GLOBAL EDITION

Probability & Statistics
for Engineers & Scientists

NINTH EDITION

Walpole • Myers • Myers • Ye

ALWAYS LEARNING

PEARSON

Probability & Statistics for Engineers & Scientists

NINTH EDITION
GLOBAL EDITION

Ronald E. Walpole
Roanoke College

Raymond H. Myers
Virginia Tech

Sharon L. Myers
Radford University

Keying Ye
University of Texas at San Antonio

PEARSON

Boston Columbus Hoboken Indianapolis New York San Francisco
Amsterdam Cape Town Dubai London Madrid Milan Munich Paris Montréal Toronto
Delhi Mexico City São Paulo Sydney Hong Kong Seoul Singapore Taipei Tokyo

7.5 If the variable X has the probability distribution of

$$f(x) = \begin{cases} 1, & 0 < x < 1 \\ 0 & \text{elsewhere,} \end{cases}$$

Show that the random variable $Y = -2 \log_e X$ has an exponential distribution with $\lambda = \frac{1}{2}$.

7.6 Given the random variable X with probability distribution

$$f(x) = \begin{cases} e^{-x}, & 0 < x < \infty \\ 0, & \text{elsewhere,} \end{cases}$$

find the probability distribution function of $Y = -3X + 5$.

7.7 The speed of a molecule in a uniform gas at equilibrium is a random variable V whose probability distribution is given by

$$f(v) = \begin{cases} kv^2 e^{-bv^2}, & v > 0, \\ 0, & \text{elsewhere,} \end{cases}$$

where k is an appropriate constant and b depends on the absolute temperature and mass of the molecule. Find the probability distribution of the kinetic energy of the molecule W, where $W = mV^2/2$.

7.8 A dealer's profit, in units of \$5000, on a new automobile is given by $Y = X^2$, where X is a random variable having the density function

$$f(x) = \begin{cases} 2(1-x), & 0 < x < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

(a) Find the probability density function of the random variable Y.

(b) Using the density function of Y, find the probability that the profit on the next new automobile sold by this dealership will be less than \$500.

7.9 The hospital period, in days, for patients following treatment for a certain type of kidney disorder is a random variable $Y = X + 4$, where X has the density function

$$f(x) = \begin{cases} \frac{32}{(x+4)^3}, & x > 0, \\ 0, & \text{elsewhere.} \end{cases}$$

(a) Find the probability density function of the random variable Y.

(b) Using the density function of Y, find the probability that the hospital period for a patient following this treatment will exceed 8 days.

7.10 The random variables X and Y, representing the weights of creams and toffees, respectively, in 1-kilogram boxes of chocolates containing a mixture of creams, toffees, and cordials, have the joint density function

$$f(x,y) = \begin{cases} 24xy, & 0 \le x \le 1,\ 0 \le y \le 1, x + y \le 1, \\ 0, & \text{elsewhere.} \end{cases}$$

(a) Find the probability density function of the random variable $Z = X + Y$.

(b) Using the density function of Z, find the probability that, in a given box, the sum of the weights of creams and toffees accounts for at least 1/2 but less than 3/4 of the total weight.

7.11 The amount of kerosene, in thousands of liters, in a tank at the beginning of any day is a random amount Y from which a random amount X is sold during that day. Assume that the joint density function of these variables is given by

$$f(x,y) = \begin{cases} 2, & 0 < x < y,\ 0 < y < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the probability density function for the amount of kerosene left in the tank at the end of the day.

7.12 Let X_1 and X_2 be independent random variables each having the probability distribution

$$f(x) = \begin{cases} e^{-x}, & x > 0, \\ 0, & \text{elsewhere.} \end{cases}$$

Show that the random variables Y_1 and Y_2 are independent when $Y_1 = X_1 + X_2$ and $Y_2 = X_1/(X_1 + X_2)$.

