Chapter 15

SIMULATION, BOOTSTRAP STATISTICAL METHODS, AND PERMUTATION TESTS

15.1 INTRODUCTION

In this chapter we introduce two powerful modern statistical techniques: bootstrap statistical methods and permutation tests. Both are nonparametric procedures in the sense that they make no specific assumptions about the form of any underlying probability distributions. Bootstrap methods enable us to measure the efficacy of an estimator of a parameter, while permutation tests yield new ways to test certain statistical hypotheses. Both, however, require a large amount of computation in their implementation. The most efficient and effective way of doing the needed computation uses simulation, the third topic of this chapter.

In Section 15.2 we introduce random numbers, which are the keys to a simulation. We show how random numbers can be used to generate random permutations and random subsets. In Section 15.2.1 we present the Monte Carlo simulation method for approximating expectations. In Section 15.3 we introduce the method of bootstrap statistics and show how the needed analysis can be done by applying the Monte Carlos simulation method. In Section 15.4 we discuss permutation tests, which are nonparametric tests for determining whether a sequence of data comes from a single population distribution. In the remaining sections we return to the study of simulation. In Sections 15.5 and 15.6 we show how random numbers can be used to generate the values of arbitrarily distributed discrete and continuous random variables, and in Section 15.7 we consider the question of when to end a Monte Carlo simulation study.

15.2 RANDOM NUMBERS

The value of a uniform $(0, 1)$ random variable is called a *random number*. Whereas in the past, mechanical devices have often been used to generate random numbers, today we commonly use random number generators to generate a sequence of pseudo random numbers. Such random number generators start with an initial value x_0, called the *seed*, and then recursively determine values by first specifying positive integers a, c, and m and then letting

$$x_{n+1} = (ax_n + c) \ \text{ modulo } m, \ \ n \geq 0$$

where the preceding means that x_{n+1} is the remainder obtained when $ax_n + c$ is divided by m. Thus each x_n is one of the values $0, 1, \ldots, m - 1$, and the quantity x_n/m is taken as the random number. It can be shown that for suitable choices of a, c, and m, the preceding gives rise to a sequence of numbers that looks as if it was generated by observing the values of independent uniform $(0, 1)$ random variables. For this reason we call the numbers x_n/m, $n \geq 1$, *pseudo random numbers*.

EXAMPLE 15.2a If $a = 3$, $c = 7$, $m = 23$, then with $x_0 = 2$

$$\begin{aligned}
x_1 &= 3(2) + 7 && \text{modulo } 23 = 13\\
x_2 &= 3(13) + 7 && \text{modulo } 23 = 0\\
x_3 &= 3(0) + 7 && \text{modulo } 23 = 7\\
x_4 &= 3(7) + 7 && \text{modulo } 23 = 5\\
x_5 &= 3(5) + 7 && \text{modulo } 23 = 22
\end{aligned}$$

and so on. Consequently, using the seed $x_0 = 2$, the pseudo random numbers obtained are $13/23$, $0, 7/23$, $5/23$, $22/23, \ldots$. ■

Most computers have built-in random number generators, and we shall take as our starting point in simulation that we can generate the values of pseudo random numbers; moreover, we will act as if these pseudo random numbers were actually true random numbers. That is, we will act as if the sequence of random numbers were actually a sequence of values of a sample from the uniform $(0, 1)$ distribution.

Random numbers are the key to any simulation study. This is illustrated in our next example, which is concerned with generating a random permutation.

EXAMPLE 15.2b Suppose we want to generate a permutation of the numbers $1, 2, \ldots, n$ in such a manner that all $n!$ possible permutations are equally likely. To accomplish this we can first randomly choose one of the numbers $1, 2, \ldots, n$ and put that number in position n. We can then randomly choose one of the remaining $n-1$ numbers and put that number in position $n-1$, and then randomly choose one of the remaining $n-2$ numbers and put that number in position $n-2$, and so on (where "randomly choose" means that each of the possible choices is equally likely to be made). However, so that we do not have to directly consider exactly which elements remain to be placed, it is convenient and effective to keep the numbers in an ordered list and then randomly choose the position of the number rather than the number itself. That is, starting with any permutation $r_1, r_2, \ldots, r_n$ of the numbers $1, 2, \ldots, n$, randomly choose one of the positions $1, \ldots, n$ and then interchange the number in that position with the number in position n. Then randomly choose one of the positions $1, \ldots, n-1$, and interchange the number in that position with the number in position $n-1$. Then randomly choose one of the positions $1, \ldots, n-2$, and interchange the number in that position with the number in position $n-2$, and so on.

To implement the preceding we need to be able to generate a random variable that is equally likely to take on any of the values $1, \ldots, k$. To accomplish this, let U denote a random number — that is, U is uniformly distributed over $(0, 1)$ — and let $\text{Int}(kU)$ be the integer part of kU — that is, it is the largest integer less than or equal to kU. Then, for $i = 1, \ldots, k$

$$\begin{aligned} P[\text{Int}(kU) + 1 = i] &= P[\text{Int}(kU) = i - 1] \\ &= P(i - 1 \le kU < i) \\ &= P(\tfrac{i-1}{k} \le U < \tfrac{i}{k}) \\ &= 1/k \end{aligned}$$

Thus, $\text{Int}(kU) + 1$ is equally likely to take on any of the values $1, \ldots, k$.

The algorithm for generating a random permutation of the numbers $1, 2, \ldots, n$ can now be written as follows:

1. Let $r_1, r_2, \ldots, r_n$ be any permutation of the numbers $1, 2, \ldots, n$. (For instance, we could have $r_j = j$, $j = 1, \ldots, n$.)
2. Set $k = n$. (The number to be put in position k is to be determined.)
3. Generate a random number U and let $I = \text{Int}(kU) + 1$.
4. Interchange the values of r_I and r_k.
5. Let $k = k - 1$.
6. If $k > 1$, go to Step 3; if $k = 1$, go to step 7.
7. $r_1, \ldots, r_n$ is the desired permutation.

For instance, suppose $n = 4$ and the initial permutation is $1, 2, 3, 4$. If the first value of I — which is equally likely to be any of the numbers $1, 2, 3, 4$ — is 2, then the number in position 2 is interchanged with the number in position 4 to give the new permutation $1, 4, 3, 2$. If the next value of I — which is equally likely to be any of the numbers $1, 2, 3$ — is 3, then the number in position 3 is interchanged with the one in position 3, so the permutation remains $1, 4, 3, 2$. If the final value of I — which is equally likely to be any of the numbers $1, 2$ — is 1, then the number in position 1 is interchanged with the one in position 2 to give the final permutation $4, 1, 3, 2$.

An important property of the preceding algorithm is that it can be used to generate a random subset of size r from the set $\{1, 2, \ldots, n\}$. For $r \le n/2$, just follow the preceding algorithm until the elements in the final r positions (that is, in positions $n, n-1, \ldots, n-r+1$) are specified, and then take the numbers in these positions as the random subset of size r. For $r > n/2$, rather than directly choosing the r numbers to be in the subset, it is quicker to choose the $n-r$ numbers that are not in the subset. So in this case, follow the preceding algorithm until the final $n-r$ positions are filled, and then take the numbers that remain as the random subset of size r. ■

15.2.1 The Monte Carlo Simulation Approach

Suppose we want to compute the expected value of a statistic $h(X_1, X_2, \ldots, X_n)$ when $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables having density function $f(x)$. Using that the joint density function of $X_1, X_2, \ldots, X_n$ is

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

we can write that

$$E[h(X_1, X_2, \ldots, X_n)] = \int\int\cdots\int h(x_1, \ldots, x_n)f(x_1)f(x_2)\cdots f(x_n)\, dx_1\, dx_2 \cdots dx_n$$

The difficulty, however, with the preceding formula is that it is often impossible to analytically compute the preceding multiple integral and also difficult to numerically evaluate it to within a specified accuracy. One approach that remains is to approximate $E[h(X_1, X_2, \ldots, X_n)]$ by a simulation.

