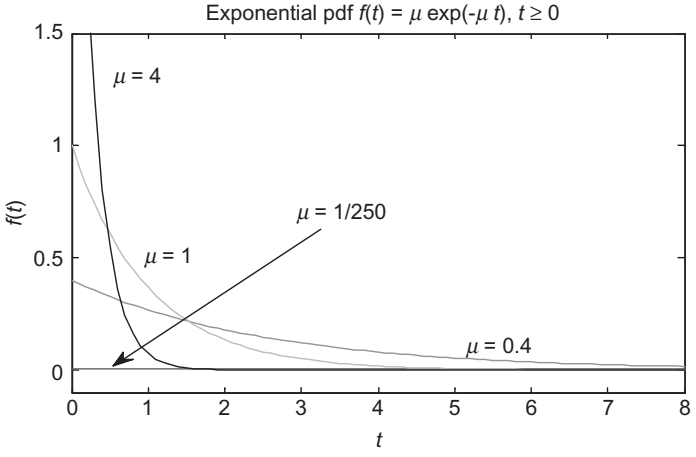


Figure 1.6.1. Exponential pdf with different values for its parameters.

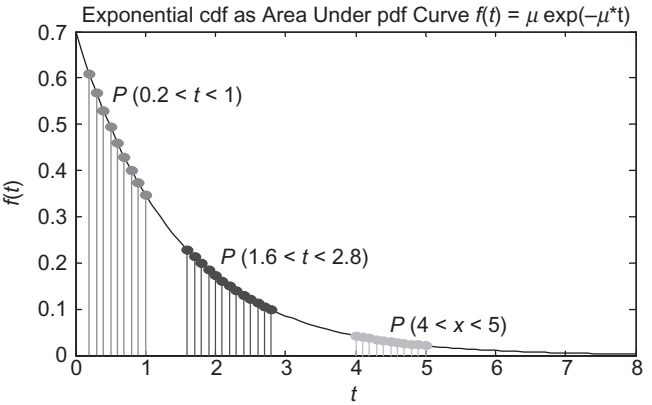


Figure 1.6.2. Exponential cdf as area under pdf with different intervals of t.

Example 1.6.3
Suppose it is known that the lifetime of a light bulb has an exponential distribution with parameter 1/250. We want to find the probability that the bulb works (a) for more than 300 hours and (b) for more than 300 if it has already worked 200 hours.

Answer
Let X be the random variable representing the lifetime of a bulb. From (1.6.3) and (1.6.4), we then have:

$$f_X(t)=\begin{cases}\dfrac{1}{250}e^{-\frac{1}{250}t}, & t\geq 0,\\ 0, & \text{elsewhere,}\end{cases} \quad \text{and} \quad F_X(t)=\begin{cases}1-e^{-\frac{1}{250}t}, & x\geq 0,\\ 0, & t<0.\end{cases}$$

Therefore,

(a) $P(X>300)=e^{-1.2}=0.3012$, and

(b) $$P(X>300|X>200)=\frac{P(X>300 \text{ and } X>200)}{P(X>200)}$$
$$=\frac{P(X>300)}{P(X>200)}=\frac{e^{-1.2}}{e^{-0.8}}=e^{-.4}=0.6703.$$

See Figure 1.6.3.

Definition 1.6.3
A continuous random variable, X, with probability density function $f(x)$ defined as:

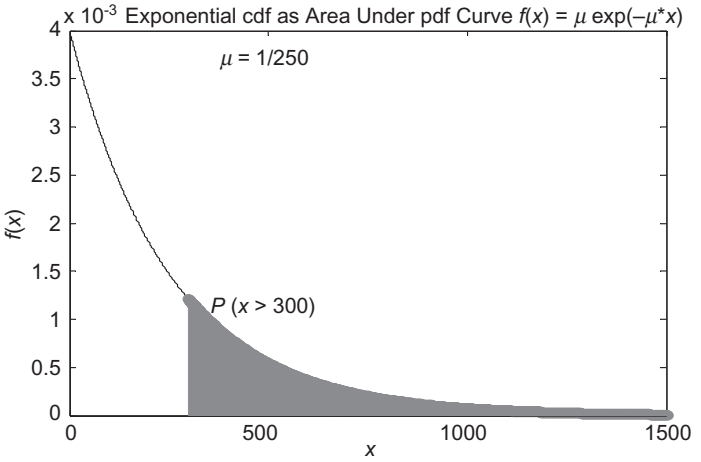


Figure 1.6.3. Exponential probability for x > 300.

$$f_X(x;\mu,\alpha)=\begin{cases}\dfrac{\mu^{\alpha}x^{\alpha-1}}{\Gamma(\alpha)}e^{-\mu x}, & x>0,\ \alpha \text{ is real and} >0,\\ 0, & x\le 0,\end{cases} \tag{1.6.5}$$

where $\Gamma(\alpha)$ is defined by:

$$\Gamma(\alpha)=\int_0^{\infty}x^{\alpha-1}e^{-x}dx, \tag{1.6.6}$$

where μ is a positive number, is called a *gamma random variable with parameters μ and α*. The corresponding distribution called *gamma distribution function* will, therefore, be:

$$F_X(x;\mu,\alpha)=\begin{cases}\dfrac{1}{\Gamma(\alpha)}\displaystyle\int_0^{x}\mu^{\alpha}u^{\alpha-1}e^{-\mu u}du, & \text{if } x\ge 0,\\ 0, & \text{if } x<0.\end{cases} \tag{1.6.7}$$

Definition 1.6.4
In (1.6.4), if μ is a nonnegative integer, say k, then the distribution obtained is called the *Erlang distribution of order k*, denoted by $E_k(\mu;x)$, that is,

$$E_k(\mu;x)=\begin{cases}\dfrac{1}{\Gamma(k)}\displaystyle\int_0^{\mu x}u^{k-1}e^{-u}du, & \text{if } x\ge 0,\\ 0, & \text{if } x<0.\end{cases} \tag{1.6.8}$$

The pdf in this case, denoted by $e_k(\mu;x)$, will be:

$$e_k(\mu;x)=\mu^k x^{k-1}e^{-\mu x}, \quad x\ge 0. \tag{1.6.9}$$

Notes:

(a) We leave it as an exercise to show that $f_X(x; \mu, \alpha)$ given by (1.6.4), indeed, defines a probability density function.

(b) The parameter μ in (1.6.8) is called the *scale parameter*, since values other than 1 either stretch or compress the pdf in the x-direction.

(c) In (1.6.4), if $\alpha = 1$, we will obtain the exponential density function with parameter μ defined by (1.6.4).

(d) $\Gamma(\alpha)$ is a positive function of α.

(e) If α is a natural number, say $\alpha = n$, then we leave it as an exercise to show that:

$$\Gamma(n) = (n-1)!, \quad n = 1, 2, \ldots, \tag{1.6.10}$$

where $n!$ is defined by:

$$n! = n(n-1)(n-2)\ldots(2)(1). \tag{1.6.11}$$

(f) We leave it as an exercise to show that from (1.6.5) and (1.6.10), one obtains:

$$0! = 1. \tag{1.6.12}$$

(g) Because of (1.6.11), the gamma function defined in (1.6.5) is called the *generalized factorial*.

(h) We leave it as an exercise to show that using double integration and polar coordinates, we obtain:

$$\int_0^\infty e^{-x^2} dx = \sqrt{2\pi}. \tag{1.6.13}$$

(i) We leave it as an exercise to show that using (1.6.13), one can obtain the following:

$$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}. \tag{1.6.14}$$

(j) Using (1.6.4), we denote the integral $\int_0^x u^{\alpha-1}e^{-u}du$, $x > 0$, by $\Gamma(\alpha, x)$, that is,

$$\Gamma(\alpha, x) = \int_0^x u^{\alpha-1}e^{-u}du, \quad x > 0. \tag{1.6.15}$$

The integral in (1.6.15), which is not an elementary integral, is called the *incomplete gamma function*.

(k) The parameter α in (1.6.4) could be a complex number whose real part must be positive for the integral to converge. There are tables available

for values of $\Gamma(\alpha, x)$, defined by (1.6.15). As x approaches infinity, the integral in (1.6.5) becomes the gamma function defined by (1.6.4).

(l) If $k = 1$, then (1.6.6) reduces to the exponential distribution function (1.6.2). In other words, exponential distribution function is a special case of gamma and Erlang distributions.

Definition 1.6.5
In (1.6.4), if $\alpha = r/2$, where r is a positive integer, and if $\mu = 1/2$, then the random variable X is called the *chi-square* random variable with r degrees of freedom, denoted by $X^2(r)$. The pdf and cdf in this case with shape parameter r are:

$$f(x) = \frac{1}{\Gamma\left(\frac{r}{2}\right)} 2^{\frac{r}{2}} x^{\frac{r}{2}-1} e^{-\frac{x}{2}}, \quad 0 \le x < \infty \tag{1.6.16}$$

and

$$F(x) = \int_0^x \frac{1}{\Gamma\left(\frac{r}{2}\right) 2^{\frac{r}{2}}} v^{\frac{r}{2}-1} e^{\frac{-v}{2}} dv, \tag{1.6.17}$$

respectively, where $\Gamma(r/2)$ is the gamma function with parameter $r/2$ defined in (1.6.4).

Notes:

(a) Due to the importance of X^2 distribution, tables are available for values of the distribution function (1.6.17) for selected values of r and x.

(b) We leave as an exercise to show the following properties of X^2 random variable: mean = r, and variance = $2r$.

The next distribution is widely used in many areas of research where statistical analysis is being used, particularly in statistics.

Definition 1.6.6
A continuous random variable X with pdf denoted by $f(x; \mu, \sigma^2)$ with two real parameters μ, $-\infty < \mu < \infty$, and σ^2, $\sigma > 0$, where:

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty, \tag{1.6.18}$$

is called a *Gaussian* or *normal* random variable.

The notation ~ is used for distribution. The letter N and character Φ are used for normal cumulative distribution. Hence,

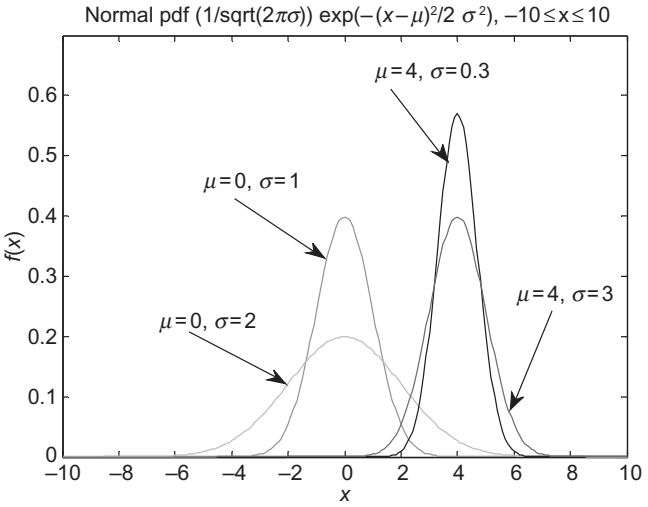


Figure 1.6.4. Normal pdf with different values for its parameters.

$$X \sim N(\mu, \sigma^2) \quad \text{or} \quad X \sim \Phi(\mu, \sigma^2) \tag{1.6.19}$$

is to show that the random variable X has a normal cumulative probability distribution with parameters μ and σ^2. The normal pdf has a *bell-shaped* curve and is *symmetric* about the line $f(x) = \mu$ (see Figure 1.6.4). The normal pdf is also asymptotic, that is, the tails of the curve from both sides get very close to the horizontal axis, but never touch it. Later, we will see that μ is the *mean* and σ^2 is the *variance* of the normal distribution function. The smaller the value of variance is, the narrower the shape of the "bell" would be. That is, the data points are clustered around the mean (i.e., the peak).

We leave it as an exercise to show that $f(x; \mu, \sigma^2)$, defined in (1.6.18), is indeed a pdf.

Definition 1.6.7
A continuous random variable Z with $\mu = 0$ and $\sigma^2 = 1$ is called a *standard normal* random variable. From (1.6.19), the cdf of Z, $P(Z \le z)$, is denoted by $\Phi(z)$. The notation $N(0, 1)$ or $\Phi(0, 1)$ is used to show that a random variable has a standard normal distribution function, which means it has the parameters 0 and 1. The pdf of Z, denoted by $\phi(z)$, therefore, will be:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad -\infty < z < \infty. \tag{1.6.20}$$

Note that any normally distributed random variable X with parameters μ and $\sigma > 0$ can be standardized using a substitution:

$$Z = \frac{X - \mu}{\sigma}. \tag{1.6.21}$$

The cdf of Z is:

$$\Phi(z) = P(Z \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{u^2}{2}} du, \tag{1.6.22}$$

which is the integral of the pdf defined in (1.6.20). We leave it as an exercise to show that $\phi(x)$ defined in (1.6.20) is a pdf.

Note that the cumulative distribution function of a normal random variable with parameters μ and σ^2, $F(x; \mu, \sigma^2)$, whose pdf was given in (1.6.18), may be obtained by (1.6.22) as:

$$F(x; \mu, \sigma^2) = \Phi\left(\frac{x-\mu}{\sigma}\right) = P(Z \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{u^2}{2}} du. \tag{1.6.23}$$

A practical way of finding normal probabilities is to find the value of z from (1.6.22), and then use available tables for values of the area under the curve of $\phi(x)$.

Figure 1.6.4 shows a graph of normal pdf for different values of its parameters. It shows the standard normal as well as the shifted mean and variety of values for standard deviation. Figure 1.6.5 shows a graph of normal cdf as area under pdf curve with different intervals.

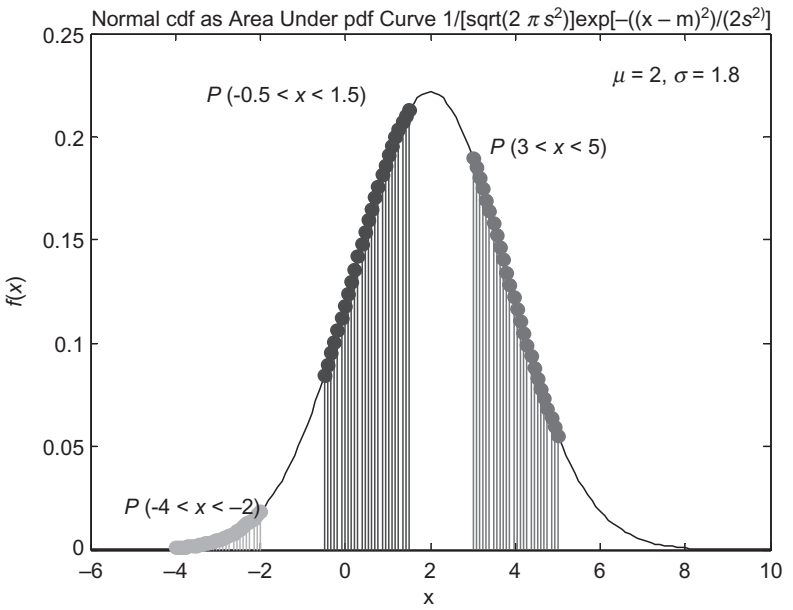


Figure 1.6.5. Normal cdf as area under pdf curve with different intervals.

Definition 1.6.8
A continuous random variable X, on the positive real line, is called a *Galton* or *lognormal* random variable if $\ln(X)$ is normally distributed.

Notes:

(a) If X is a normally distributed random variable, then $Y = e^X$ has a lognormal distribution.
(b) From Definition 1.6.8, X can be written as $X = e^{\mu+\sigma Z}$, where Z is a standard normal variable, and μ (location parameter) and σ (scale parameter) are the mean and standard deviation, respectively, of the natural logarithm of X.

The probability density function and cumulative distribution function of a lognormal random variable X, denoted by $f_X(x; \mu, \sigma)$ and $F_X(x; \mu, \sigma)$, respectively, are:

$$f_X(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, \quad x > 0, \tag{1.6.24}$$

and

$$F_X(x;\mu,\sigma) = \Phi\left(\frac{\ln x - \mu}{\sigma}\right), \tag{1.6.25}$$

where $\Phi(\cdot)$ is defined by (1.6.22).

