Statistical Validation Techniques

Introduction

In this chapter we consider some statistical procedures that are useful in validating simulation models. Sections 11.1 and 11.2 consider goodness of fit tests, which are useful in ascertaining whether an assumed probability distribution is consistent with a given set of data. In Section 11.1 we suppose that the assumed distribution is totally specified, whereas in Section 11.2 we suppose that it is only specified up to certain parameters—for example, it may be Poisson having an unknown mean. In Section 11.3 we show how one can test the hypothesis that two separate samples of data come from the same underlying population—as would be the case with real and simulated data when the assumed mathematical model being simulated is an accurate representation of reality. The results entation of reality. The results of Section 11.3 are particularly useful in testing the validity of a simulation model. A generalization to the case of many samples is also presented in this section. Finally, in Section 11.4, we show how to use real data to test the hypothesis that the process generating the data constitutes a nonhomogeneous Poisson process. The case of a homogeneous Poisson process is also considered in this section.

11.1 Goodness of Fit Tests

One often begins a probabilistic analysis of a given phenomenon by hypothesizing that certain of its random elements have a particular probability distribution. For example, we might begin an analysis of a traffic network by supposing that the daily number of accidents has a Poisson distribution. Such hypotheses can be statistically tested by observing data and then seeing whether the assumption of

Simulation. DOI: http://dx.doi.org/10.1016/B978-0-12-415825-2.00011-5

a particular probability distribution is consistent with these data. These statistical tests are called *goodness of fit* tests.

One way of performing a goodness of fit test is to first partition the possible values of a random quantity into a finite number of regions. A sample of values of this quantity is then observed and a comparison is made between the numbers of them that fall into each of the regions and the theoretical expected numbers when the specified probability distribution is indeed governing the data.

In this section we consider goodness of fit tests when all the parameters of the hypothesized distribution are specified; in the following section we consider such tests when certain of the parameters are unspecified. We first consider the case of a discrete and then a continuous hypothesized distribution.

The Chi-Square Goodness of Fit Test for Discrete Data

Suppose that n independent random variables—$Y_1, \ldots, Y_n$—each taking on one of the values $1, 2, \ldots, k$, are to be observed, and that we are interested in testing the hypothesis that $\{p_i, i = 1, \ldots, k\}$ is the probability mass function of these random variables. That is, if Y represents any of the Y_j, the hypothesis to be tested, which we denote by H_0 and refer to as the *null hypothesis,* is

$$H_0 : P\{Y = i\} = p_i, \quad i = 1, \ldots, k$$

To test the foregoing hypothesis, let N_i, $i = 1, \ldots, k$, denote the number of the Y_j's that equal i. Because each Y_j independently equals i with probability $P\{Y = i\}$, it follows that, under H_0, N_i is binomial with parameters n and p_i. Hence, when H_0 is true,

$$E[N_i] = np_i$$

and so $(N_i - np_i)^2$ is an indication as to how likely it appears that p_i indeed equals the probability that $Y = i$. When this is large, say, in relation to np_i, then it is an indication that H_0 is not correct. Indeed, such reasoning leads us to consider the quantity

$$T = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$$

and to reject the null hypothesis when T is large.

Whereas small values of the test quantity T are evidence in favor of the hypothesis H_0, large ones are indicative of its falsity. Suppose now that the actual data result in the test quantity T taking on the value t. To see how unlikely such a large outcome would have been if the null hypothesis had been true, we define the so-called p-value by

$$p\text{-value} = P_{H_0}\{T \geqslant t\}$$

where we have used the notation P_{H_0} to indicate that the probability is to be computed under the assumption that H_0 is correct. Hence, the p-value gives the probability that a value of T as large as the one observed would have occurred if the null hypothesis were true. It is typical to reject the null hypothesis—saying that it appears to be inconsistent with the data—when a small p-value results (a value less than 0.05, or more conservatively, 0.01 is usually taken to be critical) and to accept the null hypothesis—saying that it appears to be consistent with the data—otherwise.