To accomplish this approximation, start by generating the values of n independent random variables $X_1^1, X_2^1, \ldots, X_n^1$, each having density function f, and then compute

$$Y_1 = h(X_1^1, X_2^1, \ldots, X_n^1)$$

Now generate the values of a second set of n independent random variables having density function f that are also independent of the first set. Calling this second set of random variables $X_1^2, X_2^2, \ldots, X_n^2$, compute

$$Y_2 = h(X_1^2, X_2^2, \ldots, X_n^2)$$

Continue doing this until you have generated r sets of n independent random variables having density function f, and have computed the corresponding values of Y. In this way, we would have generated values of r independent and identically random variables $Y_i = h(X_1^i, X_2^i, \ldots, X_n^i)$, $i = 1, \ldots, r$. Now, by the strong law of large numbers

$$\lim_{r\to\infty} \frac{Y_1 + \cdots + Y_r}{r} = E[Y_i] = E[h(X_1, X_2, \ldots, X_n)]$$

and so we can use the average of the generated values of the Y_i's as an estimate of $E[h(X_1, X_2, \ldots, X_n)]$. This approximation method is called the *Monte Carlo* simulation approach. Each time we generate a new value of Y we say that a new *simulation run* has been completed.

Of course in order to make use of the preceding approach we need to be able to generate random variables having a specified density function. Although at present we only know how to do this for a uniform random variable — by using a random number generator — this will suffice for the needed computations both in the bootstrap method and in running permutation tests. As a result, the next two sections will present these topics. We will then return to the simulation question of how to generate random variables having arbitrary distributions, as well as how to determine when to end a simulation study, in the final sections of this chapter.

15.3 THE BOOTSTRAP METHOD

Let $X_1, \ldots, X_n$ be a sample from a population having distribution F, and suppose we want to use this sample to estimate a parameter θ of F. For instance, θ could be the common mean or variance of the X_i. Suppose we have an estimator $d = d(X_1, \ldots, X_n)$ of θ and we would like to evaluate how good an estimator of θ it is. One measure of the worth of $d(X_1, \ldots, X_n)$ as an estimator of θ is its mean square error, defined as

$$MSE_F(d) = E_F[(d(X_1, \ldots, X_n) - \theta)^2]$$

That is, $MSE_F(d)$ is the expected square of the distance between the estimator $d(X_1, \ldots, X_n)$ and the parameter θ, where we use the notation MSE_F and E_F to indicate that the expected value is to be computed under the assumption that $X_1, \ldots, X_n$ are independent random variables having distribution function F. How can this quantity be estimated?

EXAMPLE 15.3a If $\theta = E[X_i]$ is the mean of the distribution F, and $d(X_1, \ldots, X_n) = \bar{X}_n = \sum_{i=1}^{n} X_i/n$ is the sample mean of the data values $X_1, \ldots, X_n$, then because

$$E_F[d(X_1, \ldots, X_n)] = E_F[\bar{X}_n] = E_F[X_i] = \theta$$

it follows that

$$MSE_F(\bar{X}_n) = E_F[(\bar{X}_n - \theta)^2] = \mathrm{Var}_F(\bar{X}_n) = \sigma^2/n$$

where $\sigma^2 = \mathrm{Var}_F(X_i)$. Thus, in this case, $MSE_F(\bar{X}_n)$ can be estimated by the quantity S_n^2/n, where

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

is the sample variance of the data values $X_1, \ldots, X_n$, and can be used to estimate the population variance σ^2. ■

Whereas in the preceding example it was easy to estimate the mean square error of the sample mean as an estimator of a population mean, what if we initially wanted to estimate the population variance? That is, what if $\theta = \text{Var}_F(X_i)$. In this case we can use the sample variance as the estimator. However, while it was easy to come up with this estimator $d(X_1, \ldots, X_n) = S_n^2$, it is not easy to see how to estimate its mean square error. One way is to use the approach of bootstrap statistics, which we now present.

To estimate the mean square error of the estimator $d(X_1, \ldots, X_n)$ of the parameter θ, suppose that the data values are $X_i = x_i$, $i = 1, \ldots, n$. For any x, let $F_e(x)$ denote the proportion of the data values that are less than or equal to x. That is,

$$F_e(x) = \frac{\text{number of } i \leq n : x_i \leq x}{n}$$

For instance, if $n = 5$ and $X_1 = 5$, $X_2 = 3$, $X_3 = 9$, $X_4 = 2$, and $X_5 = 6$, then

$$F_e(x) = \begin{cases} 0 & \text{if} \quad x < 2 \\ 1/5 & \text{if} \quad 2 \leq x < 3 \\ 2/5 & \text{if} \quad 3 \leq x < 5 \\ 3/5 & \text{if} \quad 5 \leq x < 6 \\ 4/5 & \text{if} \quad 6 \leq x < 9 \\ 1 & \text{if} \quad x \geq 9 \end{cases}$$

The function $F_e(x)$ is called the *empirical distribution function*. When the values $x_1, \ldots, x_n$ are all distinct, F_e is the distribution function of a random variable X_e that is equally likely to be any of the values $x_1, \ldots, x_n$. That is, if the data values are all distinct, then F_e is the distribution function of the random variable X_e such that

$$P(X_e = x_i) = 1/n, \quad i = 1, \ldots, n$$

When the data values are not all distinct, then F_e is the distribution function of the random variable X_e whose probability of being equal to any specified data value is the number of times that value appears in the data set divided by n. For instance, if $n = 3$ and $x_1 = x_2 = 1, x_3 = 2$ then X_e is a random variable that takes on the value 1 with probability 2/3 and 2 with probability 1/3. With this understanding about the weight put on a distinct value, we will still say that F_e is the distribution function of a random variable that is equally likely to be any of the values $x_1, x_2, \ldots, x_n$.

Now, for any value of x, each of the data values X_i, $i = 1, \ldots, n$, will be less than or equal to x with probability $F(x)$. Hence, by the strong law of large numbers it follows that the proportion of them that are less than or equal to x will, with probability 1, converge to $F(x)$ as n goes to infinity. Thus, for n large, $F_e(x)$ should be close to $F(x)$, indicating that the empirical distribution function F_e can be used as an estimator of the population distribution function F.

Now let θ_e have the same relationship to the distribution F_e as θ has to the distribution F. For instance, if θ is the variance of a random variable X having distribution F, then θ_e is the variance of a random variable X_e having distribution F_e. Now, if F_e is close to F, then it almost always follows that θ_e will be close to θ. (Technically speaking, this will be true provided that θ is a continuous function of the distribution F.) For these reasons we can approximate the mean square error of the estimator $d(X_1, \dots, X_n)$ of θ as follows:

$$MSE_F(d) = E_F[(d(X_1, \dots, X_n) - \theta)^2] \approx E_{F_e}[(d(X_1, \dots, X_n) - \theta_e)^2]$$

where by E_{F_e} we mean that the expectation is to be taken under the assumption that $X_1, \dots, X_n$ are independent random variables, each having distribution function F_e. That is, each of $X_1, \dots, X_n$ is equally likely to be any of the values $x_1, \dots, x_n$.

The quantity

$$MSE_{F_e}(d) = E_{F_e}[(d(X_1, \dots, X_n) - \theta_e)^2]$$

is called the *bootstrap estimate of the mean square error* of $d(X_1, \dots, X_n)$ as an estimator of θ.

Let us now see how well $MSE(F_e)$ estimates $MSE(F)$ in the one case where its use as an estimator is not needed—namely, when estimating the mean of a distribution by the sample mean.