We leave it as an exercise to show that mean and variance of a lognormal random variable X are, respectively, as

$$E(X) = e^{\mu+\sigma^2/2}, \quad \text{and} \quad Var(X) = \left(e^{\sigma^2} - 1\right)e^{2\mu+\sigma^2}. \tag{1.6.26}$$

Definition 1.6.9
A continuous random variable, X, with cumulative probability distribution function $F(x)$ defined as

$$F(x) = \frac{e^x}{1+e^x}, \tag{1.6.27}$$

is called the *standard logistic probability distribution*. By adding location and scale parameters, we will have *general logistic cumulative probability distribution* and *probability density function*, respectively, defined as:

$$F(x) = \frac{e^{\frac{x-a}{b}}}{1+e^{\frac{x-a}{b}}}, \quad x \geq 0, \tag{1.6.28}$$

and

$$f(x) = \frac{e^{\frac{x-a}{b}}}{b\left[1+e^{\frac{x-a}{b}}\right]^2}, \quad x \geq 0, \tag{1.6.29}$$

where a and b are location (mean) and scale (a parameter proportion to the standard deviation), respectively.

In case there is no confusion, we refer to the general logistic as logistic distribution.

Notes:

(1) If we set $a = 0$ and $b = 1$ in (1.6.28), we obtain (1.6.29).
(2) For standard logistic, a and b are location (mean) and scale (a parameter proportion to the standard deviation), respectively.
(3) $f(x)$, defined in (1.6.29), is symmetric about $x = a$.
(4) $f(x)$ is increasing on $(-\infty, a)$ and decreasing on (a, ∞). This implies that the mode and median occur at $x = a$.

The graph of logistic distribution is very much like a normal distribution. Hence, we may approximate a logistic distribution by a normal distribution or vice versa. If we consider the standard logistic, one possibility is to set the mean of the normal distribution to zero so it is also symmetric about zero, as logistic is, then pick the variance of the normal distribution, $\sigma^2 = \pi^2/3$, so that both distributions have the same variance.

We leave it as an exercise to show that the mean and variance of standard logistic random variable X, respectively, are:

$$E(X) = a \quad \text{and} \quad Var(X) = \frac{1}{3}\pi^2 b^2. \tag{1.6.30}$$

Example 1.6.4
For values of x from –3 to 7 with increments of 0.1, $a = 2$, and $b = 2$, 3, and 4, the pdf of logistic is shown in Figure 1.6.6.

Example 1.6.5
Let there be 10 possible values available for x, in a logistic distribution, that is, the domain of the random variable X (the sample space) is $\Omega = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. If we choose $a = 5.5$ and $b = 1.5$ in (1.6.28), then we will have results as presented in Table 1.6.1.

With the chosen parameter values, we will have $F(5.50) = 0.50$ and $F(8.80) = 0.90$. In other words, we set the median of the logistic distribution at

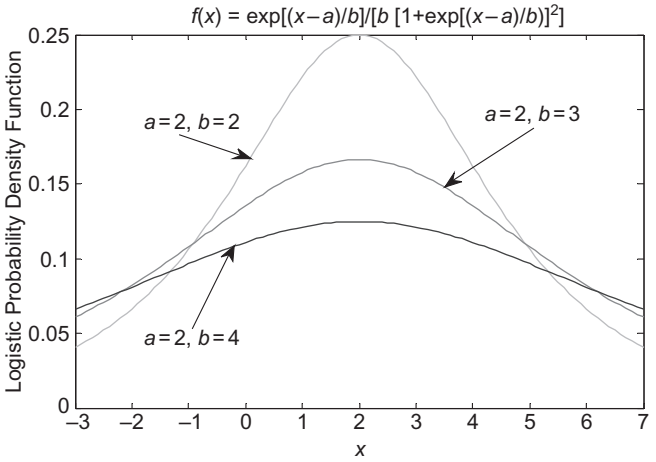


Figure 1.6.6. Logistic pdf with fixed location parameter and three values of scale parameter.

TABLE 1.6.1. Logistic cdf

Given x	1	2	3	4	5	6	7	8	9	10
Logistic cdf	0.047	0.088	0.159	0.269	0.417	0.583	0.731	0.841	0.912	0.953

5.5 and the 90th percentile at 8.80. Note that this choice of parameters centers the values of X.

We can find the inverse of the logistic distribution as follows. Let us write (1.6.28) as:

$$F(x) = \frac{1}{1+e^{-\frac{x-a}{b}}}. \tag{1.6.31}$$

By dropping x and cross-multiplying (1.6.28), we have $F + Fe^{-x-a/b} = 1$, or:

$$e^{-\frac{x-a}{b}} = \frac{1-F}{F},$$

or

$$-\frac{x-a}{b} = \ln\frac{1-F}{F},$$

or

$$\frac{x-a}{b} = \ln\frac{F}{1-F}.$$

TABLE 1.6.2. Inverse Logistic Cumulative Distribution Probability

Given values of logistic distribution probabilities, F	0.10	0.20	0.25	0.33	0.50	0.90
x values found	2.2042	3.4206	3.8521	4.4377	5.5000	8.7958

Hence, the inverse of logistic cdf (1.6.30) is:

$$x = a + b\ln\left(\frac{F}{1-F}\right). \tag{1.6.32}$$

If we choose values for F as 0.1, 0.2, 0.25, 0.33, 0.5, and 0.9, from (1.6.30) we will have results as presented in Table 1.6.2.

Definition 1.6.10
A continuous random variable X with cdf denoted by $F(x; \alpha, \beta, \gamma)$ as:

$$F(x;\alpha,\beta,\gamma) = 1 - e^{-\left(\frac{x-\gamma}{\alpha}\right)^{\beta}}, \quad \alpha, \beta > 0, x \geq \gamma \geq 0, \tag{1.6.33}$$

is called the *Weibull cumulative probability distribution function*, where α, β and γ are the *scale* (stretches/shrinks the graph), *shape* (such as skewness and kurtosis), and *location* (shifts the graph) parameters, respectively.

Without loss of generality, we let $\gamma = 0$. Thus, (1.6.33) will be reduced to the two-parameter Weibull cumulative distribution function as follows:

$$F(x;\alpha,\beta) = \begin{cases} 1 - e^{-\left(\frac{x}{\alpha}\right)^{\beta}}, & \alpha, \beta, > 0, x \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{1.6.34}$$

From (1.6.34), the two-parameter Weibull pdf, denoted by $f(x; \alpha, \beta)$, is:

$$f(x;\alpha,\beta) = \begin{cases} \frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1} e^{-\left(\frac{x}{\alpha}\right)^{\beta}}, & \alpha, \beta > 0, x \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{1.6.35}$$

In this case, we write $X \sim Weibull(\alpha, \beta)$. When $\alpha = 1$, (1.6.34) and (1.6.35) are referred to as *single-parameter* Weibull cdf and Weibull pdf, respectively. This case is referred to as *standard Weibull density* and *distribution*, respectively.

Note that when $\beta = 1$, we will have the exponential distribution, (1.6.4). In other words, exponential distribution function is a special case of Weibull distribution.

From (1.6.34), we find *the inverse*, that is, x, as follows:

$$e^{-\left(\frac{x}{\alpha}\right)^{\beta}} = 1-F, \text{or} -\left(\frac{x}{\alpha}\right)^{\beta} = \ln(1-F), \text{or}$$

$$\left(\frac{x}{\alpha}\right)^{\beta} = \ln\left(\frac{1}{1-F}\right), \text{or} \frac{x}{\alpha} = \left[\ln\left(\frac{1}{1-F}\right)\right]^{\frac{1}{\beta}}, \text{or} \tag{1.6.36}$$

$$x = \alpha\left[\ln\left(\frac{1}{1-F}\right)\right]^{\frac{1}{\beta}}.$$

Example 1.6.6

For different pair values of (α, β), we have the results presented in Table 1.6.3.

We leave it as an exercise to show that the mean and variance of the Weibull random variable, respectively, are

$$E(X) = \alpha\Gamma\left(1+\frac{1}{\beta}\right), \quad \text{and} \quad Var(X) = \alpha^2\left[\Gamma\left(1+\frac{2}{\beta}\right) - \Gamma\left(1+\frac{1}{\beta}\right)^2\right]. \tag{1.6.37}$$

See also Figure 1.6.7 and Figure 1.6.8.

TABLE 1.6.3. Inverse Weibull Probability Distribution

Given Weibull cdf, $F(x, \alpha, \beta)$	0.10	0.20	0.25	0.33	0.50	0.90
Weibull parameters (α, β) (scale, shape)	(4.0,3.0)	(4.507, 3.25)	(5.014, 3.5)	(5.521, 3.75)	(8.028, 4.0)	(6.535, 4.25)
Inverse Weibull cdf, x	1.8892	2.8409	3.5123	4.3255	5.5002	7.9519

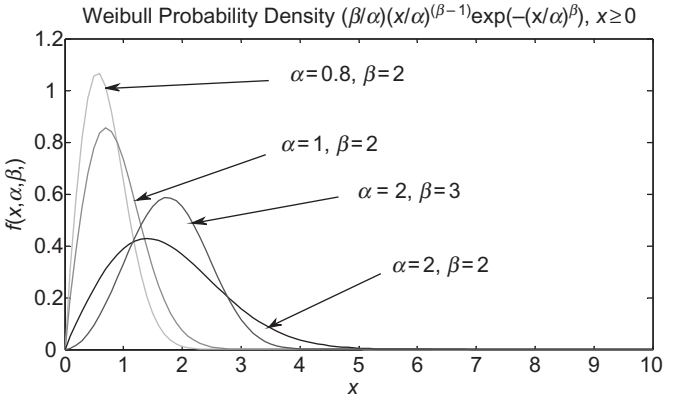


Figure 1.6.7. Two-parameter Weibull pdf with different scale and shape parameters.

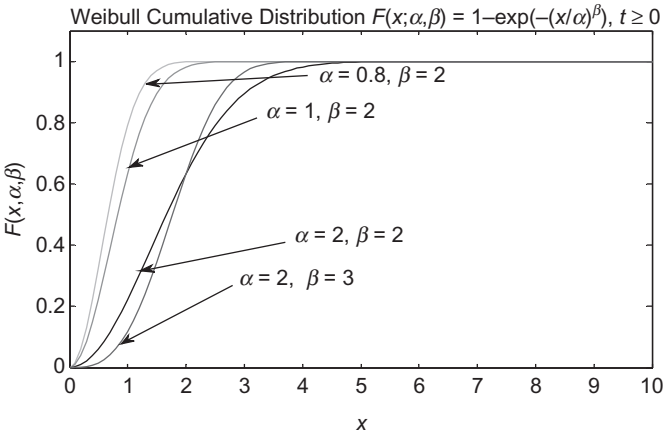


Figure 1.6.8. Two-parameter Weibull cdf with different scale and shape parameters.

Definition 1.6.11
A random variable X has an *extreme value distribution* if the distribution is of one the following three forms:

1. Type 1 (or *double exponential* or *Gumbel-type* distribution), with ν as the location parameter and θ as the scale parameter:

$$F_X(x;\nu,\theta) \equiv P(X \leq x) = e^{-e^{-(x-\nu)/\theta}}. \tag{1.6.38}$$

2. Type 2 (or *Fréchet-type distribution*) with three parameters:

$$F_X(x;\nu,\theta,\alpha) \equiv P(X \leq x) = e^{-\left(e^{-(x-\nu)/\theta}\right)^{\alpha}}, \quad x \geq \nu, \theta, \alpha > 0. \tag{1.6.39}$$

3. Type 3 (or *Weibull-type distribution*) with two parameters:

$$F_X(x;\nu,\theta,\alpha) \equiv P(X \leq x) = \begin{cases} e^{-\left(e^{-(x-\nu)/\theta}\right)^{\alpha}}, & x \leq \nu, \theta, \alpha > 0, \\ 1, & x > \nu. \end{cases} \tag{1.6.40}$$

Notes:

(a) If X is an extreme value random variable, so is $(-X)$.
(b) Type 2 and Type 3 can be obtained from each other by changing the sign of the random variable.
(c) Type 2 can be transformed to Type 1 by the following transformation: $Z = \ln(X - \nu)$.
(d) Type 3 can be transformed to Type 1 by the following transformation: $Z = -\ln(\nu - X)$.

(e) Of the three types, Type 1 is the most commonly used as *the extreme value distribution*. The pdf for Type 1 is:

$$f_X(x; \nu, \theta) = \frac{1}{\theta} e^{-e^{-(x-\nu)/\theta}} e^{-e^{-(x-\nu)/\theta}}. \tag{1.6.41}$$

(f) From (1.6.41), we have:

$$-\ln[-\ln P(X < x)] = \frac{x-\nu}{\theta}. \tag{1.6.42}$$

(g) All three types of extreme value distributions can be obtained from the following cdf:

$$F_X(x; \nu, \theta, \alpha) \equiv P(X \le x) = \left[1 + \alpha\left(\frac{x-\nu}{\theta}\right)\right]^{-1/\alpha}, \quad 1 + \alpha\left(\frac{x-\nu}{\theta}\right) > 0, -\infty < \alpha < \infty, \theta > 0. \tag{1.6.43}$$

Relation (1.6.43) is called a generalized extreme value or von Mises-type or von Mises–Jenkinson-type distribution.

1.7. RANDOM VECTOR

We have occasions where more than one random variable are considered at a time. Suppose that two distinct random experiments with sample spaces Ω_1 and Ω_2 can conceptually be combined into a single one with only one sample space, Ω. The new sample space, Ω, will be the Cartesian product of Ω_1 and Ω_2, that is, $\Omega_1 \times \Omega_2$, which is the set of all ordered pairs (ω_i, ω_j), where ω_i and ω_j are outcomes of the first and the second experiment, respectively, that is, $\omega_i \in \Omega_1$ and $\omega_j \in \Omega_2$. This idea may be extended to a finite number of sample spaces. In such cases, we are talking of a random vector. In other words, a *random vector* is defined as an n-tuple of random variables. More precisely, we have the following definition:

Definition 1.7.1
A *discrete random vector* $X(\omega) = [X_1(\omega), \ldots, X_r(\omega)]$, where $X_1, \ldots, X_r$ are r discrete random variables and $\omega \in \Omega$, is a function from the sample space Ω into the r-tuple (r-dimensional) real line, $\mathbb{R}^r$, such that for any r real numbers $x_1, x_2, \ldots, x_r$, the set $\{\omega \in \Omega: X_i(\omega) = x_i, i = 1, 2, \ldots, r\}$ is an event.

Example 1.7.1
Suppose a factory wants to conduct a quality control of its product by considering numerical factors $x_1, x_2, \ldots, x_n$ such as weight, height, volume, and color. Such a test can be done by the numerical value of the probability

$P(X_1 \le x_1, \ldots, X_n \le x_n)$, where the random vector $(X_1, X_2, \ldots, X_n)$ will describe the joint factors concerned by the factory.

When we have a random vector, the probability distribution of such a vector must give the probability for all the components at the same time. This is the joint probability distribution for the n component random variables which make up the random vector, such as the ones in previous example. Therefore, the necessary probabilistic information can be transferred to the range value from the original probability space.

Definition 1.7.2
Let X be a discrete bivariate random vector, that is, $X = (x_1, x_2)$. Suppose that x_1 and x_2 are two real numbers. Let $p_{x_1x_2} = P(X_1 = x_1, X_2 = x_2)$. $p_{x_1x_2}$ is then called the *joint probability mass function of* x_1 *and* x_2.

We have already discussed independence of two events. We now want to define it for random variables as follows:

Definition 1.7.3
Suppose (S, Ω, P) is an elementary probability space and the random vector $X = (X_1, X_2)$ is defined on the sample space S. The random variables X_1 and X_2 are *independent* if the partitions generated by X_1 and X_2 are independent.

The following theorem gives a better understanding of the independence concept.

Theorem 1.7.1
Two discrete random variables X and Y are independent if and only if the joint pmf of X and Y is the product of the marginal pmf of each X and Y. In other words, X and Y are independent if and only if:

$$P_{X,Y} = P_X P_Y, \tag{1.7.1}$$

where P_X and P_Y are the pmf of X and Y, respectively.

Proof:
See Haghighi et al. (2011a, p. 179).