7.13 A current of I amperes flowing through a resistance of R ohms varies according to the probability distribution

$$f(i) = \begin{cases} 6i(1-i), & 0 < i < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

If the resistance varies independently of the current according to the probability distribution

$$g(r) = \begin{cases} 2r, & 0 < r < 1, \\ 0, & \text{elsewhere,} \end{cases}$$

find the probability distribution for the power $W = I^2R$ watts.

7.14 Let X be a random variable with probability distribution

$$f(x) = \begin{cases} \frac{1+x}{2}, & -1 < x < 1, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the probability distribution of the random variable $Y = X^2$.

7.15 Let X have the probability distribution

$$f(x) = \begin{cases} \frac{2(x+1)}{9}, & -1 < x < 2, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the probability distribution of the random variable $Y = X^2$.

7.16 Show that the rth moment about the origin of the gamma distribution is

$$\mu_r' = \frac{\beta^r \Gamma(\alpha + r)}{\Gamma(\alpha)}.$$

[*Hint*: Substitute $y = x/\beta$ in the integral defining μ_r' and then use the gamma function to evaluate the integral.]

7.17 A random variable X has the discrete uniform distribution

$$f(x; k) = \begin{cases} \frac{1}{k}, & x = 1, 2, \ldots, k, \\ 0, & \text{elsewhere.} \end{cases}$$

Show that the moment-generating function of X is

$$M_X(t) = \frac{e^t(1 - e^{kt})}{k(1 - e^t)}.$$

7.18 A random variable X has the geometric distribution $g(x; p) = pq^{x-1}$ for $x = 1, 2, 3, \ldots$. Show that the moment-generating function of X is

$$M_X(t) = \frac{pe^t}{1 - qe^t}, \qquad t < \ln q,$$

and then use $M_X(t)$ to find the mean and variance of the geometric distribution.

7.19 A random variable X has the Poisson distribution $p(x; \mu) = e^{-\mu}\mu^x/x!$ for $x = 0, 1, 2, \ldots$. Show that the moment-generating function of X is

$$M_X(t) = e^{\mu(e^t - 1)}.$$

Using $M_X(t)$, find the mean and variance of the Poisson distribution.

7.20 The moment-generating function of a certain Poisson random variable X is given by

$$M_X(t) = e^{9(e^t - 1)}.$$

Find $P(\mu - \sigma \leq X \leq \mu + \sigma)$.

7.21 Show that the moment-generating function of the random variable X having a chi-squared distribution with v degrees of freedom is

$$M_X(t) = (1 - 2t)^{-v/2}.$$

7.22 Using the moment-generating function of Exercise 7.21, show that the mean and variance of the chi-squared distribution with v degrees of freedom are, respectively, v and $2v$.

7.23 Both X and Y independently follow a geometric distribution with a probability mass function of $P(X = r) = P(Y = r) = q^r p, r = 0, 1, 2, \ldots$, where, p and q are positive numbers such that $p + q = 1$. Find
(a) the probability distribution of $U = X + Y$;
(b) the conditional distribution of $X/(X + Y) = u$.

7.24 By expanding e^{tx} in a Maclaurin series and integrating term by term, show that

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} e^{tx} f(x)\, dx \\ &= 1 + \mu t + \mu_2' \frac{t^2}{2!} + \cdots + \mu_r' \frac{t^r}{r!} + \cdots . \end{aligned}$$

Chapter 8

Fundamental Sampling Distributions and Data Descriptions

8.1 Random Sampling

The outcome of a statistical experiment may be recorded either as a numerical value or as a descriptive representation. When a pair of dice is tossed and the total is the outcome of interest, we record a numerical value. However, if the students of a certain school are given blood tests and the type of blood is of interest, then a descriptive representation might be more useful. A person's blood can be classified in 8 ways: AB, A, B, or O, each with a plus or minus sign, depending on the presence or absence of the Rh antigen.