EXAMPLE 15.3b Consider Example 15.3a, where $\bar{X}_n = \sum_{i=1}^{n} X_i/n$ is used as an estimator of the mean of the distribution F. Because X_e puts equal weight on each of the data values $x_1, \dots, x_n$, it follows, when θ_e is the mean of this distribution, that

$$\theta_e = E[X_e] = \sum_{i=1}^{n} x_i P(X_e = x_i) = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}_n$$

Because

$$E_{F_e}[\sum_{i=1}^{n} X_i/n] = E_{F_e}[X] = \theta_e = \bar{x}_n$$

it follows that

$$\begin{aligned} MSE_{F_e}(\bar{X}_n) &= E_{F_e}[(\sum_{i=1}^{n} X_i/n - \bar{x}_n)^2] \\ &= \mathrm{Var}_{F_e}(\sum_{i=1}^{n} X_i/n) \\ &= \tfrac{1}{n}\mathrm{Var}_{F_e}(X) \end{aligned}$$

Now,

$$\begin{aligned}\mathrm{Var}_{F_e}(X) &= E_{F_e}[(X-\bar{x}_n)^2]\\ &= \sum_{i=1}^{n}(x_i-\bar{x}_n)^2 P_{F_e}(X=x_i)\\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2\end{aligned}$$

Thus, we have shown that

$$MSE_{F_e}(\bar{X}_n) = \frac{1}{n^2}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2$$

As the usual estimator of $MSE_F(\bar{X}_n) = \frac{1}{n}\mathrm{Var}_F(X)$ is S_n^2/n, whose observed value is $\frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2$, we see that the bootstrap estimate is almost identical to the usual estimate in this case. ■

As previously noted, if the data values are $X_i = x_i, i = 1, \ldots, n$, then the empirical distribution function F_e puts equal weight $1/n$ on each of the points x_i; consequently, it is usually easy to compute the value of θ_e. To compute the bootstrap estimate of the mean square error of the estimator $d(X_1, \ldots, X_n)$ of θ, we then have to compute

$$MSE_{F_e}(d) = E_{F_e}[(d(X_1, \ldots, X_n) - \theta_e)^2]$$

However, since the preceding expectation is to be computed under the assumption that $X_1, \ldots, X_n$ are all distributed according to F_e, it follows that the vector $(X_1, \ldots, X_n)$ is equally likely to be any of the n^n possible values $(x_{i_1}, x_{i_2}, \ldots, x_{i_n})$, where each i_j is one of the values $1, \ldots, n$. Consequently an exact computation of $MSE_{F_e}(d)$ is prohibitive unless n is small.

It is, however, easy to approximate $MSE_{F_e}(d)$ by a simulation. To do so, we generate n independent random variables $X_1^1, \ldots, X_n^1$ having distribution F_e and use them to compute the value of

$$Y_1 = (d(X_1^1, \ldots, X_n^1) - \theta_e)^2$$

We then repeat this process and generate a second set of n independent random variables $X_1^2, \ldots, X_n^2$ having distribution F_e and use them to compute the value of

$$Y_2 = (d(X_1^2, \ldots, X_n^2) - \theta_e)^2$$

This is then repeated a large number of times, say r, to obtain the values $Y_1, \ldots, Y_r$. The average of these values, $\sum_{i=1}^{r} Y_i/r$, would be the approximation of $MSE_{F_e}(d)$, which would then be used as the estimate of $MSE_F(d)$.

REMARK

It is easy to generate a random variable X having distribution F_e. Just generate a random number U; let $I = \text{Int}(nU) + 1$, so that I is equally likely to be any of the values $1, \ldots, n$; and then set

$$X = x_I$$

EXAMPLE 15.3c Suppose we use the sample variance $S_n^2 = \sum_{i=1}^{n}(X_i - \bar{X}_n)^2/(n-1)$ of a sample of size n from the distribution F as an estimator of σ^2, the variance of the distribution F. To estimate the mean square error of the sample variance, let the observed data be $X_i = x_i, i = 1, \ldots, n$.

Because the distribution F_e puts equal weight on all of the values $x_i, i = 1, \ldots, n$, it follows that

$$E_{F_e}[X] = \sum_{i=1}^{n} x_i P_{F_e}(X = x_i) = \sum_{i=1}^{n} x_i/n = \bar{x}_n$$

showing that θ_e, the variance of the distribution F_e, is given by

$$\theta_e = \text{Var}_{F_e}(X) = E_{F_e}[(X - \bar{x}_n)^2] = \sum_{i=1}^{n}(x_i - \bar{x}_n)^2/n$$

Consequently,

$$MSE_{F_e}(S_n^2) = E_{F_e}[(S_n^2 - \theta_e)^2] = E_{F_e}\left[\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}{n-1} - \theta_e\right)^2\right]$$

To approximate $MSE_{F_e}(S_n^2)$, we use simulation.

For instance, suppose $n = 8$ and the data values were $x_1 = 5, x_2 = 9, x_3 = 12, x_4 = 8, x_5 = 7, x_6 = 15, x_7 = 3, x_8 = 6$. Then

$$\bar{x}_8 = 8.125, \qquad \theta_e = \sum_{i=1}^{8}(x_i - \bar{x}_8)^2/8 \approx 13.11$$

In the following approach for the simulation-based approximation of $MSE_{F_e}(S_n^2)$, the $x_i, i = 1, \ldots, 8$ are as given in the preceding. There are to be a total of r simulation runs, with the variable N representing the number of the current simulation run. In each run we generate the values of 8 random variables X_M, $M = 1, \ldots, 8$, distributed according to F_e. The quantities S and SS represent running totals of, respectively, the sum of the X_M and the sum of the squares of the X_M so far generated in the run. When the run is completed, the sample variance SV is computed by using the identity

$$\frac{\sum_{i=1}^{8}(X_i - \bar{X}_8)^2}{7} = \frac{\sum_{i=1}^{8} X_i^2 - 8\bar{X}_8^2}{7} = \frac{\sum_{i=1}^{8} X_i^2 - (\sum_{i=1}^{8} X_i)^2/8}{7} = \frac{SS - S^2/8}{7}$$

The squared difference between SV and $\theta_e = 13.11$ is computed and then added to T, the sum of the $N - 1$ previous squared differences. When r runs have been completed the simulation is ended; the average of the squared differences between the sample variances and θ_e is the simulation-based approximation to $MSE_{F_e}(S_n^2)$.

```
1. Let $T = 0, N = 1$
2. Let $M = 1$
3. $S = 0,\ SS = 0$
4. Generate a random number $U$
5. Set $I = \text{Int}(8U) + 1$
6. $S = S + x_I$
7. $SS = SS + x_I^2$
8. If $M < 8$, set $M = M + 1$ and go to 4
9. $SV = (SS - S^2/8)/7$
10. Let $T = T + (SV - 13.11)^2$
11. If $N < r$, set $N = N + 1$ and go to 2
12. If $N = r$, return $T/r$ as the approximation to $MSE_{F_e}(S_n^2)$ ■
```

Now suppose we wanted to estimate not the mean square error of the estimator but rather the probability that the estimator of θ will be within h of the actual value of θ. That is, suppose we want to estimate

$$p_h \equiv P_F(|d(X_1, \ldots, X_n) - \theta| \leq h)$$

To obtain an estimator of the preceding, we use that

$$P_F(|d(X_1, \ldots, X_n) - \theta| \leq h) \approx P_{F_e}(|d(X_1, \ldots, X_n) - \theta_e| \leq h)$$

and then employ simulation to estimate the right side of the preceding. That is, after the data $X_1, \ldots, X_n$ are observed to take on the values $X_i = x_i$, $i = 1, \ldots, n$, we let F_e be the empirical distribution. That is, F_e is the distribution function of a random variable that is equally likely to take on any of the values $x_1, \ldots, x_n$. We next compute the value of θ_e. We then continually generate sets of n independent random variables from the distribution F_e. For each set of values obtained, we compute d evaluated at these values, and check whether this quantity is within h of θ_e. The fraction of times that it is within h is our simulation-based estimate of

$$P_{F_e}(|d(X_1, \ldots, X_n) - \theta_e| \leq h)$$

and is also what we use to estimate p_h.