The bivariate X can be extended to a random vector with discrete components $X_1, X_2, \ldots, X_r$. The joint probability mass function of $x_1, x_2, \ldots, x_r$ then, is similarly defined as:

$$P_{x_1,x_2,\ldots,x_r} = P(X_1 = x_1, X_2 = x_2, \ldots, X_r = x_r).$$

Notes:

(1) From the axioms of probability, a joint probability mass function has the following properties:

(a) $p_{x_1x_2,\cdots,x_r} \geq 0$, and

(b) $\sum_{x_1}\sum_{x_2}\cdots\sum_{x_r} p_{x_1x_2\cdots x_r} = 1$.

(2) The discrete bivariate, p_{x_i,y_j} means the probability that $X = x_i$, $Y = y_j$, and P is defined on the set of ordered pairs $\{(x_i, y_j), j \leq i \leq m, i \leq j \leq n\}$ by $p_{x_i,y_j} = P([X = x_i] \text{ and } [Y = y_j])$. We may obtain each individual distribution functions from the joint distribution. If $A_i = [X = x_i]$ and $B_j = [Y = y_j]$ are events, then $A_iB_j, i = 1, 2, \ldots, m$, are mutually exclusive events and $A_i = \cup A_iB_j$. Thus, we have:

$$p_{x_i} = P(A_i)\sum_{j=1}^{n} P(A_iB_j) = \sum_{j=1}^{n} p_{x_i,y_j}, \quad i = 1, 2, \ldots, m. \tag{1.7.2}$$

Similarly, we will have:

$$p_{y_j} = P(B_j)\sum_{i=1}^{m} P(A_iB_j) = \sum_{i=1}^{m} p_{x_i,y_j}, \quad i = 1, 2, \ldots, n. \tag{1.7.3}$$

Example 1.7.2

Let X and Y be two random variables with the following joint distribution:

$p_{X,Y}$		Y			
		–1	0	1	2
X	–1	0	$\frac{1}{36}$	$\frac{1}{6}$	$\frac{1}{12}$
	0	$\frac{1}{18}$	0	$\frac{1}{18}$	0
	1	0	$\frac{1}{36}$	$\frac{1}{6}$	$\frac{1}{12}$
	2	$\frac{1}{12}$	0	$\frac{1}{12}$	$\frac{1}{6}$

Questions:

(a) Find $P\{X \geq 1 \text{ and } Y \leq 0\}$.

(b) Find $P\{Y \leq 0 \mid X = 2\}$.

(c) Are X and Y independent? Why?

(d) Find the distribution of $Z = X \cdot Y$.

Answers

(a) For $X \geq 1$ and $Y \leq 0$, we have the following pairs (1, 0), (1, –1), (2, 0), and (2, –1), with probabilities 1/36, 0, 0, and 1/12, yielding:

$$P\{X \geq 1 \text{ and } Y \leq 0\} = \frac{1}{36} + 0 + 0 + \frac{1}{12} = \frac{1}{9}.$$

(b) For given X given as 2 and Y to be less than or equal to 0, we have the following pairs (2, 0) and (2, –1), with probabilities 0 and 1/12, yielding:

$$P\{Y \leq 0 \,|\, X = 2\} = 0 + \frac{1}{12} = \frac{1}{12}.$$

(c) X and Y are dependent. Here is one reason why:

$$P\{X = -1 \text{ and } Y = -1\} = 0.$$

However,

$$P\{X = -1\} = 5/18 \quad \text{and} \quad P\{Y = -1\} = 5/36.$$

Hence,

$$P\{X = -1\} \cdot P\{Y = -1\} = \frac{5}{18} \cdot \frac{5}{36} \neq P\{X = -1, Y = -1\} = 0.$$

(d) Since the values for both X and Y from the table given are as –1, 0, 1, 2, we will have $Z = X \cdot Y = -2, -1, 0, 1, 2, 4$. Pairs comprising these values and corresponding probabilities are given in the following table:

XY	–2	–1	0	1	2	4
Possible pairs	(–1,2), (2,–1)	(–1,1), (1,–1)	(0,–1), (0,0), (0,1), (0,2), (–1,0), (1,0), (2,0)	(1,1), (–1,–1)	(1,2), (2,1)	(2,2)
$P(Z = X \cdot Y)$	$\frac{1}{12} + \frac{1}{12} = \frac{1}{6}$	$\frac{1}{6} + 0 = \frac{1}{6}$	$\frac{1}{18} + 0 + \frac{1}{18}$ $0 + \frac{1}{36} + \frac{1}{36} + 0 = \frac{1}{6}$	$\frac{1}{6} + 0 = \frac{1}{6}$	$\frac{1}{12} + \frac{1}{12} = \frac{1}{6}$	$\frac{1}{6}$

Definition 1.7.4
Each probability mass function p_X and p_Y, defined by (1.2.15) and (1.2.16), respectively, is called the *marginal probability mass function.* In other words, a marginal probability mass function, p_X or p_Y, can be found from the joint distribution function p_{XY} of X and Y by summing up the joint distribution over all values of the Y or X, respectively.

Now let $X_1, X_2, \ldots, X_r$ be r random variables representing the occurrence of r outcomes among $X_1, X_2, \ldots, X_n$ possible outcomes of a random experiment, which is being repeated independently n times. Suppose the corresponding probabilities of these outcomes are $p_1, p_2, \ldots, p_r$, respectively. The joint pmf of these random variables will then be:

$$P(X_1 = x_1, \ldots X_r = x_r) = \frac{n!}{n_1! \ldots n_r!} p_1^{n_1} \ldots p_r^{n_r}, \tag{1.7.4}$$

where n_j ranges over all possible integral values subject to (1.7.4). The relation (1.7.4) does, indeed, represent a probability mass function (why?) and it is called the *multinomial probability mass function* for the random vector (X_1, $X_2, \ldots, X_r$). We denote this distribution similar to the binomial distribution as: $m(n; r: n_1, \ldots, n_r)$, subject to $p_1 + p_2 + \ldots + p_r = 1$.

Marginal mass function for each one of the random variables $X_1, X_2, \ldots,$ X_r alone will be a binomial probability mass function and can be obtained from (1.7.4). For instance, for X_1 we have:

$$P(X_1 = x_1) = b(n; p, n) = \frac{n!}{n_1!(n - n_1)!} p^{n_1} (1-p)^{n-n_1}. \tag{1.7.5}$$

As the reader recalls, we defined the conditional probability of the event and the law of total probability. Both these concepts may be extended to random variables. We leave the proof of the following theorem as an exercise.

Theorem 1.7.2. ***The Law of Total Probability***
Let X be a random variable and let the event A be represented by a discrete random variable, then we have:

$$P(A) = \sum_x P(A \mid X = x) P(X = x). \tag{1.7.6}$$

An important notion that measures the dependency of random variables is the notion of covariance, which will be defined in the following sections.

Definition 1.7.5
Let $E(X) = \mu_X$ and $E(Y) = \mu_Y$. The *covariance* of two random variables X and Y, denoted by $Cov(X, Y)$, is then defined by:

$$Cov(X, Y) \equiv E[(X - \mu_X)(Y - \mu_Y)]. \tag{1.7.7}$$

The following properties of covariance can easily be proved and are left as exercises.

Properties of Covariance

1. $Cov(X, X) = Var(X)$. (1.7.8)
2. $Cov(X, Y) = E[XY] - \mu_X \mu_Y$. (1.7.9)
3. $Cov(X, Y) = Cov(Y, X)$. (1.7.10)
4. If c is a real number, then:

$$Cov(cX, Y) = cCov(Y, X). \tag{1.7.11}$$

5. For two random variables X and Y, we have:

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y). \tag{1.7.12}$$

As an important application of covariant, we define the following:

Definition 1.7.6
The *coefficient of correlation* of two random variables X and Y, denoted by $\rho(X, Y)$, is given by:

$$\rho(X, Y) \equiv \frac{Cov(X, Y)}{\sigma(X)\sigma(Y)}, \tag{1.7.13}$$

where $\sigma(X)$ and $\sigma(Y)$ are the standard deviations of X and Y, respectively, provided that the denominator is not zero.

It can be shown that:

$$-1 \le \rho(X, Y) \le 1. \tag{1.7.14}$$

We note that when ρ is negative, it means that the random variables are dependent oppositely, that is, if one increases, the other will decrease. When ρ is positive, it means that both random variables increase and decrease together. However, if $\rho = 0$, the random variables are called *uncorrelated*. Some other properties of the correlation coefficient are as follows:

$$\rho(X, Y) = \rho(Y, X), \tag{1.7.15}$$

$$\rho(X, X) = 1, \tag{1.7.16}$$

$$\rho(X, -X) = -1, \tag{1.7.17}$$

and

$$\rho(aX + b, cY + d) = \rho(X, Y), \tag{1.7.18}$$

where, a, b, c, and d are real numbers and $a, c \neq 0$.

Example 1.7.3
Suppose X and Y are two random variables representing the "on" or "off" situation of two automatic switches that work together with the following joint distribution: $P(X = 0, Y = 0) = 1/5$, $P(X = 0, Y = 0) = 1/4$, $P(X = 1, Y = 0) = 1/4$, and $P(X = 1, Y = 1) = 3/10$. We may tabulate these values as follows:

X \ Y	0	1	Sum
0	1/5	1/4	9/20
1	1/4	3/10	11/20
Sum	9/20	11/20	

The marginal distributions of X and Y, denoted by p_X and p_Y, respectively, are as follows:

X	0	1
p_X	9/20	11/20

Y	0	1
p_Y	9/20	11/20

Hence,

$$E(XY) = (0)(0)(1/5) + (0)(1)(1/4) + (1)(0)(1/4) + (1)(1)(3/10) = 3/10 = 0.3000,$$
$$E(X) = \mu_X = (0)(9/20) + (1)(11/20) = 11/20 = 0.5500,$$
$$E(Y) = \mu_Y = (0)(9/20) + (1)(11/20) = 11/20 = 0.5500,$$
$$E(X^2) = (0)(9/20) + (1)(11/20) = 11/20 = 0.5500,$$
$$E(Y^2) = (0)(9/20) + (1)(11/20) = 11/20 = 0.5500,$$
$$Var(X) = 0.5500 - 0.3025 = 0.2475,$$
$$Var(Y) = 0.5500 - 0.3025 = 0.2475,$$
$$\sigma(X) = 0.4975,$$
$$\sigma(Y) = 0.4975,$$
$$Cov(XY) = E(XY) - \mu_X\mu_Y = 0.3000 - (0.5500)(0.5500) = -0.0025, \text{ and}$$
$$r(X, Y) = (-0.0025)/(0.4975)(0.4975) = -0.0101.$$

The value of the correlation coefficient, in this case, is very close to 0. Hence, X and Y in this case are uncorrelated.

1.8. CONTINUOUS RANDOM VECTOR

Definition 1.8.1
The *joint bivariate pdf* of two continuous random variables X and Y with pdf $f_X(x)$ and $f_Y(y)$, respectively, is an integrable function, say $f_{X,Y}(x, y)$ or just $f(x, y)$, with the following properties:

(a) $P(X = x \text{ and } Y = y) \approx f_{X,Y}(x, y)dx\,dy.$ (1.8.1)
(b) $f_{X,Y}(x, y) \geq 0.$
(c) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x, y)dx\,dy = 1.$
(d) $P\{(X, Y) \in A\} = \iint_A f_{X,Y}(x, y)dx\,dy,$ (1.8.2)
where $P\{(X, Y) \in A\}$ is an event defined in the x–y plane.

We note that property (d) implies that properties of discrete joint pmf can be extended to a continuous case using the approximation (1.8.1).

Definition 1.8.2
The *marginal pdf* of X and Y can be obtained from (1.8.1), respectively, as:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy, \tag{1.8.3}$$

and

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dx. \tag{1.8.4}$$

Definition 1.8.3
The *conditional probability density function of X given Y* is given by:

$$f_{X|Y}(x|y) = \frac{f_{X|Y}(x|y)}{f_Y(y)}. \tag{1.8.5}$$

As in the discrete case, the joint pdf can be extended for finitely many random variables.

Definition 1.8.4
Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a finite or denumerable random vector with joint pdf or pmf of $f_X(x_1, x_2, \ldots, x_n)$. We denote the marginal pdf or pmf of X_i,

$i = 1, 2, \ldots, n$, by $f_{X_i}(x_i)$. $X_1, X_2, \ldots, X_n$ are then *mutually independent random variables* if for every $(x_1, x_2, \ldots, x_n)$ we have:

$$f_{\mathbf{X}}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1) \times f_{X_2}(x_2) \times \cdots \times f_{X_n}(x_n). \tag{1.8.6}$$

If $f_{X_i}(x_i)$ is parametric pdf or pmf with, say one parameter, θ, say, denoted by $f_{X_i}(x_i; \theta)$, then the joint parametric pdf or pmf is:

$$f_{\mathbf{X}}(x_1, x_2, \ldots, x_n; \theta) = f_{X_1}(x_1; \theta) \times f_{X_2}(x_2; \theta) \times \cdots \times f_{X_n}(x_n; \theta). \tag{1.8.7}$$

Note that pairwise independence does not apply mutual independence.

1.9. FUNCTIONS OF A RANDOM VARIABLE

We now consider a random variable as a general term to include both discrete and continuous.

Definition 1.9.1
Suppose that $\phi(\cdot)$ is a function that associates real numbers onto real numbers. Next, the composite function $\phi[X(\cdot)]$ is defined and with each outcome ω, $\omega \in \Omega$, it associates the real number $\phi[X(\omega)]$. $Y(\omega) \equiv \phi[X(\omega)]$ is called the *function of the random variable X*. If $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is a random vector of n random variable that associates the sample space Ω to the space $\mathbb{R}^n$ of real n-tuples, then the function $\phi(\cdot, \ldots, \cdot)$ on n real variables associates with each point in $\mathbb{R}^n$ a real number. Hence, we define $\phi[\mathbf{X}] = \phi[X_1(\cdot), X_2(\cdot), \ldots, X_n(\cdot)]$ for each $\omega \in \Omega$ as the real number $\phi[X_1(\omega), X_2(\omega), \ldots, X_n(\omega)]$. $Y(\omega) \equiv \phi[X_1(\omega), X_2(\omega), \ldots, X_n(\omega)]$ is called the *function of n*, $n \geq 1$, *random variables*.

It can be easily proved that:

Theorem 1.9.1
If Y is a function of a discrete random variable X, say $Y = \phi[X(\omega)]$ and $p(\omega) = P(X = \omega)$, then:

$$E(Y) = E[\phi(X)] = \sum_{\omega \in \Omega} \phi(\omega)\, p(\omega). \tag{1.9.1}$$

Limiting distribution functions of certain functions of n random variables when n approaches infinity is an important class of problems in the theory of probability that serves mathematical statistics. In other words, let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a finite or denumerable random vector and $f_n(X_1, X_2, \ldots, X_n)$ is a function of $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, which itself is a random variable. The question is to find the limit of the cumulative distribution function of $\mathbf{X} = f_n(X_1, X_2, \ldots, X_n)$ as n approaches infinity, and if this is not possible, maybe find some properties of cdf, if it exists. This idea leads to considering a *stochastic process*

that is a sequence of random variables. We will discuss this process in detail in Chapter 5. But here we want to use it for a different purpose.

Definition 1.9.2
Let $(X_1, X_2, \ldots, X_n)$ be a stochastic process such that for an arbitrary small positive number ε,

$$\lim_{n\to\infty} P\{|X_n - X| > \varepsilon\} = 0. \tag{1.9.2}$$

We then say that $(X_1, X_2, \ldots, X_n)$ *converges stochastically* or *converges in probability* to the random variable X.

We note that if $P\{X = x_0\} = 1$, where x_0 is a constant, that is, X is a degenerate random variable, then $(X_1, X_2, \ldots, X_n)$ converges in probability to the constant x_0.

One of the most important examples of stochastic convergence is the weak law of large numbers. Before we state and prove this law, we present two important inequalities.

Theorem 1.9.2. ***Markov's Inequality***
Let X be a nonnegative discrete random variable with a finite mean, $E(X)$. Let a be a fixed positive number. Hence,

$$P\{X \geq a\} \leq \frac{E(X)}{a}. \tag{1.9.3}$$

Proof:
By definition,

$$\begin{aligned} E(X) &= \sum_x xf(x) \\ &= \sum_{0\leq x<a} xf(x) + \sum_{x=a}^{\infty} xf(x). \end{aligned} \tag{1.9.4}$$

The first term on the right-hand side of (1.9.4) is positive. Hence,

$$E(X) \geq \sum_{x=a}^{\infty} xf(x). \tag{1.9.5}$$

Since a is the minimum value of x, from (1.9.5) we have:

$$E(X) \geq \sum_{x=a}^{\infty} af(x) = a\sum_{x=a}^{\infty} f(x). \tag{1.9.6}$$

Hence,

$$E(X) \geq aP\{X \geq a\}. \tag{1.9.7}$$

It is clear from (1.9.7) that (1.9.3) follows.