After observing the value—call it t—of the test quantity, it thus remains to determine the probability

$$p\text{-value} = P_{H_0}\{T \geqslant t\}$$

A reasonably good approximation to this probability can be obtained by using the classical result that, for large values of n, T has approximately a chi-square distribution with $k - 1$ degrees of freedom when H_0 is true. Hence,

$$p\text{-value} \approx P\left\{X_{k-1}^2 \geqslant t\right\} \tag{11.1}$$

where X_{k-1}^2 is a chi-square random variable with $k - 1$ degrees of freedom.

Example 11a Consider a random quantity which can take on any of the possible values 1, 2, 3, 4, 5, and suppose we want to test the hypothesis that these values are equally likely to occur. That is, we want to test

$$H_0 : p_i = 0.2, \quad i = 1, \ldots, 5$$

If a sample of size 50 yielded the following values of N_i:

$$12, 5, 19, 7, 7$$

then the approximate p-value is obtained as follows. The value of the test statistic T is given by

$$T = \frac{4 + 25 + 81 + 9 + 9}{10} = 12.8$$

This yields

$$p\text{-value} \approx P\left\{X_4^2 > 12.8\right\} = 0.0122$$

For such a low p-value the hypothesis that all outcomes are equally likely would be rejected. □

If the p-value approximation given by Equation (11.1) is not too small—say, of the order of 0.15 or larger—then it is clear that the null hypothesis is not going to be rejected, and so there is no need to look for a better approximation. However, when the p-value is closer to a critical value (such as 0.05 or 0.01) we would probably want a more accurate estimate of its value than the one given by the chi-square approximate distribution. Fortunately, a more accurate estimator can be obtained via a simulation study.

To effect the simulation study we need to generate $N_1, \ldots, N_k$, where N_i is the number of $Y_1, \ldots, Y_n$, independent random variables having mass function $\{p_i, i = 1, \ldots, k\}$, that are equal to i, This can be accomplished in two different ways. One way is to generate the values $Y_1, \ldots, Y_n$ and then use these values to determine $N_1, \ldots, N_k$. Another way is to generate $N_1, \ldots, N_k$ directly by first generating N_1, then generating N_2 given the generated value of N_1, and so on. This is done by using that N_1 is binomial with parameters (n, p_1); that the conditional distribution of N_2 given that $N_1 = n_1$ is binomial with parameters $(n - n_1, \frac{p_2}{1-p_1})$; that the conditional distribution of N_3 given that $N_1 = n_1, N_2 = n_2$ is binomial with parameters $(n - n_1 - n_2, \frac{p_3}{1-p_1-p_2})$, and so on. If n is much larger than k the second approach is preferable.

The Kolmogorov–Smirnov Test for Continuous Data

Now consider the situation where $Y_i, \ldots, Y_n$ are independent random variables, and we are interested in testing the null hypothesis H_0 that they have the common distribution function F, where F is a given continuous distribution function. One approach to testing H_0 is to break up the set of possible values of the Y_j into k distinct intervals, say,

$$(y_0, y_1), (y_1, y_2), \ldots, (y_{k-1}, y_k), \quad \text{where} y_0 = -\infty, \; y_k = +\infty$$

and then consider the discretized random variables Y_j^d, $j = 1, \ldots, n$, defined by

$$Y_j^d = i \quad \text{if } Y_j \text{ lies in the interval } (y_{i-1}, y_i)$$

The null hypothesis then implies that

$$P\left\{Y_j^d = i\right\} = F(y_i) - F(y_{i-1}), \quad i = 1, \ldots, k$$

and this can be tested by the chi-square goodness of fit test already presented.

There is, however, another way of testing that the Y_j come from the continuous distribution function F which is generally more efficient than discretizing; it works as follows. After observing $Y_1, \ldots, Y_n$, let F_e be the empirical distribution function defined by

$$F_e(x) = \frac{\#i : Y_i \leqslant x}{n}$$

That is, $F_e(x)$ is the proportion of the observed values that are less than or equal to x. Because $F_e(x)$ is a natural estimator of the probability that an observation is less than or equal to x, it follows that, if the null hypothesis that F is the underlying distribution is correct, it should be close to $F(x)$. Since this is so for all x, a natural quantity on which to base a test of H_0 is the test quantity

$$D \equiv \underset{x}{\text{Maximum}} |F_e(x) - F(x)|$$

where the maximum is over all values of x from $-\infty$ to $+\infty$. The quantity D is called the *Kolmogorov–Smirnov test statistic.*