More specifically, we use the original data to obtain F_e and the resulting value of θ_e. We then decide on the number of simulation runs (typically between 10^4 and 10^5 will suffice, but see Section 15.7 for specifics on how to determine the number of runs that should be performed). With r runs, we need to generate r sets of n independent random variables from the distribution F_e. With $x_{i,1}, \ldots, x_{i,n}$ being the ith set of values generated, we compute the value of $d_i = d(x_{i,1}, \ldots, x_{i,n})$. The proportion of the values of i, $i = 1, \ldots, r$, for which $|d_i - \theta_e| \leq h$ is our estimate of $p_h \equiv P_F(|d(X_1, \ldots, X_n) - \theta| \leq h)$.

EXAMPLE 15.3d The following are the PSAT math scores of a random sample of 16 students from a certain school district.

$$522, 474, 644, 708, 466, 534, 422, 480, 502, 655, 418, 464, 600, 412, 530, 564$$

Use them to estimate

(a) the average score of all students in the district;
(b) the probability that the estimator of the district average will be within 5 of the actual district average;
(c) the probability that the estimator of the district average will be within 10 of the actual district average.

SOLUTION We suppose that the data constitute a random sample from a distribution F with mean $\theta(F) = \mu$. The natural estimator of μ is the sample average $\bar{X}$, yielding the estimate

$$\theta_e = \bar{x} = 524.7$$

The probability p_h, that the sample mean of a sample of size 16 will be within h of the population mean, is estimated by

$$P_{F_e}(|\bar{X}_{16} - \theta_e| \le h) = P_{F_e}(|\bar{X}_{16} - 524.7| \le h)$$

where $\bar{X}_{16}$ is the average of a sample of size 16 from the distribution that puts probability 1/16 on each of the original 16 data values. A simulation based on 10^5 simulation runs—with each run generating a sample of size 16 from F_e—yielded the estimates .1801 and .3542 for $h = 5$ and $h = 10$, respectively.

Because we are estimating the mean of the distribution by the sample average, we could also approximate the probability p_h by making use of the central limit theorem. With μ and σ being the mean and standard deviation of F, the probability that the sample mean of a sample of size 16 is within h of μ can be approximated by using the fact that $\bar{X}_{16}$ approximately has a normal distribution with mean μ and variance $\sigma^2/16$. Consequently, with Z being a standard normal random variable

$$\begin{aligned} P(-h \le \bar{X}_{16} - \mu \le h) &= P(\frac{-h}{\sigma/4} \le \frac{\bar{X}_{16} - \mu}{\sigma/4} \le \frac{d}{\sigma/4}) \\ &\approx P(-4h/\sigma \le Z \le 4h/\sigma) \\ &= 2\,\Phi(4h/\sigma) - 1 \end{aligned}$$

An easy calculation gives that the sample standard deviation of the 16 data values is $s = 89.1$. Taking this value as an approximation of σ yields that

$$2\,\Phi(4h/\sigma) - 1 \approx 2\,\Phi(4h/89.1) - 1$$

Thus, the estimate of the probability that the sample mean is within 5 of the population mean is $2\Phi(.2245) - 1 = .1776$, whereas the estimate that it is within 10 of the population mean is $2\Phi(.4490) - 1 = .3466$, which are quite close to the ones obtained by the nonparametric bootstrap approach. However, it should be noted that the central limit theorem approximation would not be available to us if we were estimating some other parameter of the distribution aside from its mean.

For instance, suppose we wanted to use the 16 data values to estimate σ, the standard deviation of the scores of all the student in the district. Using the sample standard deviation as the estimator yields the estimate $s = 89.1$. Now suppose we wanted to estimate the probability that our estimator will be within 10 of σ. That is, suppose we wanted to estimate the probability that the sample standard deviation of a sample of size 16 from the distribution F will be within 10 of the actual standard deviation of F. To do so, we estimate this by the probability that the sample standard deviation of a sample of size 16 from the empirical distribution F_e is within 10 of the standard deviation of F_e. Now, because the distribution F_e puts equal weight on each of the 16 values $x_1, \ldots, x_{16}$, its mean is $\bar{x}$ and its standard deviation is

$$\sigma_e = \sqrt{E_{F_e}[(X - \bar{x})^2]} = \sqrt{\frac{1}{16}\sum_{i=1}^{16}(x_i - \bar{x})^2} = 89.1\sqrt{15/16} = 86.27$$

Consequently the estimate of the probability that the sample standard deviation differs from σ by at most 10 is

$$P_{F_e}(|S_{16} - \sigma_e| \leq 10) = P_{F_e}(|S_{16} - 86.27| \leq 10)$$

where S_{16} is the sample standard deviation of a sample of size 16 from the distribution F_e. This probability can be approximated by a simulation. Indeed, a simulation performed with 10^5 runs yielded the result

$$P_{F_e}(|S_{16} - 86.27| \leq 10) \approx .5424$$

so there is roughly a 54 percent chance that the actual standard deviation of all student scores is within 79.1 and 99.1. ■

15.4 PERMUTATION TESTS

Suppose we want to test the null hypothesis H_0 that the data $X_1, \ldots, X_N$ is a sample from some unspecified distribution. *Permutation tests* are tests of this hypothesis in which the p-value is computed conditional on knowing the set $\boldsymbol{\mathcal{S}}$ of data values observed but without knowing which data value corresponds to X_1, which corresponds to X_2 and so on. For instance, if $N = 3$ and $X_1 = 5, X_2 = 7, X_3 = 2$, then the p-value is computed conditional on the information that the set of data values is $\boldsymbol{\mathcal{S}} = \{2, 5, 7\}$. The computation of the

p-value makes use of the fact that, conditional on the set of data values $\boldsymbol{\mathcal{S}}$, each of the $N!$ possible ways of assigning these N values to the original data is equally likely when the null hypothesis is true. That is, suppose that $N = 3$ and the set of data values is, as in the preceding, $\boldsymbol{\mathcal{S}} = \{2, 5, 7\}$. Now the null hypothesis H_0 states that X_1, X_2, X_3 are independent and identically distributed. Consequently, if H_0 is true then, given the data set $\boldsymbol{\mathcal{S}}$, it follows that the vector (X_1, X_2, X_3) is equally likely to equal any of the 3! permutations of the values $2, 5, 7$.

The implementation of a permutation test is as follows. Depending on the alternative hypothesis, a test statistic $T(X_1, \ldots, X_N)$ is chosen. Suppose, for the moment, that large values of the test statistic are evidence for the alternative hypothesis. The data values are then observed, say that $X_i = x_i, i = 1, \ldots, N$, and the value of $T(x_1, \ldots, x_N)$ is calculated. Now let $\boldsymbol{\mathcal{S}} = \{x_1, \ldots, x_N\}$ be the unordered set consisting of the N observed values. Then, if the value of the test statistic is $T(x_1, \ldots, x_N) = t$, the resulting p-value of the null hypothesis that results from these data is

$$p\text{-value} = P_{H_0}(T(X_1, \ldots, X_N) \geq t | \boldsymbol{\mathcal{S}} = \{x_1, \ldots, x_N\})$$

Now, under H_0, $X_1, \ldots, X_N$ is equally likely to equal any of the $N!$ permutations of $x_1, \ldots, x_N$. Consequently, letting $I_1, \ldots, I_N$ be a random vector that is equally likely to be any of the $N!$ permutations of $1, \ldots, N$, we can write the preceding p-value as

$$\begin{aligned} p\text{-value} &= P\{T(x_{I_1}, x_{I_2}, \ldots, x_{I_N}) \geq t\} \\ &= \frac{\text{number of permutations } (i_1, \ldots, i_N) : T(x_{i_1}, x_{i_2}, \ldots, x_{i_N}) \geq t}{N!} \end{aligned}$$