Theorem 1.9.3. ***Chebyshev's Inequality***
Let X be a nonnegative random variable with finite mean μ and variance σ^2. Let k also be a fixed positive number. Hence:

$$P\{|X-\mu| > k\} \leq \frac{\sigma^2}{k^2}. \tag{1.9.8}$$

Proof:
Consider the random variable $(X - \mu)^2$, which is positive. From Marko's inequality (Theorem 1.9.2), we now have:

$$P\left\{(X-\mu)^2 \geq k^2\right\} \leq \frac{E\left[(X-\mu)^2\right])}{k^2}. \tag{1.9.9}$$

Now $(X - \mu)^2 \geq k^2$ implies that $|X - \mu| \geq k$. Hence,

$$P\left\{(X-\mu)^2 \geq k^2\right\} \leq \frac{E\left[(X-\mu)^2\right])}{k^2} = \frac{\sigma^2}{k^2}. \tag{1.9.10}$$

Example 1.9.1
Let us assume that the number of production of an item per week by a factory is a random variable with a mean of 600. We want to find (a) the probability that the number of production per week be at least 1200 and (b)the number of production to be between 500 and 600, if the variance of the number of production is 120.

Answer

(a) By Markov's inequality (Theorem 1.9.2) we have:

$$P\{X \geq 1200\} \leq \frac{E(X)}{1200} = \frac{600}{1200} = \frac{1}{2}.$$

(b) By Chebyshev's inequality (Theorem 1.9.3) we have:

$$P\{|X-600| \geq 120\} \leq \frac{\sigma^2}{(120)^2} = \frac{1}{120}.$$

Therefore,

$$P\{|X-600|<120\}\geq 1-\frac{1}{120}=\frac{119}{120}.$$

Theorem 1.9.4. ***The Weak Law of Large Numbers***
Let $\{X_1, X_2, \ldots, X_n, \ldots\}$ be a sequence of independent and identically distributed (iid) random variables with mean μ. Next, for an arbitrary small positive number ε, we have:

$$\lim_{n\to\infty} P\left\{\left|\frac{X_1+X_2+\cdots+X_n}{n}-\mu\right|>\varepsilon\right\}=0. \tag{1.9.11}$$

Proof:
Since $\{X_1, X_2, \ldots, X_n, \ldots\}$ is iid, due to additivity property of expected value, we have:

$$E\left(\frac{X_1+X_2+\cdots+X_n}{n}\right)=\frac{n\mu}{n}=\mu.$$

Similarly,

$$Var\left(\frac{X_1+X_2+\cdots+X_n}{n}\right)=\frac{n\sigma^2}{n^2}=\sigma^2/n.$$

Next, by Chebyshev inequality, we have:

$$P\left\{\left|\frac{X_1+X_2+\cdots+X_n}{n}-\mu\right|\geq\varepsilon\right\}\leq\frac{Var(X_1+X_2+\cdots+X_n)}{\varepsilon^2}=\frac{\sigma^2}{n\varepsilon^2}. \tag{1.9.12}$$

Thus, $\lim_{n\to\infty}(\sigma^2/n\varepsilon^2)=0$, regardless how small ε is.

We note that (1.9.11) states that under the conditions of the theorem, the stochastic process $\{\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_n, \ldots\}$ converges in probability to 0.

Theorem 1.9.4 may be stated differently and will be called differently, as follows:

Theorem 1.9.5. ***The Strong Law of Large Numbers***
Let $\{X_1, X_2, \ldots, X_n, \ldots\}$ be a sequence of iid random variables with mean μ. We then have:

$$P\left\{\lim_{n\to\infty}\frac{X_1+X_2+\cdots+X_n}{n}=\mu\right\}=1. \tag{1.9.13}$$

Proof:
See Neuts (1973, p. 304).

We note that (1.9.13) states that under the conditions of the theorem, the sample mean, repeated infinitely many times, converges almost surely to the

expected value. It is sometimes denoted as $\bar{X}_n \xrightarrow{a.s.} \mu$ when $n \to \infty$. Note also that in this convergence we are using "almost surely," while in the previous case, we used the expression "in probability." The Weal law essentially says that for a given small positive number, with a sufficient large number of sample, there will be a very high probability that the mean of observations will be within the given number of the expected value. This, of course, leaves open the possibility that $|\bar{X}_n - \mu| > \varepsilon$ happens many times, although at infrequent intervals. The strong law prevents this to happen, that is, $|\bar{X}_n - \mu| < \varepsilon$ will hold when n is large enough.

Example 1.9.2
Consider flipping a fair coin. Let A be an event. Let X_i, $i = 1, 2, \ldots,$ represent the ith trial such that:

$$X_i = \begin{cases} 1, & \text{if } A \text{ occurs on the } i^{th} \text{ trial,} \\ 0, & \text{otherwise.} \end{cases}$$

Based on the strong law of large numbers (Theorem 1.9.5), we then have:

$$\frac{X_1 + X_2 + \cdots + X_n}{n} \xrightarrow{n \to \infty} E(X) = P(A).$$

In other words, with probability 1, the limiting proportion of times the event A occurs is $P(A)$.

1.10. BASIC ELEMENTS OF STATISTICS

The purpose of most statistical studies is to obtain information about the population from a random sample by generalization. It is a common practice to identify a population and a sample with a distribution of their values: *parameters* and *statistic*, respectively. A *parameter* is a number that describes a character of the population. Numbers such as the mean and variance of a distribution are examples of a parameter. A characteristic of a sample is called a *statistic*. The smallest and largest values of a data set are referred to as the *minimum* and the *maximum*, respectively. The absolute value of difference between the maximum and minimum is called the *range*.

A set of data points may be summarized or grouped in a simple way or into a suitable number of classes (or categories). However, this grouping may cause the loss of some information in the data. This is because instead of knowledge of an individual data point, knowledge of its belonging to a group will be known.

A *simple frequency distribution* is a grouping of a data set according to the number of repetitions of a data point. The ratio of a frequency to the total number of observations is called *relative frequency*. This is the same

terminology we used for outcomes earlier in this chapter. A relative frequency multiplied by 100 will yield a *percent relative frequency*.

Example 1.10.1
Consider the set of 30 observations of daily emission (in tons) of sulfur from an industrial plant that is given by:

20.2, 12.7, 18.3, 18.3, 23.0, 23.0, 12.7, 11.0, 21.5, 10.2, 17.1, 07.3, 11.0, 18.3, 17.1, 12.7, 18.3, 20.2, 20.2, 20.2, 18.3, 12.7, 21.5, 17.1, 11.0, 12.7, 11.0, 23.0, 07.9, 07.9

We rewrite the data in ordered points as follows:

7.3	7.9	7.9	10.2	11.0
11.0	11.0	11.0	12.7	12.7
12.7	12.7	12.7	17.1	17.1
17.1	18.3	18.3	18.3	18.3
18.3	20.2	20.2	20.2	20.2
21.5	21.5	23.0	23.0	23.0

Data Point	Frequency	Relative Frequency	Percent Frequency
07.3	1	0.033	03.3
0.79	2	0.067	06.7
10.2	1	0.033	03.3
11.0	4	0.133	13.3
12.7	5	0.167	16.7
17.1	3	0.100	10.0
18.3	5	0.167	16.7
20.2	4	0.133	13.3
21.5	2	0.067	06.7
23.0	3	0.100	10.0
Total	30	1.000	100

In presenting a set of data points (in the form of a table, in percents, or a graph, for instance), one must be careful for any misinterpretation. In fact, two very important words should be in mind for such presentation. They are *abuse* and *misuse*. The word "abuse" means the use of a wrong concept intentionally. But "misuse" is often referred to as the unintentional use of a wrong concept. However, the "misuse" of statistics is not limited to unintentionally using a wrong concept; it could be a misrepresentation of a set of data points for a regular audience who does not have much knowledge of statistics. Even scientists have been known to fool themselves with statistics due to lack of knowledge of *probability theory* and the lack of mathematical statistical concepts. Thus, it is important to make it clear for what purpose the statistical representation is for and avoid all ambiguity as much as possible.

Example 1.10.2
Just recently, the following table was posted on *Facebook* regarding the 2012 Summer Olympics in London, England. The table of medals received by different countries was to show the ranking of countries based on the percent of the number of medals versus the number of athletes that participated in that country. The following table was sourced from the *Los Angeles Times* (August 13, 2012):

London Olympics 2012 Medal Winners			
% of medals per athletes			
Source: LA Times Aug. 13, 2012			
Country	% of athletes with medals	Total Athletes	Total Medals
CHINA	22.83465	381	87
IRAN	22.64151	53	12
USA	19.29499	539	104
Russia	18.7643	437	82
Japan	12.45902	305	38
UK	11.883	547	65
Germany	11.11111	396	44
S. Korea	10.85271	258	28
France	10.36585	328	34
Italy	9.824561	285	28

Here is a simple frequency distribution and percent of ratios of two variables: "Total Athletes" and "Total Medals." The second column shows the percent ratios. The ranking caught the first author's attention since Iran is ranked second, receiving **12** medals with only **53** athletes, while the United States is ranked third, receiving **104** medals with **539** athletes participating in the games. To the author, this was an example of "misuse" of statistics. It gives a good feeling to a person from Iran, seeing his/her country ranked higher than the United States, but what does this ranking tell us? To get a sense of how other statisticians felt, the following question was posted on an online statistics site, *ResearchGate*, for statisticians to comment on (note that due to the limitation of the number of characters you are allowed in a post, the description of the question is brief):

Is it misuse of statistics?

Countries 1 and 2 send team to Olympic. C1 team has 1 athlete and C2 has 500. C1 gets a medal. C2 gets 100 medals. C1's percent of medals versus its athletes is 100% but C2 is 20%. Billboard ranks C1 # 1 and C2 # 2 for percents. Is it misuse of statistics? If so, how should it be corrected?

There were 13 responses by worldwide statisticians, ranging from academic statisticians to medical doctors. For instance, here are four excerpts:

Someone from the IFF of India said:

> No, it is not misuse of statistics, this is the statistics.

Someone from the Danish methodological Institute said:

> I like to worry about significance levels—if a big country gets 100 medals and a small one gets a single medal, then at the very least it is not certain that the small country with its single medal is a 'robust signal'—that small country might easily be replaced by any of several other small countries with a good chance of getting about one medal, while the big country with its 100 medals is almost certainly a 'stable' result.

A part of a reply from someone from the Université René Descartes of Paris reads as follows:

> Then you must think if what you observed is really a random sample or not. I would personally say here that not in that context and that instead we have an exhaustive examination of the results, but let's assume than yes. In that case, you can ask "does a C1's athletes significantly has a higher probability to obtain a medal than a C2's athletes?". And here, you can apply "usual" statistical methods. But, with only n = 1 in the C1 sample, there cannot be any significant difference here at usual alpha levels. So, in that case, concluding that "C1 is better than C2" may be seen as misuse of statistics.

A medical doctor from Shiraz University of Medical Science said:

> I believe that the interpretation of the data is more important. For instance, in Olympic Games, the total number of medals is important not the percentage of medal winners. But in another issue (e.g., incidence of a disease) the percentage is important not the number. So you should interpret the data based on your variable and background.

Hence, as it can be seen, statisticians are not unanimous in interpreting a statistical presentation. This, of course, is because they look much deeper in the presented than just a presentation itself. However, people not in the profession of statistics take it at face value, and that is where misleading and "misuse" enters the equation.

Another grouping of data is a *frequency distribution with classes* (or *intervals*). Here is how we construct it:

1. Find the range.
2. Decide how many classes you want to have. However, we don't want too many or too few classes.

3. Classes may be chosen with equal lengths. If that is the case, divide the range by the number of classes selected and choose the rounded up number. This number will be the *length* (or *size*) *of a class*.
4. Now to avoid overlaps, move 0.5 (if data are with integral digits and 0.05 if data are with no digits and start with the tenth decimal place, and 0.005 if the data start with the hundredth decimal place, and so on), down and up from each end of the interval (*class boundaries*) and choose the interval half-open on the right end.
5. If, however, the length of an interval is given, then divide the range by the length of an interval and choose the rounded up number to find the number of classes.
6. After the intervals have been determined, count the number of data points in each class to find the frequency for each class. We note that a sample point such that there is an unusually large gap between the largest and the smallest data point is called an *outlier* (or an *extreme data point*).

In addition to grouping data, there are several graphic ways that a set of data may be presented such as *histogram*, *dot plot*, *box plot*, *stem-and-leaves*, and *scatter plot*. The *histogram* is one of the most common graphic presentations of a data set. It displays data that have been summarized into class intervals. The histogram is a graph that indicates the "shape" of a sample. It can be used to assess the *symmetry* or *skewness* of the data. To construct a histogram: (1) the horizontal axis is divided into equal intervals, (2) a vertical bar (or a strip) is drawn at each interval to represent its frequency (the number of data points that fall within the interval), and (3) adjacent rectangles are constructed, with (4) the bases of the rectangles representing the end points of the class intervals and the heights representing the frequencies of the classes.

Using the concept of a random variable, we can redefine the random sample we defined in Section 1.1 as follows: a random sample $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is defined as a sample consisting of n independent random variables with the same distribution. Each component of a random sample is a random variable representing observations. The term "random sample" is also used for a set of observed values $x_1, x_2, \ldots, x_n$ of the random variables. We should, however, caution that it is not always easy to select a random sample, particularly when the population size is very large and the sample size is to be small. For instance, to select a sample of size 10 cartons of canned soup to inspect thousands of cartons in the storage, it is almost impossible to number all these cartons and then choose 10 at random. Hence, in cases like this, we do not have many choices; we have to do the best we can and hope that we are not seriously violating the randomness property of the sample.

A statistic is itself a random variable because the value of it is uncertain prior to gathering data. Denoted by, say λ, a statistic is usually calculated based on a random sample. In other words, a statistic is a function of the random vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, say $\hat{\Lambda}(\mathbf{X}) = f(X_1, X_2, \ldots, X_n)$.

In Section 1.3, we discussed moments of a random variable that included mean and variance. We now define the concept of random sampling and its properties that includes sample mean and sample variance.

Definition 1.10.1
Let the random variables $X_1, X_2, \ldots, X_n$ be a sample of size n chosen from a population (of *infinite size*) in such a way that each sample has the same chance to be selected. This type of sampling is called *random sampling* (or *random sampling from an infinite population*) and the result is called a *random sample*. If the population is *finite of size N*, then a sample of size n from this population such that each sample of the combination:

$$\binom{N}{n},$$

would be referred to as a *random sample*.

The reader is encouraged to show that in sampling from a finite population, if selections are without replacement, then the random variables $X_1, X_2, \ldots, X_n$ are not mutually independent.

A population may be identified by its distribution $F(x)$. In that case, each X_i, $i = 1, 2, \ldots, n$ is an observation and has a marginal distribution $F(X)$. Additionally, each observation is taken such that its value has no effect or relationship with any other observation. In other words, $X_1, X_2, \ldots, X_n$ are mutually independent.

1.10.1. Measures of Central Tendency

Here in this section, we define the mean, median, and the mode for a distribution as examples of *measures of central tendency*. Each measure of central tendency defined loses some information of data points.

In Section 1.3, we defined *arithmetic average* or simply the *average*. In fact that is the mean of a sample as we define next.

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a random vector with $x_1, x_2, \ldots, x_n$ as values of observations. The statistic *sample mean* denoted by $\bar{x}$ is then defined as:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{1.10.1}$$

Now suppose we have a sample of size n, and p is a number between 0 and 1, exclusive. The $(100p)$th sample *percentile* is then a sample point at which there are approximately np sample points below it and $n(1 - p)$ sample point above it. The 25th *percentile* (also called the *first quartile*, denoted by Q_1) and the 75th *percentile* (also called the *third quartile*, denoted by Q_3) are the highest value for the lowest 25% of the data points and the lowest value for the highest 25% of the data points, respectively. Similarly, we could have *deciles* that are

the percentiles in the 10th increments, such as the *first deciles as the 10th percentile*, *the fifth deciles as the 50th percentile* or *median*, and so on. The *interquartile range*, denoted by IQR, is the range of the middle 50% of the data points and is calculated as the difference between Q_3 and Q_1, that is, $Q_3 - Q_1$. The *median* (also called the *second quartile*) of n (nonmissing) observations $x_1, x_2, \ldots, x_n$ is loosely defined as the *middlemost* of the observed values.