To compute the value of D for a given data set $Y_j = y_j,\ j = 1, \ldots, n$, let $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ denote the values of the y_j in increasing order. That is,

$$y_{(j)} = j\text{th smallest of } y_1, \ldots, y_n$$

For example, if $n = 3$ and $y_1 = 3,\ y_2 = 5,\ y_3 = 1$, then $y_{(1)} = 1,\ y_{(2)} = 3,\ y_{(3)} = 5$. Since $F_e(x)$ can be written

$$F_e(x) = \begin{cases} 0 & \text{if } x < y_{(1)} \\ \frac{1}{n} & \text{if } y_{(1)} \leqslant x < y_{(2)} \\ \vdots & \\ \frac{j}{n} & \text{if } y_{(j)} \leqslant x < y_{(j+1)} \\ \vdots & \\ 1 & \text{if } y_{(n)} \leqslant x \end{cases}$$

we see that $F_e(x)$ is constant within the intervals $(y_{(j-1)}, y_{(j)})$ and then jumps by $1/n$ at the points $y_{(1)}, \ldots, y_{(n)}$. Since $F(x)$ is an increasing function of x which is bounded by 1, it follows that the maximum value of $F_e(x) - F(x)$ is nonnegative and occurs at one of the points $y_{(j)},\ j = 1, \ldots, n$ (see Figure 11.1). That is,

$$\underset{x}{\text{Maximum}}\{F_e(x) - F(x)\} = \underset{j=1,\ldots,n}{\text{Maximum}}\left\{\frac{j}{n} - F(y_{(j)})\right\} \tag{11.2}$$

Similarly, the maximum value of $F(x) - F_e(x)$ is also nonnegative and occurs immediately before one of the jump points $y_{(j)}$, and so

$$\underset{x}{\text{Maximum}}\{F(x) - F_e(x)\} = \underset{j=1,\ldots,n}{\text{Maximum}}\left\{F(y_{(j)}) - \frac{(j-1)}{n}\right\} \tag{11.3}$$

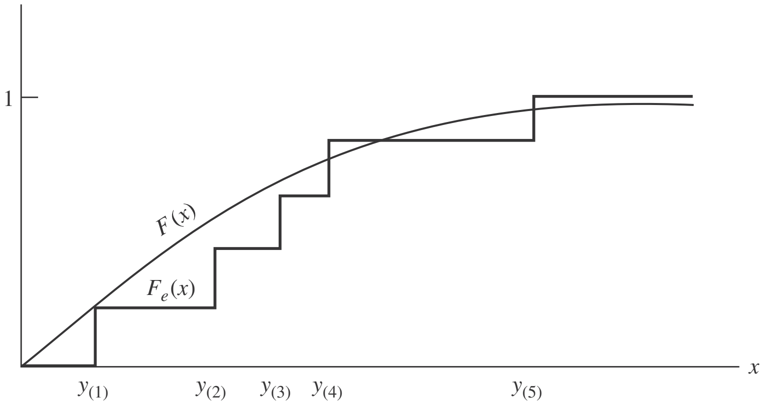

Figure 11.1. $n = 5$.

From Equations (11.2) and (11.3) we see that

$$\begin{aligned} D &= \underset{x}{\text{Maximum}}\, |F_e(x) - F(x)| \\ &= \text{Maximum}\{\text{Maximum}\{F_e(x) - F(x)\}, \text{Maximum}\{F(x) - F_e(x)\}\} \\ &= \text{Maximum}\left\{\frac{j}{n} - F(y_{(j)}), F(y_{(j)}) - \frac{(j-1)}{n}, j = 1, \ldots, n\right\} \end{aligned} \tag{11.4}$$

Equation (11.4) can be used to compute the value of D.

Suppose now that the Y_j are observed and their values are such that $D = d$. Since a large value of D would appear to be inconsistent with the null hypothesis that F is the underlying distribution, it follows that the p-value for this data set is given by

$$p\text{-value} = P_F\{D \geqslant d\}$$

where we have written P_F to make explicit that this probability is to be computed under the assumption that H_0 is correct (and so F is the underlying distribution).