For an illustration, suppose we are to observe data over N weeks, with X_i being the data value observed in week i, $i = 1, \ldots, N$, and that we want to use these data to test the null hypothesis

$$H_0 : X_1, \ldots, X_N \text{ are independent and identically distributed}$$

against

$$H_1 : X_i \text{ tends to increase as } i \text{ increases}$$

Now if the null hypothesis is true and the data are independent and identically distributed, then, conditional on knowing the set of values $X_1, \ldots, X_N$, but not knowing which value corresponds to X_1 or which corresponds to X_2 and so on, the statistic $\sum_{j=1}^{N} jX_j$ would be distributed as if we randomly paired up the two data sets $\{1, \ldots, N\}$ and $\{X_1, \ldots, X_N\}$ and then summed the products of the N paired values. On the other hand, if the alternative hypothesis were true, then $\sum_{j=1}^{N} jX_j$ would tend to be larger than if we just randomly paired the values $1, \ldots, N$ with the values $X_1, \ldots, X_N$ and then summed the products of the N pairs. This is because the sum of the paired values of two sets of equal size is largest when the largest values are paired with each other, the second largest are paired with each other, and so on. (In statistical terms the correlation coefficient of data pairs

$(j, X_j), j = 1, \ldots, N$ is large when the X_j tend to increase as j increases.) Consequently, one possible permutation test of H_0 versus H_1 is to

1. Observe the data values—say that $X_j = x_j$, $j = 1, \ldots, N$
2. Let $t = \sum_{j=1}^{N} jx_j$
3. Determine the p-value given by

$$p\text{-value} = P(\sum_{j=1}^{N} I_j x_j \geq t)$$

where $I_1, \ldots, I_N$ is equally likely to be any of the $N!$ permutations of $1, \ldots, N$.

The p-value in the preceding can be approximated by a simulation that uses the method of Example 15.2b to generate random permutations.

EXAMPLE 15.4a To determine if the weekly sales of DVD players is on a downward trend, the manager of a large electronics store has been tracking such sales for the past 12 weeks, with the following sales figures from week 1 to week 12 (the current week) resulting:

$$22, 24, 20, 18, 16, 14, 15, 15, 13, 17, 12, 14$$

Are the data strong enough to reject the null hypothesis that the distribution of sales is unchanging in time, and so enable the manager to conclude that there is a downward trend in sales?

SOLUTION Let the null hypothesis be that the distribution of sales is unchanged over time, and let the alternative hypothesis be that there is a downward trend in sales. Thus, if the alternative hypothesis is true then there would be a negative correlation between X_j, the sales during week j, and j. So a relatively small value of $\sum_{j=1}^{12} jX_j$ would be evidence in favor of the alternative hypothesis. Now, with x_j equal to the observed value of X_j, the sales data give that

$$\sum_{j=1}^{12} jx_j = 1{,}178$$

Hence, the p-value of the permutation test of the null hypothesis that the data come from the same distribution versus the alternative that the data tend to be decreasing in time is given by

$$p\text{-value} = P(\sum_{j=1}^{12} I_j x_j \leq 1{,}178)$$

where $I_1, \ldots, I_{12}$ is equally likely to be any of the $12!$ permutations of $1, \ldots, 12$. A simulation, using 10^5 runs, yielded that

$$p\text{-value} \approx .00039$$

leading us to reject the null hypothesis that the distribution is unchanging over time. ■

Although $\sum_{j=1}^{N} jX_j$ is the test statistic most commonly used to test the null hypothesis that $X_1, \ldots, X_n$ are independent and identically distributed against the alternative that X_j tends to increase as j increases, it is not the only possibility. Indeed, we could have chosen any test statistic of the form $\sum_{j=1}^{N} a_j X_j$, where $a_1 < a_2 < \ldots < a_n$. (For instance, we could have chosen $a_j = j^2$.) Analogous to the preceding, the value of the statistic would first be determined, say it is t. Because the alternative hypothesis will tend to make $\sum_{j=1}^{N} a_j X_j$ larger than it would be under the null hypothesis—since large values of the a_j would tend to be paired with large data values when the alternative hypothesis is true—we would again want to reject the null hypothesis when t is large. Consequently, the resulting p-value would be

$$p\text{-value} = P\left(\sum_{j=1}^{N} a_{I_j} x_j \geq t\right)$$

where $I_1, \ldots, I_N$ is equally likely to be any of the $N!$ permutations of $1, \ldots, N$.

Depending on the alternative hypothesis, we could choose other constants $a_j, j = 1, \ldots, N$ to test the null hypothesis that the data values are independent and identically distributed. For instance, if the alternative was that the data tended to be higher in the middle values and lower in the extremes, then we could let the test statistic be of the form $T = \sum_{j=1}^{N} a_j X_j$, where $a_1, \ldots, a_N$ is such that its middle values tend to be larger than its earlier or later values. For instance, we could use $a_j = j(N - j)$, $j = 1, \ldots, N$. As this would again make it more likely that larger data values are paired with larger constants when the alternative hypothesis is true, we would again want to reject the null hypothesis when T is large.

15.4.1 Normal Approximations in Permutation Tests

Although not as accurate as doing a simulation, the p-value of a permutation test can be approximated by assuming that the test statistic is approximately normally distributed. Now, under the null hypothesis that the data values are independent and identically distributed, it follows that, given the data set $\mathcal{S} = \{x_1, \ldots, x_N\}$, the random variable X_i is equally likely to be any of these N values and the random vector $(X_i, X_j), i \neq j$ is equally likely to take on any of the $N(N-1)$ values $x_k x_r$, $r \neq k$. Consequently, given $\mathcal{S} = \{x_1, \ldots, x_N\}$,

$$E[X_i] = \frac{1}{N}\sum_{i=1}^{N} x_i = \bar{x}$$

$$E[X_i^2] = \frac{1}{N}\sum_{i=1}^{N} x_i^2$$

$$E[X_i X_j] = \frac{1}{N(N-1)}\sum_{k}\sum_{r \neq k} x_k x_r$$

$$= \frac{1}{N(N-1)}\Big(\sum_k \sum_r x_k x_r - \sum_k \sum_{r=k} x_k x_r\Big)$$

$$= \frac{1}{N(N-1)}\Big(\sum_k x_k \sum_r x_r - \sum_k x_k^2\Big)$$

$$= \frac{1}{N(N-1)}\Big(N^2 \bar{x}^2 - \sum_{k=1}^{N} x_k^2\Big)$$

So, with $v = \text{Var}(X_i)$ and $c = \text{Cov}(X_i, X_j)$, $i \neq j$, the preceding yields

$$E[X_i] = \bar{x}$$

$$v = \text{Var}(X_i) = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2$$

$$\begin{aligned} c = \text{Cov}(X_i, X_j) &= \frac{1}{N(N-1)}\,(N^2 \bar{x}^2 - \sum_{k=1}^{N} x_k^2) - \bar{x}^2 \\ &= \frac{\bar{x}^2}{N-1} - \frac{1}{N(N-1)}\sum_{k=1}^{N} x_k^2 \\ &= \frac{1}{N-1}\,(\bar{x}^2 - \sum_{k=1}^{N} x_k^2/N) \end{aligned}$$

which also shows that

$$v - c = \frac{\sum_{i=1}^{N} x_i^2 - N\bar{x}^2}{N-1}$$

Therefore, when H_0 is true, the test statistic $T = \sum_{j=1}^{N} jX_j$ has mean

$$E[T] = \frac{N(N+1)}{2}\bar{x}$$

and variance

$$\begin{aligned} \text{Var}(T) &= \text{Var}\Big(\sum_{j=1}^{N} jX_j\Big) \\ &= \sum_{j=1}^{N} \text{Var}(jX_j) + \sum_i \sum_{j \neq i} \text{Cov}(iX_i, jX_j) \\ &= v\sum_{j=1}^{N} j^2 + c\sum_i \sum_{j \neq i} ij \end{aligned}$$

$$
\begin{aligned}
&= v\sum_{j=1}^{N} j^2 + c\left(\sum_i\sum_j ij - \sum_i\sum_{j=i} ij\right)\\
&= v\sum_{j=1}^{N} j^2 + c\left(\sum_{i=1}^{N} i\sum_{j=1}^{N} j - \sum_{i=1}^{N} i^2\right)\\
&= (v-c)\sum_{j=1}^{N} j^2 + \frac{cN^2(N+1)^2}{4}\\
&= (v-c)\,\frac{N(N+1)(2N+1)}{6} + \frac{cN^2(N+1)^2}{4}
\end{aligned}
$$

Using the preceding we can approximate the p-value of a permutation test by assuming that the distribution of T, when H_0 is true, is approximately normal.