To find the median, arrange $x_1, x_2, \ldots, x_n$ according to their values in ascending or descending order, and then pick the middle value if n is odd, that is, $(n + 1)/2$, and the average (mean) of the middle two if n is even, that is, $[(n/2 + (n + 2)/2]/2$. In general, to find a percentile, we consider three cases for the $(n + 1)p$, namely, (1) $(n + 1)p$ as an integer, (2) $(n + 1)p$ as an integer plus a proper fraction, and (3) $(n + 1)p < 1$. Then:

1. If $(n + 1)p$ is an integer, the $(100p)$th sample percentile is the $(n + 1)p$th ordered data point display.
2. If $(n + 1)p$ is not an integer, but is equal to $r + a$, where r is the whole part and a is the proper fraction part of $(n + 1)p$, then take the weighted average of the rth and $(r + 1)$st ordered data points. In other words, let D_r = the rth ordered data point and the D_{r+1} = $(r + 1)$st ordered data point. Next, denoting the weighted average by π_p, we will have $\pi_p = D_r + a(D_{r+1} - D_r) = (1 - a)D_r + aD_{r+1}$.
3. If $(n + 1)p < 1$, then the sample percentile is not defined.

The median is less sensitive to extreme values than the mean. Thus, when data contain outliers, or are *skewed* (lack of symmetry), the median is used instead of the mean. If one tail extends farther than the other, we say the distribution is *skewed*. Outliers cause *skewness*. An outlier on the far left will cause a *negative or left skew*, while on the far right will cause a *positive* or *right skew*. To avoid outliers, it is customary to *trim* the data. The *trimmed mean* is the mean of the *trimmed data*, that is, of the *remaining data set* by cutting the data about 5–10% (rounded to the nearest integer) on each end (after being sorted in ascending or descending order). The measure of the sharpness of the peak of a distribution is referred to as *kurtosis*. Similar to skewness, positive or negative values of kurtosis will cause a peak flatter than or sharper than the peak of the normal curve.

The most frequent of the observed values is called *mode*. When a distribution has more than one mode, the data set is said to have *multimodes*. If this number is two, it is called *bimodal*.

1.10.2. Measure of Dispersion

Now that we have discussed measures of central tendency, we will discuss measures of dispersion. Range, as defined before, is a statistic that is often used to describe dispersion in data sets. As another measure of dispersion, let

$\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be a random vector of observations with sample mean $\bar{x}$. The statistic *sample variance* (sometimes called *mean square*) denoted by S^2 with its values as s^2, is then defined as:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{1.10.2}$$

Note that in the denominator of (1.10.2), the term $n - 1$ instead of n has been used. Although logically we should use n (and some authors in their publications a couple of decades ago used n), it tends to underestimate the population variance σ^2. Since the primary use of the sample variance s^2 is to estimate the population variance σ^2, by replacing $n - 1$ and, thus, enlarging s^2, the tendency is corrected. The same correction is made in the use of sample *standard deviation*, which is the positive square root of the sample variance.

In practice, the following formula (Formula 1.10.3) for sample variance is used instead of (1.10.2). It is easy to see that (1.10.2) is equivalent to (1.10.3):

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]. \tag{1.10.3}$$

In case that data are grouped in a simple frequency format, the sample variance would be calculated using:

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{k} f_i x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} f_i x_i\right)^2\right], \quad 1 \le k \le n, \tag{1.10.4}$$

where $f_i, 1 \le k \le n$ denote frequencies of each group of data points.

We note that it is usually desired to have the variance as small as possible so that data points are gathered around the mean. However, there might be cases that we have to deal with highly scattered data such that the mean is smaller than the variance or the standard deviation. There is no theory avoiding such cases. Particularly, since mean and variance are with different units of measurement, it would be difficult to compare the two.

In the following example, we will construct a histogram using the statistical software MINITAB. Using MINITAB has the advantage of providing information such as the measure of central tendencies. We note that this example is taken from Haghighi et al. (2011a, p. 265).

Example 1.10.3

Consider Example 1.7.1. The median for this data set is Q_2, that is, $p = 0.50$. Hence, $(n + 1)p = (31)(0.5) = 15.5$. Thus, $r = 15$ and $a = 0.5$. Therefore, Q_2 is the average of the 15th and the 16th data points, which is 17.1. In contrast, for the 53rd percentile, $(31)(0.53) = 16.43$, yielding $r = 16$ and $a = 0.43$. Hence:

$$\overline{\pi_{0.53}} = (1 - 0.43)D_{16} + 0.43D_{17} = (0.57)(17.1) + (0.43)(18.3) = 17.616.$$

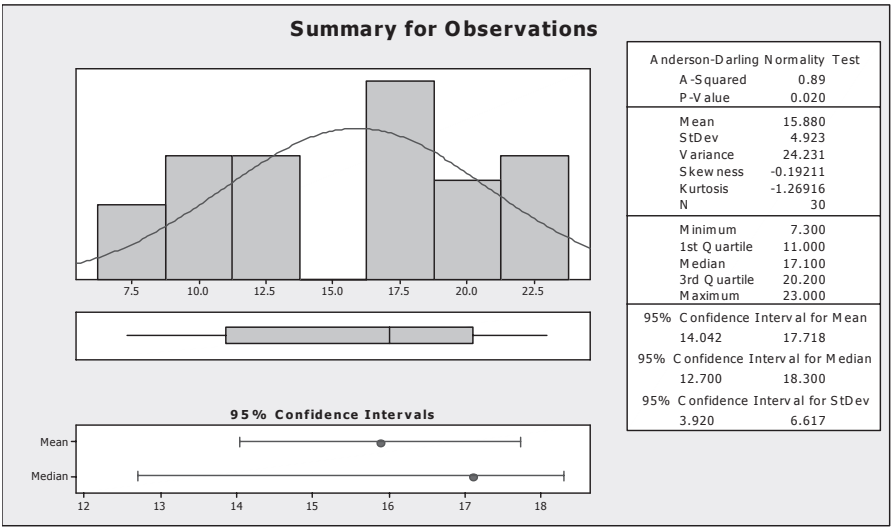
In addition, the 95th percentile is:

$$\overline{\pi_{0.95}} = (1-0.45)D_{29} + 0.45D_{30} = (0.55)(23) + (0.45)(23) = 23.$$

Using MINITAB, we obtain the following information as well as a histogram. Note that the "confidence interval" will be discussed later in this chapter.

Descriptive Statistics: Observations

```
Sum of
Variable      Mean     SE Mean TrMean StDev   Variance CoefVar Squares
Observations 15.880   0.899   15.969 4.92    24.231   31.00   8267.940
Variable      Minimum Q1      Median Q3      Maximum  Range   IQR Skewness
Observations 7.300    11.000  17.100 20.200 23.000   15.700  9.200 -0.19
```



1.10.3. Properties of Sample Statistics

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a population with mean and variance μ and $\sigma^2 < \infty$, respectively. Denoting the sample mean and variance by $\bar{X}$ and S^2, respectively, we then leave it as an exercise to prove the following properties of sample mean:

$$E(\bar{X}) = \mu, \tag{1.10.5}$$

and

$$Var\left(\bar{X}\right) = \frac{\sigma^2}{n}. \tag{1.10.6}$$

A measure of variability of the sampling distribution of the sample mean is called the *standard error of the sample mean*. If the population standard deviation, σ, is *known*, then the standard error of the sample mean, denoted by $\sigma_{\bar{X}}$, is defined by:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}. \tag{1.10.7}$$

If the population standard deviation is *unknown*, and denoting S as the *sample standard deviation*, then the *standard error of the sample mean*, denoted by $S_{\bar{X}}$, is:

$$S_{\bar{X}} = \frac{S}{\sqrt{n}}. \tag{1.10.8}$$

We now want to show that the mean of sample variance, as we defined in (1.10.2), is, indeed, the variance of the population.

Theorem 1.10.1

$$E\left(S^2\right) = \sigma^2, \tag{1.10.9}$$

where S^2 is defined in (1.10.2).

Proof:
Recall from (1.10.2) that the observed value of the sample variance is defined as:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

as defined in (1.10.1). Now, we add and subtract μ in $\sum_{i=1}^{n}(x_i-\bar{x})^2$ and do some algebra as follows:

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x})^2 &= \sum_{i=1}^{n}(x_i-\mu+\mu-\bar{x})^2 = \sum_{i=1}^{n}[(x_i-\mu)-(\bar{x}-\mu)]^2\\ &= \sum_{i=1}^{n}\left[(x_i-\mu)^2-2(x_i-\mu)(\bar{x}-\mu)+n(\bar{x}-\mu)^2\right]\\ &= \sum_{i=1}^{n}(x_i-\mu)^2-2(\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\mu)+n(\bar{x}-\mu)^2\\ &= \sum_{i=1}^{n}(x_i-\mu)^2-2(\bar{x}-\mu)\left[n\frac{\sum_{i=1}^{n}x_i}{n}-n\mu\right]+n(\bar{x}-\mu)^2\\ &= \sum_{i=1}^{n}(x_i-\mu)^2-2(\bar{x}-\mu)(nx_i-n\mu)+n(\bar{x}-\mu)^2\\ &= \sum_{i=1}^{n}(x_i-\mu)^2-n(\bar{x}-\mu)^2.\end{aligned}$$

Thus,

$$\begin{aligned}E(s^2) &= E\left[\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}\right] = \frac{1}{n-1}\left(E\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]-nE\left[(\bar{x}-\mu)^2\right]\right)\\ &= \frac{1}{n-1}\left[n\sigma^2-nVar(\bar{x})\right] = \frac{n}{n-1}\left[\sigma^2-Var(\bar{x})\right]\\ &= \frac{n}{n-1}\left(\sigma^2-\frac{\sigma^2}{n}\right) = \sigma^2.\end{aligned}$$

Now consider a population with a mean and variance of μ and σ^2, respectively. As n increases without bound, we see from (1.10.6) that the variance of $\bar{X}$ will decrease. Hence, the distribution of $\bar{X}$ depends on the sample size n. In other words, for different sizes of n, we will have a sequence of distributions to deal with. To see the limit of such a sequence as n increases without bound, let us consider the random variable W defined by:

$$W=\frac{X-\mu}{\sigma/\sqrt{n}},\quad n=1,2,\ldots. \tag{1.10.10}$$

We then leave it as an exercise to show that W is a standard normal random variable for each positive integer n. Therefore, the limiting distribution of W

is also a standard normal, which we will state as the following theorem known as the *central limit theorem*, and we refer the reader to Hogg and Tanis (1993) for the proof.

Theorem 1.10.2. ***The Central Limit Theorem***
Suppose $X_1, X_2, \ldots, X_n$ is a random sample of size n from a distribution with finite mean μ and a finite positive variance σ^2. Then the limiting distribution (i.e., as $n \to \infty$) of

$$W = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$$

is the standard normal.

As we noted before, we are using $n - 1$ instead of n in the sample variance. We stated the reason for such a choice. That choice actually causes the obtaining of (1.10.10), which makes the sample variance the so-called *unbiased* estimator for the population variance.

To find the variance of the sample variance, we remind the reader of equations (1.10.2) and (1.10.9). We leave it as an exercise to show that the second moment of S^2 is:

$$E\left[\left(S^2\right)^2\right] = \frac{\mu_4}{n} + \frac{(n-1)^2 + 2}{n(n-1)}\sigma^4, \tag{1.10.11}$$

where μ_4 is the fourth central moment of X. Hence, from (1.10.2) we will have:

$$\begin{aligned} Var\left(\sigma^2\right) &= \frac{\mu_4}{n} + \frac{(n-1)^2 + 2}{n(n-1)}\sigma^4 - \left(\sigma^2\right)^2 \\ &= \frac{\mu_4}{n} + \sigma^4\left[\frac{(n-1)^2 + 2}{n(n-1)} - 1\right] \\ &= \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right). \end{aligned} \tag{1.10.12}$$

Example 1.10.4
As an example, let us find the mean and variance of the sampling distribution of the standard deviation for Gaussian (normal) distributions.

Answer
What the question says is that we have a standard normal population. Take a sequence of samples from this population, calculate the standard deviation for each sample taken, and consider the values of these standard deviations as a set of data or new sample points. The question is, what are the mean and standard deviation of this set of data?

Let $X_1, X_2, \ldots, X_n$ be a simple random sample (i.e., a set of n independent random variables) from a normal population with mean μ and variance σ^2. Now, multiplying both sides of (1.10.2) by $(n-1)$ and dividing by σ^2, we will have:

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{\sigma^2}. \tag{1.10.13}$$

We leave it as an exercise to show that:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1). \tag{1.10.14}$$

Therefore, let:

$$W \equiv \frac{(n-1)S^2}{\sigma^2}. \tag{1.10.15}$$

From (1.10.13) and (1.10.15), W is $\chi^2(n-1)$. Thus, the pdf of W, denoted by $g_W(w)$ is:

$$g_W(w) = \frac{w^{\frac{n-1}{2}-1} e^{\frac{-w}{2}}}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)}, \quad w \geq 0$$

or

$$g_W(w) = \frac{w^{\frac{n-3}{2}} e^{\frac{-w}{2}}}{2^{\frac{n-1}{2}}\Gamma\left(\frac{n-1}{2}\right)}, \quad w \geq 0. \tag{1.10.16}$$

From (1.10.15), we have:

$$S^2 = \frac{\sigma^2 W}{n-1}. \tag{1.10.17}$$

Letting S denote the sample standard deviation for the normal population, from (1.10.17) we have:

$$S = \frac{\sigma}{\sqrt{n-1}}\sqrt{W}. \tag{1.10.18}$$

We define the random variable Y as:

$$Y \equiv \sqrt{W} \tag{1.10.19}$$

To find the pdf of Y, we note that since W is a continuous-type random variable with pdf of $g_W(w)$ as defined in (1.10.15), for $w \in [0, \infty)$, Y is defined as a

function of W, as in (1.10.19), say $Y = r(W)$, and Y and its inverse $W = f(Y)$ are increasing continuous functions, then the pdf of Y is:

$$h_Y(y) = g[f(y)]f'(y). \tag{1.10.20}$$

Hence, since from (1.10.19), $W = f(Y) = Y^2$ and thus $f'(Y) = 2Y$. From (1.10.20), we have:

$$h_Y(y) \equiv g_W(y^2)2y = \frac{y^{n-2}e^{\frac{-y^2}{2}}}{2^{\frac{n-3}{2}}\Gamma\left(\frac{n-1}{2}\right)}, \quad y \geq 0. \tag{1.10.21}$$

Then:

$$\begin{aligned} E(Y) &= \int_0^\infty yh(y)dy = \int_0^\infty \frac{y^{n-1}e^{\frac{-y^2}{2}}}{2^{\frac{n-3}{2}}\Gamma\left(\frac{n-1}{2}\right)}dy \\ &= \frac{1}{2^{\frac{n-3}{2}}\Gamma\left(\frac{n-1}{2}\right)}\int_0^\infty y^{n-1}e^{\frac{-y^2}{2}}dy. \end{aligned} \tag{1.10.22}$$

Now, let $u = y^2/2$, from which we have $du = ydy$ and $y^{n-1} = 2^{n-1}u^{(n-1)/2}$. Then, from (1.10.22), we have:

$$\begin{aligned} E(Y) &= \frac{1}{2^{\frac{n-3}{2}}\Gamma\left(\frac{n-1}{2}\right)}\int_0^\infty 2^{\frac{n-1}{2}}u^{\frac{n-1}{2}}e^{-u}\frac{du}{2^{1/2}u^{1/2}} \\ &= \frac{2^{\frac{n-2}{2}}}{2^{\frac{n-3}{2}}\Gamma\left(\frac{n-1}{2}\right)}\int_0^\infty u^{\frac{n-1}{2}}e^{-u}u^{-\frac{1}{2}}du \\ &= \frac{2^{\frac{n-2}{2}}}{2^{\frac{n-3}{2}}\Gamma\left(\frac{n-1}{2}\right)}\int_0^\infty u^{\frac{n}{2}-1}e^{-u}du \\ &= \frac{2^{\frac{n-2}{2}}\Gamma\left(\frac{n}{2}\right)}{2^{\frac{n-3}{2}}\Gamma\left(\frac{n-1}{2}\right)} \\ &= \frac{2^{\frac{1}{2}}\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}. \end{aligned} \tag{1.10.23}$$

Now, from (1.10.18) and (1.10.19) we have:

$$E(S) = \frac{\sigma}{\sqrt{n-1}} E(Y). \tag{1.10.24}$$

Substituting (1.10.23) in (1.10.24), the mean of standard deviation of the normal population is:

$$E(S) = \sqrt{\frac{2}{n-1}} \frac{\sigma \Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}. \tag{1.10.25}$$

It is known that:

$$Var(S) = E\left(S^2\right) - (E(S))^2. \tag{1.10.26}$$

Since the mean of $\chi^2(n-1)$ is $n-1$, from (1.10.14), we will have:

$$E\left(S^2\right) = \frac{\sigma^2}{n-1} E(W) = \sigma^2. \tag{1.10.27}$$

Thus, from (1.10.25) and (1.10.27), the variance of the standard deviation of the normal population is:

$$Var(S) = \sigma^2 \left\{ 1 - \frac{2}{n-1} \left[\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \right]^2 \right\}. \tag{1.10.28}$$

For the standard normal population, $\mu = 0$ and $\sigma = 1$. Therefore, the mean of standard normal standard deviation remains as in (1.10.25), and its variance from (1.10.26) will be:

$$Var(S) = 1 - \frac{2}{n-1} \left[\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \right]^2. \tag{1.10.29}$$

1.11. INFERENTIAL STATISTICS

The purpose of most statistical investigations is to generalize information contained in samples to the populations from which the samples were drawn. This is the essence of statistical inference. The term *inference* means a conclusion or a deduction. Two major areas are main components of the methods of *statistical inference* in the classical approach: (1) *hypotheses testing* and (2)

estimation (*point* and *interval*). In summary, an *estimate* of a distribution based on a sample and drawing a conclusion about a parameter based on a sample is called a *statistical inference*. Of the different methods of estimation that are available, we discuss *point* and *interval estimations*.