In this chapter, we focus on sampling from distributions or populations and study such important quantities as the *sample mean* and *sample variance*, which will be of vital importance in future chapters. In addition, we attempt to give the reader an introduction to the role that the sample mean and variance will play in statistical inference in later chapters. The use of modern high-speed computers allows the scientist or engineer to greatly enhance his or her use of formal statistical inference with graphical techniques. Much of the time, formal inference appears quite dry and perhaps even abstract to the practitioner or to the manager who wishes to let statistical analysis be a guide to decision-making.

Populations and Samples

We begin this section by discussing the notions of *populations* and *samples*. Both are mentioned in a broad fashion in Chapter 1. However, much more needs to be presented about them here, particularly in the context of the concept of random variables. The totality of observations with which we are concerned, whether their number be finite or infinite, constitutes what we call a **population**. There was a time when the word *population* referred to observations obtained from statistical studies about people. Today, statisticians use the term to refer to observations relevant to anything of interest, whether it be groups of people, animals, or all possible outcomes from some complicated biological or engineering system.

Definition 8.1: A **population** consists of the totality of the observations with which we are concerned.

The number of observations in the population is defined to be the size of the population. If there are 600 students in the school whom we classified according to blood type, we say that we have a population of size 600. The numbers on the cards in a deck, the heights of residents in a certain city, and the lengths of fish in a particular lake are examples of populations with finite size. In each case, the total number of observations is a finite number. The observations obtained by measuring the atmospheric pressure every day, from the past on into the future, or all measurements of the depth of a lake, from any conceivable position, are examples of populations whose sizes are infinite. Some finite populations are so large that in theory we assume them to be infinite. This is true in the case of the population of lifetimes of a certain type of storage battery being manufactured for mass distribution throughout the country.

Each observation in a population is a value of a random variable X having some probability distribution $f(x)$. If one is inspecting items coming off an assembly line for defects, then each observation in the population might be a value 0 or 1 of the Bernoulli random variable X with probability distribution

$$b(x; 1, p) = p^x q^{1-x}, \qquad x = 0, 1$$

where 0 indicates a nondefective item and 1 indicates a defective item. Of course, it is assumed that p, the probability of any item being defective, remains constant from trial to trial. In the blood-type experiment, the random variable X represents the type of blood and is assumed to take on values from 1 to 8. Each student is given one of the values of the discrete random variable. The lives of the storage batteries are values assumed by a continuous random variable having perhaps a normal distribution. When we refer hereafter to a "binomial population," a "normal population," or, in general, the "population $f(x)$," we shall mean a population whose observations are values of a random variable having a binomial distribution, a normal distribution, or the probability distribution $f(x)$. Hence, the mean and variance of a random variable or probability distribution are also referred to as the mean and variance of the corresponding population.

In the field of statistical inference, statisticians are interested in arriving at conclusions concerning a population when it is impossible or impractical to observe the entire set of observations that make up the population. For example, in attempting to determine the average length of life of a certain brand of light bulb, it would be impossible to test all such bulbs if we are to have any left to sell. Exorbitant costs can also be a prohibitive factor in studying an entire population. Therefore, we must depend on a subset of observations from the population to help us make inferences concerning that same population. This brings us to consider the notion of sampling.

Definition 8.2: A **sample** is a subset of a population.

If our inferences from the sample to the population are to be valid, we must obtain samples that are representative of the population. All too often we are tempted to choose a sample by selecting the most convenient members of the population. Such a procedure may lead to erroneous inferences concerning the population. Any sampling procedure that produces inferences that consistently overestimate or consistently underestimate some characteristic of the population is said to be **biased**. To eliminate any possibility of bias in the sampling procedure, it is desirable to choose a **random sample** in the sense that the observations are made independently and at random.