EXAMPLE 15.4b Again consider Example 15.4a. A calculation yields that, under H_0,

$$E[T] = 1{,}300 \quad \text{Var}(T) = 1{,}958.81$$

Thus, the normal approximation yields that

$$
\begin{aligned}
p\text{-value} &= P_{H_0}(T \le 1{,}178)\\
&= P_{H_0}\Big(\frac{T - 1{,}300}{\sqrt{1{,}958.81}} \le \frac{1{,}178 - 1{,}300}{\sqrt{1{,}958.81}}\Big)\\
&\approx \Phi(-2.757)\\
&= .0029
\end{aligned}
$$

which is quite close to the value given by the simulation.

Let us now suppose that whereas the set of 12 data values was as before, they now appeared in the order

$$22, 14, 14, 16, 24, 20, 18, 15, 17, 15, 12, 13$$

With these data, the value of the test statistic is $\sum_{j=1}^{12} jX_j = 1{,}233$, and the normal approximation yields that

$$
\begin{aligned}
p\text{-value} &= P_{H_0}(T \le 1{,}233)\\
&= P_{H_0}\Big(\frac{T - 1, 300}{\sqrt{1{,}958.81}} \le \frac{1{,}233 - 1, 300}{\sqrt{1{,}958.81}}\Big)\\
&\approx \Phi(-1.514)\\
&= .065
\end{aligned}
$$

Finally, suppose again that the set of 12 data values was as before, but suppose that they now appeared in the order

$$22, 14, 14, 16, 24, 13, 18, 15, 17, 15, 12, 20$$

In this case, the value of the test statistic is $\sum_{j=1}^{12} jX_j = 1275$. Thus, the normal approximation yields that

$$\begin{aligned}
p\text{-value} &= P_{H_0}(T \leq 1{,}275) \\
&= P_{H_0}\left(\frac{T - 1{,}300}{\sqrt{1{,}958.81}} \leq \frac{1{,}275 - 1{,}300}{\sqrt{1{,}958.81}}\right) \\
&\approx \Phi(-.565) \\
&\approx .286
\end{aligned}$$

A simulation of 10^5 runs yielded values quite similar to the preceding. The simulation gave

$$P_{H_0}(T \leq 1{,}233) \approx .068$$

and

$$P_{H_0}(T \leq 1{,}275) \approx .299$$

which are quite close to the values given by the normal approximation. ■

EXAMPLE 15.4c For another indication as to the validity of a normal approximation, suppose that $N = 4$, with the data appearing in the following order:

$$13, 7, 5, 3$$

Suppose that we want to use these data to test the null hypothesis that the data are a sample from some distribution against the alternative hypothesis that the data tend to be decreasing. The value of the test statistic is $T = \sum_{j=1}^{r} jX_j = 54$. An easy computation yields

$$c = -4.667, \quad v = 14$$

showing that, under H_0,

$$E[T] = 70, \quad \text{Var}(T) = 93.33$$

Consequently, with Z being a standard normal random, the normal approximation yields that

$$\begin{aligned}
p\text{-value} &= P_{H_0}(T \leq 54) \\
&= P_{H_0}\left(\frac{T - 70}{\sqrt{93.33}} \leq \frac{54 - 70}{\sqrt{93.33}}\right) \\
&\approx P(Z \leq -1.656) \\
&= .049
\end{aligned}$$

whereas the exact value is

$$p\text{-value} = P_{H_0}(T \leq 54) = 1/4! \approx .042 \quad ■$$

15.4.2 Two-Sample Permutation Tests

Permutation tests are also useful in the two-sample problems where we test whether samples from two populations have the same underlying distribution. Specifically, let $X_1, \ldots, X_n$ be a sample from an unknown population distribution F, and let $X_{n+1}, \ldots, X_{n+m}$ be an independent sample from an unknown population distribution G, and suppose we want to use these data to test the hypothesis that the two population distributions are identical against the alternative hypothesis that data from the second distribution tend to be larger than those from the first. That is, we want to use these data to test the null hypothesis

$$H_0 : F = G$$

against the alternative

$$H_1 : \text{ data from } G \text{ tend to be larger than data from } F$$

If the data values are $X_i = x_i$, $i = 1, \ldots, n+m$, then a permutation test of the preceding null hypothesis is done conditional on knowing $\mathcal{S} = \{x_1, \ldots, x_{n+m}\}$, the set of these $n + m$ numbers in no particular order. Then if H_0 is true, and so all $n + m$ random variables $X_1, \ldots, X_{n+m}$ are independent and identically distributed, then given the set of values $\mathcal{S}$, each subset of size n of this set is equally likely to be the set of the data values of $X_1, \ldots, X_n$. Because the alternative hypothesis is that data from the population distribution F tend to be smaller than data from the population distribution G, a reasonable test would be to reject the null hypothesis if the sum of the data values from the population distribution F is smaller than might be expected by chance when n values are randomly chosen from the data set $\mathcal{S}$. More specifically, we can test H_0 by computing $\sum_{i=1}^{n} x_i$; say its value is t. Then the p-value of this permutation test of H_0 versus H_1 would equal the probability that a random selection of n of the values $x_1, \ldots, x_{n+m}$ would be less than or equal to t. That is,

$$p\text{-value} = P\left(\sum_{i \in R} x_i \leq t\right)$$

where R is equally likely to be any of the $\binom{n+m}{n}$ subsets of size n from the set $\{1, 2, \ldots, n+m\}$. Whereas an exact computation of the preceding is possible only when $\binom{n+m}{n}$ is small, a precise approximation is easily obtained by simulation. In each simulation run we use the method of Example 15.2a to randomly generate a subset of n of the values $1, \ldots, n + m$. If R is the subset obtained, then we check whether $\sum_{i \in R} x_i$ is less than or equal to t. The fraction of simulation runs for which this is the case is our estimate of the preceding p-value.

REMARK

We can again use a normal approximation, rather than a simulation, to estimate the p-value. Starting with the random subset R, equally likely to be any of the $\binom{n+m}{n}$ subsets of size n from the set $\{1, 2, \ldots, n+m\}$, let, for $i = 1, \ldots, n+m$,

$$I_i = \begin{cases} 1, & \text{if } i \in R \\ 0, & \text{if } i \notin R \end{cases}$$

Then,

$$\sum_{i \in R} x_i = \sum_{i=1}^{n+m} x_i I_i$$

By a similar analysis as in Section 15.4.1, we can now show that

$$E\left[\sum_{i \in R} x_i\right] = E\left[\sum_{i=1}^{n+m} x_i I_i\right] = n\bar{x}$$

and

$$\text{Var}\left(\sum_{i \in R} x_i\right) = \text{Var}\left(\sum_{i=1}^{n+m} x_i I_i\right) = \frac{nm}{n+m-1}\left(\frac{\sum_{i=1}^{n+m} x_i^2}{n+m} - \bar{x}^2\right)$$

where $\bar{x} = \sum_{i=1}^{n+m} x_i/(n+m)$.