1.11.1. Point Estimation

Definition 1.11.1
The most plausible value of a parameter μ (population value) is called the *point estimate* of μ. The statistic (sample value) that estimates this parameter is called the *point estimator* of μ and is denoted by $\hat{\mu}$. In other words, we want to find a number $\hat{\mu}$ as a function of observations $(x_1, x_2, \ldots, x_n)$, that is, $\hat{\mu} = Y(x_1, x_2, \ldots, x_n)$. The function Y is a statistic that estimates μ, that is, a *point estimator* for μ. We want the computed value $\hat{\mu} = Y(x_1, x_2, \ldots, x_n)$ to be closed to the actual value of the parameter μ.

Note that Y is a random variable and thus has a pdf by its own.

It is known that a good estimator is the one that, on the average, is equal to the parameter (*unbiased*) and its *variance is as small as possible* (*minimum variance*). In other words, a point estimator, denoted by $\hat{\mu}$, of a parameter μ is *unbiased* if $E(\hat{\mu}) = \mu$, otherwise it is said to be *biased*. The amount of *bias* of $\hat{\mu}$, denoted by $B(\hat{\mu})$, is defined by $B(\hat{\mu}) = E(\hat{\mu}) - \mu$.

Let μ be a parameter and $\hat{\mu}$ an estimator of it. Then the *mean square error* of $\hat{\mu}$, denoted by MSE, (and if there is confusion, by $MSE(\hat{\mu})$) is defined by:

$$MSE(\hat{\mu}) \equiv MSE = [E(\hat{\mu}) - \mu]^2 + Var(\hat{\mu}). \tag{1.11.1}$$

There are different methods of point estimation. One common one is the *method of moments*. However, the maximum likelihood estimation method (MLE), which is defined later, is the most widely used method of estimation, especially when the sample size is large. It is a method that allows one to choose a value for the unknown parameter that most likely is the closest to the observed data.

We note that in the maximum likelihood function described later, for discrete distribution function we use the probability mass function, and for continuous distribution function we use the probability density function. However, we will use the pdf notation for both cases. Hence, we will use $f(x; \theta)$ for a density function with one parameter and $f(x; \theta_1, \theta_2)$ for a density function with two parameters.

Definition 1.11.2
Suppose $(X_1, X_2, \ldots, X_n)$ is a random sample of size n, with observed values $(x_1, x_2, \ldots, x_n)$ from a cumulative distribution function (cdf) $F(x; \theta)$ with pdf $f(x; \theta)$, where θ is a vector of k parameters $\theta_1, \theta_2, \ldots, \theta_k$. The cdf of this random sample denoted by $F_n(x_1, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_k)$ is:

$$F_n(x_1, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_k) = \prod_{i=1}^{n} F(x_i; \theta_1, \theta_2, \ldots, \theta_k). \tag{1.11.2}$$

For the given sample $(x_1, x_2, \ldots, x_n)$, the quantity $dF_n(x_i; \theta) = \prod_{i=1}^{n} dF(x_i; \theta)$ is called the *likelihood function* of $\theta_1, \theta_2, \ldots, \theta_k$ for $(x_1, x_2, \ldots, x_n)$. We denote the likelihood function of θ by $L(\theta)$,

$$L(\theta) = L(x_1, \ldots, x_n; \theta_1, \theta_2, \ldots, \theta_k) = \prod_{i=1}^{n} f(x_i; \theta_1, \theta_2, \ldots, \theta_k), \tag{1.11.3}$$

that is, the joint density function of n random variables and k parameters. For each sample point x, let $\hat{\theta}(x)$ be a value of the parameter that maximizes $L(\theta)$. The *maximum likelihood estimator* (MLE) of the parameter θ based on a sample $(X_1, X_2, \ldots, X_n)$ is denoted by $\hat{\theta}(X)$, where $X = (X_1, X_2, \ldots, X_n)$.

We note that in order find the value of θ that maximizes $L(\theta)$, we take the derivative of $L(\theta)$, set it equal to zero, and find θ. If θ is a vector of, say, size k, then we need to take partial derivatives with respect to each element of θ, set it equal to zero, and solve the system of k equations with k unknowns. That is, if the likelihood function is differentiable with respect to θ, then letting x represent the random sample, potential candidates for the MLE are the values of $(\theta_1, \theta_2, \ldots, \theta_k)$ that solve:

$$\frac{\partial L(x; \theta_i)}{\partial \theta_i} = 0, \quad i = 1, 2, \ldots, k.$$

In most cases, it is easier to use derivatives of the natural logarithm of $L(\theta)$ rather than $L(\theta)$, which is called the *log likelihood function*. This is because the logarithmic function is strictly increasing on $(0, \infty)$ and that implies that the extrema of $\ln L(\theta)$ and $L(\theta)$ coincide (we leave the proof of this statement as an exercise).

Example 1.11.1. ***Estimating Poisson Parameter by MLE***
Suppose $x_1, x_2, \ldots, x_n$ are observed values of a Poisson random variable with parameterλ representing the random sample $X_1, X_2, \ldots, X_n$ of size n. We want to estimate λ using MLE.

Answer
From the Poisson assumption with parameter λ, the probability of observing x_i events in the ith trial is:

$$p_X(x_i; \lambda) \equiv P(X_i = x_i; \lambda) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}, \quad i = 1, 2, \ldots, n; x_i = 0, 1, 2, \ldots. \tag{1.11.4}$$

We now take a random sample of size n, say $X_1, X_2, \ldots, X_n$. Let $(x_1, x_2, \ldots, x_n)$ denote the set of observations of $(X_1, X_2, \ldots, X_n)$. Then, from (1.11.3) and (1.11.4), we have the likelihood function as:

$$L(\lambda) = L(x_1, \ldots, x_n; \lambda) = \prod_{i=1}^{n} \left(\frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \right), \tag{1.11.5}$$

and the log likelihood function as:

$$\ln L(\lambda) = \sum_{i=1}^{n} (x_i \ln \lambda) - n\lambda - \sum_{i=1}^{n} \ln (x_i!). \tag{1.11.6}$$

We regard (1.11.6) as a function of λ and will find the value of λ that maximizes this likelihood function. Taking derivative with respect to λ of (1.11.6) and set it equal to zero, we obtain:

$$\frac{d \ln L(\lambda)}{d\lambda} = \frac{1}{\lambda} \sum_{i=1}^{n} x_i - n = 0,$$

from which

$$\lambda = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}. \tag{1.11.7}$$

Thus, MLE for λ is $\bar{x}$, which is the sample mean.

Example 1.11.2. ***Estimating Weibull Parameters by MLE***
In Example 1.11.1, we estimated the parameter of a discrete distribution with one parameter. In this example, we want to show how to apply MLE to estimate parameters of a continuous-type distribution with two parameters.

We note that the contents of this example are straightforward materials, and perhaps this is why similar formulae and discussions appear in different textbooks in the literature without any reference to each other. It should also be noted that unfortunately, some properties of Weibull distribution have appeared in some published papers that are incorrect. Hence, readers must be careful when referring to such papers without making necessary corrections.

We consider the two-parameter Weibull distribution function defined in (1.6.34) and its pdf by (1.6.35). Thus, from (1.11.3), the likelihood function is:

$$\begin{aligned} L(\alpha, \beta) = L(x_1, \cdots, x_n; \alpha, \beta) &= \prod_{i=1}^{n} f(x_i; \alpha, \beta) \\ &= \prod_{i=1}^{n} \frac{\beta}{\alpha} \left(\frac{x_i}{\alpha} \right)^{\beta - 1} e^{-\left(\frac{x_i}{\alpha} \right)^{\beta}}. \end{aligned} \tag{1.11.8}$$

Hence, for $n \geq 1$, using log likelihood functions, the *likelihood equations* for α and β are as follows:

$$\begin{cases} \dfrac{\partial \ln L}{\partial \beta} = \dfrac{n}{\beta} - n \ln \alpha + \displaystyle\sum_{i=1}^{n} \ln x_i - \sum_{i=1}^{n} \left[\left(\frac{x_i}{\alpha} \right)^{\beta} \ln \left(\frac{x_i}{\alpha} \right) \right] = 0, \\ \dfrac{\partial \ln L}{\partial \alpha} = -\dfrac{n\beta}{\alpha} + \dfrac{\beta}{\alpha} \displaystyle\sum_{i=1}^{n} \left(\frac{x_i}{\alpha} \right)^{\beta} = 0. \end{cases} \tag{1.11.9}$$

To solve the system (1.11.9), we eliminate α from the first equation of (1.11.9). From the second equation of (1.11.9), we have:

$$\sum_{i=1}^{n} \left(\frac{x_i}{\alpha} \right)^{\beta} = n \quad \text{and} \quad \frac{1}{\alpha^{\beta}} = \frac{1}{n} \sum_{i=1}^{n} x_i^{\beta}. \tag{1.11.10}$$

Now from the first equation of (1.11.9), we have:

$$\frac{n}{\beta} - n \ln \alpha + \sum_{i=1}^{n} \ln x_i - \sum_{i=1}^{n} \left(\frac{x_i}{\alpha} \right)^{\beta} \ln x_i + \ln \alpha \sum_{i=1}^{n} \left(\frac{x_i}{\alpha} \right)^{\beta} = 0. \tag{1.11.11}$$

Using (1.11.10) and cancelling $n \ln \alpha$ terms, (1.11.11) will yield:

$$\frac{n}{\beta} + \sum_{i=1}^{n} \ln x_i - \sum_{i=1}^{n} \left(\frac{x_i}{\alpha} \right)^{\beta} \ln x_i = \frac{n}{\beta} + \sum_{i=1}^{n} \ln x_i - \frac{1}{\alpha^{\beta}} \sum_{i=1}^{n} x_i^{\beta} \ln x_i = 0,$$

$$\frac{n}{\beta} + \sum_{i=1}^{n} \ln x_i - \frac{1}{\dfrac{\sum_{i=1}^{n} \ln x_i^{\beta}}{n}} \sum_{i=1}^{n} x_i^{\beta} \ln x_i = 0.$$

Hence:

$$\frac{\displaystyle\sum_{i=1}^{n} x_i^{\hat{\beta}} \ln x_i}{\displaystyle\sum_{i=1}^{n} x_i^{\hat{\beta}}} - \frac{1}{\hat{\beta}} - \frac{1}{n} \sum_{i=1}^{n} \ln x_i = 0, \tag{1.11.12}$$

from which $\hat{\beta}$ should be found. Substituting $\hat{\beta}$ in the second equation of (1.11.9), will give $\hat{\alpha}$ as:

$$\hat{\alpha} = \left[\frac{1}{n} \sum_{i=1}^{n} x_i^{\hat{\beta}} \right]^{1/\hat{\beta}}. \tag{1.11.13}$$

Since (1.11.12) cannot be solved analytically, we apply Newton's iterative method:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \tag{1.11.14}$$

where

$$f\left(\hat{\beta}\right) = \frac{1}{\hat{\beta}} + \frac{1}{n}\sum_{i=1}^{n} \ln x_i - \frac{\sum_{i=1}^{n} x_i^{\hat{\beta}} (\ln x_i)}{\sum_{i=1}^{n} x_i^{\hat{\beta}}}, \tag{1.11.15}$$

and

$$f'\left(\hat{\beta}\right) = -\frac{1}{\hat{\beta}^2} - \frac{\sum_{i=1}^{n} x_i^{\hat{\beta}} (\ln x_i)^2}{\sum_{i=1}^{n} x_i^{\hat{\beta}}} + \left(\frac{\sum_{i=1}^{n} x_i^{\hat{\beta}} (\ln x_i)}{\sum_{i=1}^{n} x_i^{\hat{\beta}}} \right)^2, \tag{1.11.16}$$

to find $\hat{\beta}$ numerically. To do that, we would need to choose a sample of observations. That can be done by simulation.

We should note that we went through the lengthy process to show the MLE method. However, we could use the MATLAB computer program, for instance, to find the MLE estimate of parameters numerically.

1.11.2. Interval Estimation

Although point estimation has its usage, it also has its limitations. For instance, point estimation needs to be accompanied by an MSE, as we mentioned earlier. In reality, a point estimate cannot be expected to coincide with the quantity we are to estimate. Hence, to avoid limitations, that is, to determine the precision of the estimate, the *interval estimator* is used. That is, we can assert with some level of certainty that the interval contains the parameter under consideration. Such an interval is called a *confidence interval* for the parameter. The lower and upper limits for this interval are constructed in such a way that the true value of the estimate falls within the interval. The upper end point is usually obtained by adding a multiple value of the standard deviation to the estimated value and the lower end point is obtained by subtracting the same multiple values from the estimated mean. The likeliness of the true value falling within the interval is referred to as the *level* (or *degree*) *of confidence*.

Example 1.11.1

We use this example to show how to construct a confidence interval. Suppose we have a large random sample of size n (i.e., $n \geq 30$), say $X_1, X_2, \ldots, X_n$, from a normal population with a *known variance* σ^2 and we want to estimate the unknown mean, μ. We want to estimate μ and find a 95% confidence for the estimate.