In selecting a random sample of size n from a population $f(x)$, let us define the random variable X_i, $i = 1, 2, \ldots, n$, to represent the ith measurement or sample value that we observe. The random variables $X_1, X_2, \ldots, X_n$ will then constitute a random sample from the population $f(x)$ with numerical values $x_1, x_2, \ldots, x_n$ if the measurements are obtained by repeating the experiment n independent times under essentially the same conditions. Because of the identical conditions under which the elements of the sample are selected, it is reasonable to assume that the n random variables $X_1, X_2, \ldots, X_n$ are independent and that each has the same probability distribution $f(x)$. That is, the probability distributions of $X_1, X_2, \ldots, X_n$ are, respectively, $f(x_1), f(x_2), \ldots, f(x_n)$, and their joint probability distribution is $f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n)$. The concept of a random sample is described formally by the following definition.

Definition 8.3: Let $X_1, X_2, \ldots, X_n$ be n independent random variables, each having the same probability distribution $f(x)$. Define $X_1, X_2, \ldots, X_n$ to be a **random sample** of size n from the population $f(x)$ and write its joint probability distribution as

$$f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n).$$

If one makes a random selection of $n = 8$ storage batteries from a manufacturing process that has maintained the same specification throughout and records the length of life for each battery, with the first measurement x_1 being a value of X_1, the second measurement x_2 a value of X_2, and so forth, then $x_1, x_2, \ldots, x_8$ are the values of the random sample $X_1, X_2, \ldots, X_8$. If we assume the population of battery lives to be normal, the possible values of any X_i, $i = 1, 2, \ldots, 8$, will be precisely the same as those in the original population, and hence X_i has the same identical normal distribution as X.

8.2 Some Important Statistics

Our main purpose in selecting random samples is to elicit information about the unknown population parameters. Suppose, for example, that we wish to arrive at a conclusion concerning the proportion of coffee-drinkers in the United States who prefer a certain brand of coffee. It would be impossible to question every coffee-drinking American in order to compute the value of the parameter p representing the population proportion. Instead, a large random sample is selected and the proportion $\hat{p}$ of people in this sample favoring the brand of coffee in question is calculated. The value $\hat{p}$ is now used to make an inference concerning the true proportion p.

Now, $\hat{p}$ is a function of the observed values in the random sample; since many random samples are possible from the same population, we would expect $\hat{p}$ to vary somewhat from sample to sample. That is, $\hat{p}$ is a value of a random variable that we represent by P. Such a random variable is called a **statistic**.

Definition 8.4: Any function of the random variables constituting a random sample is called a **statistic**.

Location Measures of a Sample: The Sample Mean, Median, and Mode

In Chapter 4 we introduced the two parameters μ and σ^2, which measure the center of location and the variability of a probability distribution. These are constant population parameters and are in no way affected or influenced by the observations of a random sample. We shall, however, define some important statistics that describe corresponding measures of a random sample. The most commonly used statistics for measuring the center of a set of data, arranged in order of magnitude, are the **mean**, **median**, and **mode**. Although the first two of these statistics were defined in Chapter 1, we repeat the definitions here. Let $X_1, X_2, \ldots, X_n$ represent n random variables.

(a) Sample mean:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Note that the statistic $\bar{X}$ assumes the value $\bar{x} = \frac{1}{n}\sum\limits_{i=1}^{n} x_i$ when X_1 assumes the value x_1, X_2 assumes the value x_2, and so forth. The term *sample mean* is applied to both the statistic $\bar{X}$ and its computed value $\bar{x}$.

(b) Sample median:

$$\tilde{x} = \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ is odd,} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}), & \text{if } n \text{ is even.} \end{cases}$$

The sample median is also a location measure that shows the middle value of the sample. Examples for both the sample mean and the sample median can be found in Section 1.3. The sample mode is defined as follows.

(c) The sample mode is the value of the sample that occurs most often.

Example 8.1: Suppose a data set consists of the following observations:

$$0.32\ 0.53\ 0.28\ 0.37\ 0.47\ 0.43\ 0.36\ 0.42\ 0.38\ 0.43.$$

The sample mode is 0.43, since this value occurs more than any other value.

As we suggested in Chapter 1, a measure of location or central tendency in a sample does not by itself give a clear indication of the nature of the sample. Thus, a measure of variability in the sample must also be considered.