15.5 GENERATING DISCRETE RANDOM VARIABLES

Suppose we want to generate the value of a random variable X having probability mass function

$$P(X = x_i) = p_i, i = 1, \ldots, \sum_i p_i = 1$$

To generate the value of X, generate a random number U and set

$$X = x_i \quad \text{if} \quad p_1 + \ldots p_{i-1} < U \leq p_1 + \ldots p_{i-1} + p_i$$

That is,

$$X = \begin{cases} x_1, & \text{if } U \leq p_1 \\ x_2, & \text{if } p_1 < U \leq p_1 + p_2 \\ x_3, & \text{if } p_1 + p_2 < U \leq p_1 + p_2 + p_3 \\ . & \\ . & \\ . & \\ x_i, & \text{if } p_1 + \ldots + p_{i-1} < U \leq p_1 + \ldots + p_{i-1} + p_i \\ . & \\ . & \\ . & \end{cases}$$

Because U is uniformly distributed on $(0, 1)$, it follows that for $0 < a < b < 1$

$$P(a < U \leq b) = b - a$$

Consequently,

$$P\left(\sum_{j=1}^{i-1} p_j < U \leq \sum_{j=1}^{i} p_j\right) = p_i$$

which shows that X has the desired probability mass function. This method of generating X is called the *discrete inverse transform method.*

EXAMPLE 15.5a To generate a Bernoulli random variable X such that

$$P(X = 1) = p = 1 - P(X = 0)$$

generate a random number U, and set

$$X = \begin{cases} 1, & \text{if } U \leq p \\ 0, & \text{if } U > p. \end{cases}$$ ■

EXAMPLE 15.5b Suppose now that we wanted to generate a binomial random variable X with parameters n and p. Recalling that X represents the number of successes in n independent trials when each trial is a success with probability p, we can generate X by generating the results of the n trials. That is, we can generate n random numbers $U_1, \ldots, U_n$, say that trial i is a success if $U_i \leq p$, and then set

$$X = \text{number of } i : U_i \leq p$$

Another possibility is to use the inverse transform method.

To efficiently use the inverse transform method we need an efficient method to recursively compute the values

$$p_i = P(X = i) = \binom{n}{i} p^i (1 - p)^{n-i}, \; i = 0, \ldots, n$$

This is accomplished by first noting that

$$\begin{aligned} \frac{\binom{n}{i+1}}{\binom{n}{i}} &= \frac{n!}{(n - i - 1)!\,(i + 1)!} \frac{(n - i)!\,i!}{n!} \\ &= \frac{n - i}{i + 1} \end{aligned}$$

which yields that

$$\begin{aligned}\frac{p_{i+1}}{p_i} &= \frac{n-i}{i+1}\,\frac{p^{i+1}(1-p)^{n-i-1}}{p^i(1-p)^{n-i}}\\ &= \frac{n-i}{i+1}\,\frac{p}{1-p}\end{aligned}$$

Thus,

$$p_{i+1} = \frac{n-i}{i+1}\,\frac{p}{1-p}\,p_i$$

Using the preceding, we are now ready to give the inverse transform method for generating a binomial (n, p) random variable X. In the following, i represents the possible value of X, the variable P is the probability that $X = i$, and the variable F is the probability that $X \leq i$. (That is, for given i, $P = p_i$ and $F = \sum_{j=0}^{i} p_j$.) Also, let $\alpha = p_0 = (1-p)^n$, and let $b = \frac{p}{1-p}$.

1. Set $i = 0,\ \ P = \alpha,\ \ F = \alpha$
2. Generate a random number U
3. If $U \leq F$ set $X = i$ and stop
4. $P = \frac{n-i}{i+1}\,b\,P$
5. $F = F + P$
6. $i = i + 1$
7. Go to 3

(In the preceding, when we say that $P = \frac{n-i}{i+1}\,b\,P$, we don't mean this literally as an algebraic identity; rather we mean that the value of P is to be changed. Its new value is its old value multiplied by $\frac{n-i}{i+1}\,b$. Similarly, when we write $F = F + P$ we mean that the value of F is to be changed by adding P to its old value.)

Because the algorithm first checks whether $X = 0$, then whether $X = 1$, and so on, it follows that the number of iterations needed (that is, the number of times that it goes to step 3) is one more than the final value of X. So, on average, this algorithm requires $E[X + 1] = np + 1$ iterations to generate the value of X. ■

15.6 GENERATING CONTINUOUS RANDOM VARIABLES

Let F be the distribution function of a continuous random variable. For any u between 0 and 1, the quantity $F^{-1}(u)$ is defined to be that value x such that $F(x) = u$. That is, $F(F^{-1}(u)) = u$. Because the distribution function of a continuous random variable is strictly increasing, it follows that there is a unique value of $F^{-1}(u)$. We call F^{-1} the inverse function of F.

A general method for generating a continuous random variable having distribution function F, known as the *inverse transformation method*, is based on the following proposition.

PROPOSITION 15.6.1 Let U be a uniform $(0, 1)$ random variable. For any continuous distribution function F, if we define

$$X = F^{-1}(U)$$

then X has distribution function F.

Proof

Because a distribution function F is nondecreasing, it follows that for any numbers a and b the inequality $a \leq b$ is equivalent to the inequality $F(a) \leq F(b)$. Consequently,

$$\begin{aligned} P(F^{-1}(U) \leq x) &= P(F(F^{-1}(U)) \leq F(x)) \\ &= P(U \leq F(x)) \\ &= F(x) \end{aligned}$$

thus showing that $F^{-1}(U)$ has distribution F. ■

EXAMPLE 15.6a (Generating an Exponential Random Variable) Let

$$F(x) = 1 - e^{-\lambda x}, \quad x \geq 0$$

be the distribution function of an exponential random variable with parameter λ. Then $F^{-1}(u)$ is that value x such that

$$u = F(x) = 1 - e^{-\lambda x}$$

or, equivalently,

$$e^{-\lambda x} = 1 - u$$

or

$$-\lambda x = \log(1 - u)$$

or

$$x = -\frac{1}{\lambda} \log(1 - u)$$

So, by Proposition 15.6.1, we can generate an exponential random variable X with parameter λ by generating a uniform $(0, 1)$ random variable U and setting

$$X = -\frac{1}{\lambda} \log(1 - U)$$

Because $1 - U$ is also a uniform $(0, 1)$ random variable, it follows that $-\frac{1}{\lambda}\log(1 - U)$ and $-\frac{1}{\lambda}\log(U)$ have the same distribution, thus showing that

$$X = -\frac{1}{\lambda}\log(U)$$

is also exponential with parameter λ. ■

15.6.1 Generating a Normal Random Variable

Because inverting the distribution function of a normal random variable is computationally involved, special methods are used for generating normal random variables. The following one is known as the *Box-Muller method.*

To begin, suppose that X and Y are independent standard normal random variables, so their joint density function is

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

Let R, Θ be the polar coordinates of the point (X, Y). Now $R^2 = X^2 + Y^2$ is, by definition, a chi-square random variable with 2 degrees of freedom, and as shown in Section 5.8.1.1 this distribution is the same as an exponential distribution with parameter 1/2 (that is, with mean 2). Consequently, the density function of R^2 is

$$f_{R^2}(r) = \frac{1}{2}e^{-r/2}, \quad 0 < r < \infty$$

Consider now the conditional joint density function of X, Y given that $R^2 = r$. Because

$$f(x,y) = \frac{1}{2\pi}e^{-r/2} \quad \text{when} \quad x^2 + y^2 = r$$

is a constant when $x^2 + y^2 = r$, it is intuitive (and can be proven) that conditional on $R^2 = r$, the vector X, Y is uniformly distributed on the circumference of the circle of radius $\sqrt{r}$. But this implies that, conditional on $R^2 = r$, the polar coordinate Θ of the point (X, Y) is uniformly distributed over $(0, 2\pi)$. Because this is true for all r, it follows that the polar coordinates R and Θ are independent, with R distributed as the square root of an exponential random variable with mean 2, and Θ being a uniform random variable on $(0, 2\pi)$.