Answer

Let $\bar{X}$ be the sample mean. We have seen that it is an unbiased point estimator of the population mean μ. We also have seen that:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{1.11.17}$$

is a random variable that has an approximately standard normal distribution. The end points of the interval for which the value of the estimator falls within, with, say, 95% probability, can be looked up from the standard normal tables. Hence, we will see that $\bar{X}$ falls within the 1.96 standard deviation of mean 0.95. That is,

$$P(-1.96 < z < 1.96) = 0.95. \tag{1.11.18}$$

This is because 2.5% of the area under the standard normal density curve is to the right of the point $z = 1.96$ ($z_{\alpha/2} = z_{.025} = 1.96$) and 2.5% of the area is to the left of –1.96. Hence, 95% of the area is between –1.96 and +1.96. Thus, from (1.11.17) and (1.11.18), we have:

$$P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95, \tag{1.11.19}$$

or

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95. \tag{1.11.20}$$

Therefore, since 1.96 is 97.5% of the standard normal distribution, a 95% confidence interval estimator of the population mean μ is:

$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right). \tag{1.11.21}$$

In other words, 95 times out of every 100 times we sample, μ will fall within the interval $\bar{X} \pm 1.96\left(\sigma/\sqrt{n}\right)$. For instance, suppose X represent the lifetime of a light bulb. Suppose also that X is normally distributed with mean μ and standard deviation of 42. The manufacturer chose 36 light bulbs and lighted them to observe their lifetime until each burned out. The mean of this sample was 1500 hours. To estimate the mean of all such light bulbs, a 95% confidence interval is decided. Thus, from (1.11.21) we will have:

$$\left(1500 - 1.96\frac{42}{\sqrt{36}}, 1500 + 1.96\frac{42}{\sqrt{36}}\right) = (1486.28, 1513.72).$$

We note that we can use the central limit theorem to approximate the confidence interval. This is because when the sample size is large enough, the ratio

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is $N(0,1)$. Hence, a $100(1 - \alpha)\%$ confidence interval for μ is:

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha, \tag{1.11.22}$$

and that implies that the *standard normal confidence interval* is:

$$\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right), \tag{1.11.23}$$

where $z_{\alpha/2}$ is the value of the area under the normal curve to its right. We note that the amount of error of the estimate is $\left|\bar{X} - \mu\right|$ and:

$$\text{the maximum error of estimate} < z_{\alpha/2}\sigma/\sqrt{n}\text{, with probability of } 1-\alpha. \tag{1.11.24}$$

Other percent confidence intervals can be found similarly.

In case the *variance is unknown*, the sample variance, S^2, can be used. In this case, however, we can again use the quantity $\left(\bar{X} - \mu\right)/\left(\sigma/\sqrt{n}\right)$ with the population standard deviation σ replaced by the sample standard deviation, S, that is,

$$T \equiv \frac{\bar{X} - \mu}{S/\sqrt{n}} \tag{1.11.25}$$

The T defined in (1.11.25) is a random variable since it depends upon S and $\bar{X}$; however, it is not normally distributed; rather, it has *Student's t-distribution* (or just simply, *t-distribution*). T may be treated as a standard normal as long as the sample size is large (30 or more), since in that case there is a high probability that S is very close to σ.

So far, we have discussed the confidence interval for the population mean based on large sample sizes ($n \geq 30$). Those discussions should ensue regardless of the distribution because of the central limit theorem. When the sample size is small ($n < 30$), S may not be close to σ and so $\bar{X}$ may not be normally distributed, unless the sample is selected from a normal population. For a small sample size, there is no good general method for finding the confidence interval. In such a case, the Student's t-distribution will be used along with the t-distribution tables. The *Student's t-distribution confidence interval*, then, would be:

$$\left(\bar{x} - t_{n-1,\alpha/2} S\sqrt{1+\frac{1}{n}}, \bar{x} + t_{n-1,\alpha/2} S\sqrt{1+\frac{1}{n}}\right). \tag{1.11.26}$$

Example 1.11.3
Suppose that to determine a value of interest (the melting point of tin, for instance), an experiment is conducted six times (sample size). The mean and standard deviation are found to be 0.14 and 232.26, respectively. If the researcher is to use the mean to estimate the actual value of interest (meting point) with high probability, say 98%, what would be the expected maximum error?

Answer
Since $1 - \alpha = 98\%$, $\alpha/2 = 0.01$, thus, from the t-table, for $n - 1 = 5$ (degrees of freedom) is $i_{0.01} = 3.365$. Hence, from (1.8.13), with probability of 98%, the maximum error will be less than $3.365 \times \left(0.14/\sqrt{6}\right) = 0.19$.

We note that one must be careful about rounding off numbers; the least significant digit should reflect the precision of the estimate. The precision depends on the variable, on the technique of measurement, and also on the sample size. It is not sufficient to round off just by taking into consideration the sample size alone.

1.12. HYPOTHESIS TESTING

A question such as if the average lifetime of a light bulb is more than 2000 hours is answered by first hypothesizing it and then testing it. This is the objective of this section.

By a *statistical hypothesis*, we mean a statement about the nature of a parameter of a population. By *testing of a hypothesis*, we mean to determine the truth or falsity of the statement. The way the testing is set up is that a default position is set, called the *null hypothesis*, denoted by H_0. The negation of the null hypothesis, H_0, is called the *alternative hypothesis*, denoted by H_1. A statistic whose value is determined from the sample observations is called a *test statistic*. The test statistic is used to decide to reject or accept the null hypothesis. *Critical region* (or *rejection region*) is the set of values of the test statistics for which H_0 is rejected. The *critical value* is the dividing point of the region between values for which the null hypothesis is rejected and not rejected. Testing a null hypothesis may cause two types of errors, called *Type I error* and *Type II error*.

If H_0 is erroneously rejected, it is said that the Type I error has occurred. In contrast, if H_0 is erroneously accepted, it is said that the Type II error has occurred. To assure that the decision of rejecting H_0 is the correct one, one requires that the probability of rejection be less than or equal to a small number, say α. The preassigned small number α, which is the *probability of a*

Type I error (i.e., the probability that H_0 is true and H_1 was not rejected), and is often chosen as 0.01, 0.05, or 0.10, is called the *level of significance of the test* (or *significance level of the test*). The smallest significance level at which the data lead to the rejection of H_0 is called the *p-value*. A small *p*-value (0.05 or less) strongly suggests that H_0 is not true. In contrast, if the *p*-value is large, it would mean that there is strong evidence not to reject H_0.

Let the *probability of Type II error* be denoted by β (i.e., the probability that H_1 is true and H_0 was not rejected). The probability of erroneously rejecting H_0, $1 - \beta$, is then called the *power of the test*. The ideal power function is $\beta = 0$. However, that is only possible in trivial cases. Hence, the practical hope is that the power function is near 1.

We note that increasing the sample size will reduce Type II error and increase power. However, it will not affect Type I error.

The comparison of means of several populations is called the *analysis of variance* (*ANOVA*). By that, it is meant estimating the means of more than one population simultaneously. ANOVA is one of the most widely used statistical methods. It is not about variance; rather, it is about analyzing variations in means. In statistical design, ANOVA is used to determine how one can get the most information on the most populations with a fewer observations. Of course, this desire is extremely important in the industry due to minimizing the cost.

In addition to ANOVA, the *t-test* or *error bar* (a graphical representation of the variability of data, which gives a general idea of how accurate a measurement is), when one uses multipliers based on the Student distribution, may be used for the same purpose. The error bar may also be used with normal, Poisson, and other distributions as well. For most straightforward situations, the multiplier lies between 1.95 and 2.00 for *p*-value less than the 0.05 test (two sided against a known chance/null baseline or one sided against an alternative result expected to beat). Usually, the multiplier is applied to the standard error of the cases one regards as the baseline or null hypothesis, but even when comparing two or more arbitrary conditions/systems in an ad hoc way with no strong a priori idea that a particular one should be better, perhaps it is preferred to make the error bars 1 standard error around points from each source, and regard differences as tentatively significant if the error bars do not cross, corresponding to an averaging of the conditions where each is considered the baseline for the other. By "tentative," is it meant that this acts as a filter and formal tests can be done for those that are interesting and potentially significant.

For a two-sided test, for a *p*-value to be considered significant, one needs to judge that 1.5 standard errors would not meet. This problem may be taken care of by including additional whiskers, so at ±1 and ±1.5 standard error.

A word of caution: One should not compare means in cases like repeated measures or pre- or posttest designs. In such cases, the individual differences across the system should be used. Comparing means is often too conservative and is not sensitive to correlation between the individual measurements.

Often there is interest in knowing the relationship between two variables. Finding the relationship and degree of dependence are what *regression analysis* does. Regression (in different forms such as linear, nonlinear, parametric, nonparametric, simple, and multiple) is widely used in different disciplines such as engineering, biological sciences, social and behavioral sciences, and business. For instance, we might be interested in analyzing the relation between two variables such as the strength of and stress on a beam in building a bridge; or a student's grade point average and his/her grade in a statistics course. The analysis of regression is not only used to show the relationship between variables, but it is also for prediction.

The simplest relation between two variables is a straight line, say $Y = \beta_0 + \beta_1 x$. Hence, for each value of x, the value of Y will be predicted, but not exactly when estimations are involved. In those cases, the values are subject to *random error*. Thus, in general, the linear equation considered is of the type $Y = \beta_0 + \beta_1 x + \varepsilon$, where β_0 and β_1 are *parameters*, called *regression coefficients*, and ε is what is called the *random error* with mean 0.

We note that the error terms ε for different trials are assumed to be uncorrelated. Thus, the outcome in any trial has no effect on the error term for any other trial. This implies that Y_i and Y_j, $i \neq j$, are uncorrelated. It should also be noted that, essentially, regression and correlation (that we discussed earlier) are the same. However, we cannot conclude causation from either one. Causes can only be established by an experiment. Causality cannot be determined by statistics; it needs to be assumed. In fact R.A. Fisher, one of the pioneers of statistics, has tried and failed. Correlation and regression are analyses of relationships among mathematical constructs. Causal models are hypothetical statements which guide both the choice of which variables should be measured as potentially relevant and the interpretation of associations that are found in the data.

We also make a general note on rounding numbers that usually causes error of some type. There are no standards to how to round a number to the nearest tenth or others. Essentially, the precision depends on the variable, on the technique of measurement, and also on the sample size. One cannot rule just by sample size alone. Some have set as their basic underlying principle the effect of rounding should always be less than 1% of the standard error of the observation or estimate. In other words, set the rounding error to be less than 1% of the calculated standard error.

One should make sure that rounding numbers is done only after all computations have been completed.

Some use the following rule, called the *rounding interval*, denoted, say, by r. It is defined as the smallest possible positive difference between two rounded values for the same statistic as reported in a study report. If results are reported in two decimal places, the rounding interval $r = 0.01$. Admissible values for r are powers of 10 only, that is, they all belong to the set $\{\ldots; 0.001; 0.01; 0.1; 1; 10; 100; 1000; \ldots\}$. For r, the maximum value should be selected from this set that does not exceed half the standard error of the observation or of the statistic. So $r \leq 1/2$ (standard error) $< 10r$. For instance, if some

statistic has a standard error of 0.042, then r should be selected as 0.01, since $0.01 < 0.021 < 0.1$.

Numbers are rounded to the nearest rounded value. If the choice is not confusing, the last digit after rounding should be even. For instance, for $r = 0.01$, we should have the following: 5.233 rounded to 5.23; 4.268 rounded to 4.27; 6.445 rounded to 6.43; and 6.435 rounded to 6.44.

1.13. RELIABILITY

It is now well known to the industry that for manufacturing goods, along with all factors considered such as ease of manufacturing, cost, size, weight, and maintenance, they must pay great attention to the reliability of their products. If a system is multicomponent, then reliability of each component should be considered.

In simple words, the *reliability* of a product is the probability that the product will function within specified limits for at least a specified period of time under specified environmental conditions. These components may be installed in parallel, series, or a mixture of both. A system and each of its components may be in *functioning*, or *partially functioning*, or *failed* state.

If we assume a system with n independent components that are connected in series and only with two states, namely *function* and *failed*, then the probability of the system in series, denoted by R_S, is the product of reliabilities of its components, denoted by R_i, $i = 1, 2, \ldots, n$. To increase the reliability of a system, the components may be connected in parallel. In such a case, the system fails only if all components fail. Hence, if we denote the probability of failure of each component by $F_i = 1 - R_i$, then the probability of the system failure, denoted by F_P, will be the product of F_i's, that is, the reliability of the system in parallel, denoted by R_P, is $R_P = 1 - F_P$.

Suppose that each component x_i, $i = 1, 2, \ldots, n$, of the system has only two states, namely *functioning*, denoted by 1, and *nonfunctioning* (*failed*), denoted by 0. If we assume that the state of the system is determined completely by the states of its components and denote the state of the system by Ψ, then:

$$\Psi = \Psi(\mathbf{x}), \quad \mathbf{x} = (x_1, x_2, \ldots, x_n). \tag{1.13.1}$$

The function $\Psi(\mathbf{x})$ defined by (1.13.1) is called the *structure function* of the system and the number of components, n, is called the *order of the system*. A *series structure* (*n out of n*) functions if and only if each component functions, while a *parallel structure* (**1** *out of n*) functions if and only if at least one component functions. Thus, for these two cases, structure functions are given by:

$$\Psi(\mathbf{x}) = \prod_{i=1}^{n} x_i = \min(x_1, x_2, \ldots, x_n), \text{ series structure.} \tag{1.13.2}$$

and

$$\Psi(\mathbf{x}) = \coprod_{i=1}^{n} x_i \equiv 1 - \prod_{i=1}^{n} (1 - x_i) = \max(x_1, x_2, \ldots, x_n), \text{ parallel structure.} \tag{1.13.3}$$

A *k-out-of-n structure* functions if and only if k out of n components function. In this case, the structure function is:

$$\Psi(\mathbf{x}) = \begin{cases} 1, & \text{if} \quad \sum_{i=1}^{n} x_i \geq k, \\ 0, & \text{if} \quad \sum_{i=1}^{n} x_i < k. \end{cases} \tag{1.13.4}$$

Relation (1.13.4) is equivalent to:

$$\Psi(\mathbf{x}) = \prod_{i=1}^{k} x_i, \; k\text{-out-of-}n \text{ structure.} \tag{1.13.5}$$

As mentioned earlier, the probability R that the system functions is called the *reliability of the system*, that is,

$$R \equiv P\{\Psi(\mathbf{X}) = 1\}, \mathbf{X} = (X_1, X_2, \ldots, X_n). \tag{1.13.6}$$

Thus,

$$E[\Psi(\mathbf{X})] = R. \tag{1.13.7}$$

We define $\mathbf{r}$ as the random vector, $\mathbf{r} = (r_1, r_2, \ldots, r_n)$. Based on the independence assumption, the system reliability R will be a function of $\mathbf{r}$, that is,

$$R = R(\mathbf{r}). \tag{1.13.8}$$

Let a random variable, say T, represent the lifetime of a component or a system with $f_T(t)$ and $F_T(t)$ as the pdf and cdf, respectively. The *reliability function*, at a time t, denoted by $R(t)$, is the probability that the component or the system is still functioning at time t, that is,

$$R(t) = P(T > t). \tag{1.13.9}$$

From (1.13.9), it is clear that:

$$R(t) = P(T > t) = 1 - P(T \leq) = 1 - F_T(t). \tag{1.13.10}$$

Hence,

$$R'(t) = -f_T(t). \tag{1.13.11}$$

Also, the *mean functioning* time or *survival* is:

$$E(T) = \int_0^\infty f_T(t)\,dt = \int_0^\infty R_T(t)\,dt. \tag{1.13.12}$$

The *failure rate function* (or *force of mortality*), denoted by $r(t)$, is defined by:

$$r(t) = f_T(x \mid T > t)_{x=t}. \tag{1.13.13}$$

The function $r(t)$ is an increasing function of t, that is, the older the unit is, the better the chance of failure within a short interval of length h, namely $r(t)h$. Thus, (1.13.13) is equivalent to:

$$r(t) = f_T(t \mid T > t) = \frac{-R'(t)}{R(t)}. \tag{1.13.14}$$

In general, the failure rate function and the reliability are related by:

$$R(t) = e^{-\int_0^t r(\tau)d\tau}, \tag{1.13.15}$$

and

$$f_T(t) = r(t)e^{-\int_0^t r(\tau)d\tau}. \tag{1.13.16}$$

As an example, suppose that the time to failure of two units 1 and 2 of a system is exponentially distributed with parameters $\lambda_1 = 2$ and $\lambda_2 = 2.5$, respectively. From (1.13.15), the reliability functions of the units are $R_1(t) = e^{-0.5t}$ and $R_2(t) = e^{-0.4t}$, respectively. Assume that the units fail independently. Thus, the reliability functions for the series and parallel systems, respectively, are:

$$R_S = \prod_{i=1}^{n} R_i(t) = \left(e^{-0.5t}\right)\left(e^{-0.4t}\right) = e^{-0.9t},$$

and

$$\begin{aligned} R_P &= \prod_{i=1}^{n} R_i(t) = 1 - \left(e^{-0.5t}\right)\left(e^{-0.4t}\right) = 1 - \left(e^{-0.4t} - e^{-0.5t} + e^{-0.9t}\right) \\ &= e^{-0.4t} + e^{-0.5t} - e^{-0.9t}. \end{aligned}$$

For instance, if a unit of time is 2000 hours, then the probability that the series and parallel systems function for more than a unit of time each is $R_S(1) = e^{-0.9} = 0.4066$ and $R_P = e^{-0.4t} + e^{-0.5t} - e^{-0.9t} = 0.8703$, respectively.

The estimation of the reliability has become a concern for many quality control professionals and statisticians. When Y represents the random value of a stress (or supply) that a device (or a component) will be subjected to in service, X represents the strength (or demand) that varies from product to product in the population of devices. The device fails at the instant that the stress applied to it exceeds the strength and functions successfully whenever $X > Y$. The reliability (or the measure of reliability) R is then defined as $P(Y < X)$, that is, the probability that a randomly selected device functions successfully.

Consider a system with only one component. Suppose the random variable X that represents the *strength* takes place in the lifetime, T, of a device. Suppose also that Y that represents the random value of a *stress* takes place in the time of failure. The reliability, then, that a randomly selected device survives under the stress exerted is the probability that the device functions successfully, that is,

$$R = P(Y < X). \tag{1.13.17}$$

The Weibull distribution is commonly used as a life-lengths model and in the study of breaking strengths of materials. The value of R has been calculated in the literature under the assumed distributions and using different methods of estimations such as MLE, shrinkage estimation procedures, and method of moments.

As an example, for the random variables X and Y as strength and stress, respectively, we choose a random sample for each, say, $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$. To conduct a simulation, due to the involved and tedious calculations of the MLE of α, we choose $\alpha = 2, 3$, and 4 when the ratio of β_1/β_2 is 1, 1/2, and 2, with β_1 and β_2 as parameters for X and Y, respectively. We choose the ratio of the scale parameters to compare the effect measures of the reliability. We also choose $m = n$, and $m = 5(1)10$, (5–10 with increment 1). For instance, based on 1000 runs, estimated values of the different estimators can be calculated.

EXERCISES

1.1. Suppose we want to seat two persons on two chairs.

a. What are the possible outcomes?

b. List all possible simple events.

1.2. Three pairs of "before" and "after" pictures of three different people are given. Holding the "before" pictures and randomly matching the "after" pictures, what is the probability of:

a. all matching?

b. none matching?

1.3. Suppose an election is to choose a set of 4 persons from a group of 10 people without replacement. What is the number of ways to choose the set?

1.4. Suppose a repairman has a diagnostic instrument which can identify problems with a machine three out of four times. Suppose he runs the diagnostic eight times. What is the probability that he identifies the problem

a. less than four times?

b. exactly four times?

1.5. A "fair" coin is tossed three times. What is the probability that two heads turn up?

1.6. A random number generator generates numbers at random from the unit interval [0, 1]. Find the following:

a. $P(A)$, where $A = (1/2, 2/3]$.

b. $P(B)$, where $B = [0,1/2)$.

c. the partition generated by A and B.

d. the event algebra.

1.7. For mutually exclusive events $A_1, A_2, \ldots, A_n$ from a probability space $(\Omega, \mathcal{B}, P)$, show that: $P(\bigcup_i A_i) = \sum_i P(A_i)$.

1.8. If A and B are events and A is a subset of B, prove that $P(B - A) = P(B) - P(A)$ and $P(A) \leq P(B)$.

1.9. If A and B are two events, prove that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

1.10. Manufactured articles in a manufacturing company are required to pass two inspections by two inspectors. Experience shows that one inspector will miss 5% of the defective articles, whereas the second inspector will miss 4% of them. If good articles always pass inspection and if 10% of the articles turned out in manufacturing process are defective, what percentage of the articles that passes both inspections will be defective?

1.11. Suppose an institution of higher learning has equal numbers of male and female students, The chance of male students majoring in science, technology, engineering and mathematics (STEM) is1/5 and for female students this probability is 1/25. A student is chosen at random; what is the probability that:

a. the chosen student will be a male majoring in STEM?

b. the chosen student will be a STEM major?

c. a science student selected at random will be a male student?

1.12. Suppose two technicians T_1 and T_2 independently inspect a unit for a problem. The probability that technician T_1 finds the problem is 1/3 and technician T_2 is 1/6. What is the probability that:

a. both T_1 and T_2 find the problem?

b. at least one will find the problem?

c. neither will find the problem?

d. only T_2 finds the problem?

1.13. A gambler has three coins in his pocket, two of which are fair and one is two headed. He selects a coin. If he tosses the selected coin

a. and it turns up a head, what is the probability that the coin is a fair?

b. five times, what is the probability that he gets five heads?

c. six times, what is the probability of getting five heads followed by a tail?

1.14. Suppose a polling institution finds the approval rating of the handling of foreign policy by the president of the United States during the last 4 years is 23%. Ten Americans are chosen randomly and each is asked about the president's handling of foreign policy. What is the probability that there are at least 2 among the 10 who approve of the handling of foreign policy by the President?

1.15. Suppose that telephone calls enter the department of mathematics' switchboard, on the average, 2 every 10 minutes. Arriving calls are known to have a Poisson pmf with parameter 3. What is the probability of less than 7 calls in a 20-minute period?

1.16. A material is known for 20% of breakage under stress. Five different samples of such materials are tested. What is the probability that:

a. the second sample resists the stress?

b. the second and third samples resist the stress?

1.17. Show that if a random sample of size n, $X_1, X_2, \ldots, X_n$, is drawn from a finite population without replacement, then the random variables $X_1, X_2, \ldots, X_n$ are not mutually independent.

1.18. In a game, a fair die is rolled. To win or lose dollar amounts equal to the number that shows up depends on whether an even or odd number turns up, respectively.

a. What is the expected win or loss in the game?

b. What is the probability if the order of winning and losing is changed?

1.19. Suppose that X is a discrete random variable whose values are 0, 1, 2, 3, 4, and 5, with probabilities 0.2, 0.3, 0.1, 0.1, 0.2, and 0.1, respectively. Find the mean and standard deviation of X.

1.20. Let X be the number of modules with programming errors in a piece of computer software. Let Y be the number of days it takes to debug the software. Suppose X and Y have the following joint probability mass function:

		X				
	p_X	0	1	2	3	4
Y	0	0.20	0.08	0.03	0.02	0.01
	1	0	0.06	0.09	0.04	0.01
	2	0	0.04	0.09	0.06	0.02
	3	0	0.02	0.06	0.04	0.03
	4	0	0	0.03	0.02	0.02
	5	0	0	0	0.02	0.01

Find the following:

a. $E(XY)$

b. The marginal probability mass function for each X and Y

c. The $Cov(X, Y)$

1.21. Let X denote the number of hotdogs and Y the number of sodas consumed by an individual at a game. Suppose X and Y have the following joint probability mass function:

		X			
	$p_{X,Y}$	0	1	2	3
Y	0	0.06	0.15	0.06	0.03
	1	0.04	0.20	0.12	0.04
	2	0.02	0.08	0.06	0.04
	3	0.01	0.03	0.04	0.02

a. Find $E(X)$, $STD(X)$, $E(Y)$, and $STD(Y)$.

b. Find $Cov(X, Y)$ and $Corr(X, Y)$.

c. If hotdogs cost \$4.00 each and soda cost \$2.50 a can, find an individual's expected value and standard deviation of total costs for sodas and hotdogs at a game.

1.22. A number is to be chosen at random from the interval $[0, 1]$. What is the probability that the number is less than 2/5?

1.23. Let X is a standard logistic random variable. Show that

$$Y = \frac{1}{1+e^{-X}}, \quad -\infty < X < \infty,$$

has a uniform distribution with mean 0 and variance 1.

1.24. Show that for n events $A_1, A_2, \ldots, A_n$, we have:

$$P(A_1A_2\cdots A_n) = P(A_1)\times P(A_2|A_1)\times P(A_3|A_1A_2)\times\cdots\times P(A_n|A_1A_2\cdots A_{n-1}).$$

1.25. Show that if events A and B are independent and $P(B) > 0$, then $P(A \mid B) = P(A)$.

1.26. Show the following properties of random variables:

a. if X and Y are two discrete random variables, then $X \pm Y$ and XY are also random variables, and

b. if $\{Y = 0\}$ is empty, then X/Y is also a random variable.

1.27. Let X be a binomial random variable with distribution function B_k. Prove that:

$$B_k = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, 2, \ldots, n,$$

where $q = 1 - p$.

1.28. Show that (1.2.11) is equivalent to (1.2.9).

1.29. Let X be a binomial random variable with distribution function B_k and $\lambda = np$ be fixed. Show that:

$$B_k = \lim_{\substack{n\to\infty,\\ p\to 0}} = \frac{\lambda^k e^{-k}}{k!}, \quad k = 0, 1, 2, \ldots.$$

1.30. Prove:

a. The expected value of the indicator function $I_A(\omega)$ defined in (1.2.3) is $P(A)$, that is, $E(I_A) = P(A)$.

b. If c is a constant, then $E(c) = c$.

c. If c, c_1, and c_2 are constants and X and Y are two random variables, then $E(cX) = cE(X)$ and $E(c_1X + c_2Y) = c_1E(X) + c_2E(Y)$.

d. If $X_1, X_2, \ldots, X_n$ are n random variables, then $E(X_1 + X_2 + \ldots + X_n) = E(X_1) + E(X_1) + \ldots + E(X_n)$.

e. Let X_1 and X_2 be two independent random variables with marginal mass (density) functions p_{x1} and p_{x_2}, respectively. Then, if $E(X_1)$ and $E(X_2)$ exist, we will have $E(X_1X_2) = E(X_1)E(X_2)$.

f. For a finite number of random variables, that is, if $X_1, X_2, \ldots, X_n$ are n independent random variables, then:

$$E(X_1X_2 \cdots X_n) = E(X_1)E(X_2) \cdots E(X_n).$$

1.31. Show that the variance of a binomial random variable X with parameters n and p is $Var(X) = np(1 - p)$.

1.32. Verify that $F_X(-\infty) = 0$ and $F_X(\infty) = 0$.

1.33. Show that $f_X(x; \mu, \alpha)$ given by:

$$f_X(x) = \begin{cases} \mu e^{-\mu x}, & x \geq 0, \\ 0, & \text{elsewhere,} \end{cases}$$

indeed, defines a probability density function.

1.34. If n is a natural number, show that $\Gamma(n) = (n-1)!$, $n = 1, 2, \ldots$, where $n!$ is defined by $n! = n(n-1)(n-2)\ldots(2)(1)$.

1.35. Show that $0! = 1$.

1.36. Show that $\int_0^\infty e^{-x^2} dx = \sqrt{2\pi}$.

1.37. Show that $\Gamma(1/2) = \sqrt{\pi}\Gamma(1/2) = \sqrt{\pi}$.

1.38. Show the following properties of X^2 (chi-square) random variable: mean = r, and variance = $2r$.

1.39. Show that $f(x; \mu, \sigma^2)$ given by:

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

is indeed a pdf.

1.40. Show that $\phi(z)$ given by:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad -\infty < z < \infty$$

is indeed a pdf.

1.41. Show that the mean and variance of a lognormal random variable X are, respectively,

$$E(X) = e^{\mu+\sigma^2/2} \quad \text{and} \quad Var(X) = \left(e^{\sigma^2} - 1\right) e^{2\mu+\sigma^2}.$$

1.42. Show that the mean and variance of standard logistic random variable X, respectively, are $E(X) = a$ and $Var(X) = 1/3(\pi^2 b^2)$.

1.43. Show that the mean and variance of the Weibull random variable, respectively, are:

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

1.44. Show that the relation (1.7.4) represents a probability mass function.

1.45. Prove Theorem 1.7.2, the law of total probability.

1.46. Prove the following properties of covariance:

a. $Cov(X, X) = Var(X)$.

b. $Cov(X, Y) = E[XY] - \mu_X \mu_Y$.

c. $Cov(X, Y) = Cov(Y, X)$.

d. If c is a real number, then $Cov(cX, Y) = cCov(Y, X)$.

e. For two random variables X and Y we have:

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X, Y).$$

1.47. Prove the following properties of sample mean $E(\bar{X}) = \mu$ and $Var(\bar{X}) = \sigma^2/n$.

1.48. Show that the random variable

$$W = \frac{X - \mu}{\sigma/\sqrt{n}}, \quad n = 1, 2, \ldots.$$

is a standard normal random variable for each positive integer n.

1.49. Show that the second moment of S^2 is:

$$E\left[\left(S^2\right)^2\right] = \frac{\mu_4}{n} + \frac{(n-1)^2 + 2}{n(n-1)}\sigma^4,$$

where μ_4 is the fourth central moment of X.

1.50. Show that:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

1.51. Use derivatives of the natural logarithm of $L(\theta)$ rather than ln $L(\theta)$ ln $L(\theta)$ to show that the extrema of ln $L(\theta)$ and ln $L(\theta)$ coincide.

1.52. Is it possible for the standard deviation to be 0? If so, give an example; if not, explain why not.

1.53. Let X be a continuous random variable with pdf $f(x) = 1 - |x - 1|, 0 \le x \le 2$.

a. Find the 20th and 95th percentiles and show them on the graph of cdf of X.

b. Find the IQR.

1.54. Suppose a random variable X represents a large population consisting of three measurements 0, 3, and 12 with the following distribution:

X	0	3	12
P_X	1/3	1/3	1/3

a. Write every possible sample of size 3.

b. Assuming equiprobable property for every possible sample of size 3, what is the probability of each?

c. Find the sample mean, $\bar{X}$, for each possible sample.

d. Find the sample median, M_m, for each possible sample.

e. Find the sampling distribution of the sample mean, $\bar{X}$.

1.55. For Exercise 1.54, show that $\bar{X}$ is an unbiased estimator of the population parameterμ.

1.56. Let $X_1, X_2, \ldots, X_n$ represent a random sample with its n independent observed values $x_1, x_2, \ldots, x_n$ from a normal random variable with mean μ and variance σ^2. Find the MLE of $\theta(\mu, \sigma^2)$.

1.57. Suppose $X_1, X_2, \ldots, X_n$ is a random sample of size n with pdf $f(x; \lambda) = (1/\lambda)e^{-x/\lambda}$, $x \geq 0$, $\lambda > 0$. Find the MLE for the parameter λ.

1.58. Consider the pdf of a one-parameter *Cauchy distribution* as:

$$f(x;\lambda) = \frac{\lambda}{\pi\left(x^2 + \lambda^2\right)}, \lambda > 0.$$

Find the maximum likelihood estimate of the parameter λ.

1.59. Find $z_{\alpha/2}$ for each of the following values of α:

a. 0.10

b. 0.05

c. 0.01

1.60. A random sample of 80 shoppers at an automotive part store showed that they spent an average of \$20.5 with variance of \$39.2. Find a 95% confidence interval for the average amount spent by a shopper at the store.

1.61. The time it takes for a manufacturer to assemble an electronic instrument is a normal random variable with mean of 1.2 hours and a variance of 0.04 hour. To reduce the assembly time, the manufacturer implements a new procedure. A random sample of size 35 is then taken and it shows the mean assembly time as 0.9 hour. Assuming the variance remains unchanged, form a 95% confidence interval for the mean assembly time under the new procedure.

1.62. Find the value of $t_{n-1,\alpha}$ to construct a two-sided confidence interval of the given level with the indicated sample size:

a. 95% level, $n = 5$.

b. 99% level, $n = 29$.

1.63. Suppose that at a company it is known that over the past few years, employees' sick days averaged 5.4 days per year. To reduce this number, the company introduces telecommuting (allowing employees to work at home on their computers). After implementing the new policy, the Human Resource Department chooses a random sample of 50 employees at the end of the year and found an average of 4.5 sick days with a standard deviation of 2.7 days. Let μ be the mean sick days of all employees of the company. Find the p-value for testing hypothesis H_0: $\mu \geq 5.4$ versus H_1: $\mu < 5.4$.

1.64. In a comparison of the effectiveness of online leaning with the traditional in-classroom instruction, 12 students enrolled in a business course online and 14 enrolled in a course with traditional in-classroom instruction. The final exam scores were as follows:

Classroom	80	77	74	64	71	80	68	85	83	59	55	75	81	81
Online	64	66	74	69	75	72	77	83	77	91	85	88		

Does the mean score differ between the two type of course?