Using the preceding, we can generate independent standard normal random variables X and Y by first generating their polar coordinates R and Θ. Because $-\log(U)$ is exponential with mean 1, we can generate the polar coordinates of (X, Y) by generating independent uniform $(0, 1)$ random variables U_1 and U_2 and then setting

$$R^2 = -2\log(U_1)$$

and

$$\Theta = 2\pi U_2$$

Using the formula for going from the polar coordinates R, Θ back to the rectangular coordinates

$$X = R\cos(\Theta), \quad Y = R\sin(\Theta)$$

shows that

$$\begin{aligned} X &= \sqrt{-2\log(U_1)}\cos(2\pi U_2) \\ Y &= \sqrt{-2\log(U_1)}\sin(2\pi U_2) \end{aligned}$$

are independent standard normal random variables.

To generate normal random variables with mean μ and variance σ^2, just generate the independent standard normals X and Y and then take the variables $\mu+\sigma X$ and $\mu+\sigma Y$.

15.7 DETERMINING THE NUMBER OF SIMULATION RUNS IN A MONTE CARLO STUDY

Suppose we are going to generate r independent and identically distributed random variables $Y_1, \ldots, Y_r$ having mean μ, so as to use

$$\bar{Y}_r = \sum_{i=1}^{r} Y_i/r$$

as an estimator of μ. Now, with σ^2 being the variance of the Y_i, it follows by the central limit theorem that $\bar{Y}_r$ will approximately have a normal distribution with mean μ and variance σ^2/r. Consequently, we can be 95 percent certain that μ will lie in the interval

$$(\bar{Y}_r - 1.96\,\sigma/\sqrt{r}, \quad \bar{Y}_r + 1.96\,\sigma/\sqrt{r}).$$

(More generally, we can be $100(1-\alpha)$ percent confident that μ will be between $\bar{Y}_r \pm z_{\alpha/2}\,\sigma/\sqrt{r}$.)

Thus, if σ^2 were known we could choose r to give ourselves the desired level of accuracy. However, it is almost always the case that σ^2, like μ, will be unknown. To get around this difficulty, we can do a two-stage simulation experiment. In the first stage, we generate k runs where k is typically much smaller than the number we expect to use in the study. Doing these runs generates the values of the random variables $Y_1, \ldots, Y_k$. We then use the sample variance of these values,

$$S_k^2 = \frac{1}{k-1}\sum_{i=1}^{k}(Y_i - \bar{Y}_k)^2$$

to estimate σ^2. Then, acting as if that were the actual value of σ^2, we determine an appropriate value for r. Then, in the second stage of the simulation, we generate an additional $r - k$ runs.

Problems

1. If $x_0 = 5$, and

$$x_n = 3\,x_{n-1} \quad \text{mod } 5$$

find $x_1, x_2, \ldots, x_{10}$.

2. Another method of generating a random permutation, different from the one given in Example 15.2b, is to successively generate a random permutation of the numbers $1, 2, \ldots, n$ starting with $n = 1$, then $n = 2$, and so on. (Of course, the random permutation when $n = 1$ is 1.) Once we have a random permutation of the numbers $1, \ldots, n - 1$ — call it $P_1, P_2, \ldots, P_{n-1}$ — the random permutation of the numbers $1, \ldots, n$ is obtained by starting with the permutation $P_1, P_2, \ldots, P_{n-1}, n$, then interchanging the element in position n (namely, n) with the element in a randomly chosen position that is equally likely to be any of the positions $1, 2, \ldots, n$.

(a) Write an algorithm that accomplishes the preceding.
(b) Verify when $n = 2$ and when $n = 3$ that all $n!$ possible permutations are equally likely.

3. Suppose that we are to observe the independent and identically distributed vectors $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, and that we want to use these data to estimate $\theta \equiv E[X_1]/E[Y_1]$.

(a) Give an estimator of θ.
(b) Explain how you could estimate the mean square error of this estimator.

4. Suppose that $X_1, \ldots, X_n$ is a sample from a distribution whose variance σ^2 is unknown. Suppose we are planning to estimate σ^2 by the sample variance $S^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1)$, and we want to use the bootstrap technique to estimate $\text{Var}(S^2)$.

(a) If $n = 2$ and $X_1 = 1$ and $X_2 = 3$, what is the bootstrap estimate of $\text{Var}(S^2)$?
(b) If $n = 15$, and the data values are

$$5, 4, 9, 6, 21, 17, 11, 20, 7, 10, 21, 15, 13, 8, 6$$

use simulation to obtain the bootstrap estimate of $\text{Var}(S^2)$.

5. Let $X_1, \ldots, X_8$ be independent and identically distributed random variables with mean μ. Let

$$p = P\left(\sum_{i=1}^{8} X_i/8 < \mu\right)$$

Estimate p if the values of the X_i are $5, 2, 8, 6, 24, 6, 9, 4$.

6. The following are a student's weekly exam scores. Do they prove that the student improved (as far as exam score) as the semester progressed?

$$68, 64, 72, 80, 72, 84, 76, 86, 94, 92$$

7. A baseball player has the reputation of starting slowly at the beginning of a season but then continually improving as the season progresses. Do the following data, which indicate the number of hits he has in consecutive five-game strings of the season, strongly validate the player's reputation?

$$8, 3, 7, 12, 4, 7, 13, 6, 0, 9, 12, 4, 4, 6, 10$$

8. A group of 16 mice were exposed to 300 rads of radiation at the age of 5 weeks. The group was then randomly divided into two subgroups. Mice in the first subgroup lived in a normal laboratory environment, whereas those from the second subgroup were raised in a special germ-free environment. The following data give the lifetimes, in days, of the mice in each group:

Group 1 lifetimes:	133, 145, 156, 159, 164, 202, 208, 222
Group 2 lifetimes:	145, 148, 157, 171, 178, 191, 200, 204

Use a permutation test to test the hypothesis that the lifetime distributions are identical. Use the normal approximation to approximate the p-value.

9. Do Problem 13 in Chapter 12 by using a permutation test. Use the normal approximation to approximate the p-value.

10. Do Problem 16 in Chapter 12 by using a permutation test. Use the normal approximation to approximate the p-value.

11. Write an algorithm, similar to what was done in the text to generate a binomial random variable, that uses the discrete inverse transform algorithm to generate a Poisson random variable with mean λ.

12. Show that the discrete inverse transform algorithm for generating a geometric random variable with parameter p reduces to the following:

    1. Generate a random number U
    2. Set $X = \text{Int}(\frac{\log(1-U)}{\log(1-p)}) + 1$

Give a second algorithm for generating a geometric random variable with parameter p that takes into account the probabilistic interpretation of such a random variable.

13. Give a method for generating a random variable having density function

$$f(x) = e^x/(e-1), \quad 0 < x < 1$$

14. Give a method for generating a random variable having distribution function

$$F(x) = x^n, \quad 0 < x < 1$$

15. Give a method for generating a random variable having distribution function

$$F(x) = \frac{1}{2}(x + x^2), \quad 0 < x < 1$$

16. Suppose that the following are the generated values of 20 random variables from the distribution F, whose mean μ is unknown:

$$5, 4, 9, 6, 21, 12, 7, 14, 17, 11, 20, 7, 10, 21, 15, 26, 9, 13, 8, 6$$

How many additional random variables from F will we need to generate if we want to be 99 percent certain that our estimate of μ is correct to within ± 0.1?