The above p-value can be approximated by a simulation that is made easier by the following proposition, which shows that $P_F\{D \geqslant d\}$ does not depend on the underlying distribution F. This result enables us to estimate the p-value by doing the simulation with any continuous distribution F we choose [thus allowing us to use the uniform (0,1) distribution].

Proposition *$P_F\{D \geqslant d\}$ is the same for any continuous distribution F.*

Proof

$$\begin{aligned} P_F\{D \geqslant d\} &= P_F\left\{\underset{x}{\text{Maximum}}\left|\frac{\#i\colon Y_i \leqslant x}{n} - F(x)\right| \geqslant d\right\} \\ &= P_F\left\{\underset{x}{\text{Maximum}}\left|\frac{\#i\colon F(Y_i) \leqslant F(x)}{n} - F(x)\right| \geqslant d\right\} \\ &= P\left\{\underset{x}{\text{Maximum}}\left|\frac{\#i\colon U_i \leqslant F(x)}{n} - F(x)\right| \geqslant d\right\} \end{aligned}$$

where $U_1, \ldots, U_n$ are independent uniform (0, 1) random variables. The first equality follows because F is an increasing function and so $Y \leqslant x$ is equivalent to $F(Y) \leqslant F(x)$, and the second because of the result (whose proof is left as an exercise) that if Y has the continuous distribution F then the random variable $F(Y)$ is uniform on (0, 1).

Continuing the above, we see, by letting $y = F(x)$ and noting that as x ranges from $-\infty$ to $+\infty$, $F(x)$ ranges from 0 to 1, that

$$P_F\{D \geqslant d\} = P\left\{\underset{0 \leqslant y \leqslant 1}{\text{Maximum}}\left|\frac{\#i\colon U_i \leqslant y}{n} - y\right| \geqslant d\right\}$$

which shows that the distribution of D, when H_0 is true, does not depend on the actual distribution F. □

It follows from the preceding proposition that after the value of D is determined from the data, say, $D = d$, the p-value can be obtained by doing a simulation with the uniform (0, 1) distribution. That is, we generate a set of n random numbers $U_1, \ldots, U_n$ and then check whether or not the inequality

$$\underset{0 \leqslant y \leqslant 1}{\text{Maximum}} \left| \frac{\#i\colon U_i \leqslant y}{n} - y \right| \geqslant d \tag{11.5}$$

is valid. This is then repeated many times and the proportion of times that it is valid is our estimate of the p-value of the data set. As noted earlier, the left side of the inequality (11.5) can be computed by ordering the random numbers and then using the identity

$$\text{Max} \left| \frac{\#i\colon U_i \leqslant y}{n} - y \right| = \text{Max} \left\{ \frac{j}{n} - U_{(j)}, U_{(j)} - \frac{(j-1)}{n}, j = 1, \ldots, n \right\}$$

where $U_{(j)}$ is the jth smallest value of $U_1, \ldots, U_n$. For example, if $n = 3$ and $U_1 = 0.7$, $U_2 = 0.6$, $U_3 = 0.4$, then $U_{(1)} = 0.4$, $U_{(2)} = 0.6$, $U_{(3)} = 0.7$ and the value of D for this data set is

$$D = \text{Max} \left\{ \frac{1}{3} - 0.4, \frac{2}{3} - 0.6, 1 - 0.7, 0.4, 0.6 - \frac{1}{3}, 0.7 - \frac{2}{3} \right\} = 0.4$$

Example 11b Suppose we want to test the hypothesis that a given population distribution is exponential with mean 100; that is, $F(x) = 1 - e^{-x/100}$. If the (ordered) values from a sample of size 10 from this distribution are

$$66, 72, 81, 94, 112, 116, 124, 140, 145, 155$$

what conclusion can be drawn?

To answer the above, we first employ Equation (11.4) to compute the value of the Kolmogorov–Smirnov test quantity D. After some computation this gives the result $D = 0.4831487$. To obtain the approximate p-value we did a simulation which gave the following output:

```
RUN
THIS PROGRAM USES SIMULATION TO APPROXIMATE THE
   p-value OF THE KOLMOGOROV-SMIRNOV TEST
Random number seed (−32768 to 32767) ? 4567
ENTER THE VALUE OF THE TEST QUANTITY
? 0.4831487
ENTER THE SAMPLE SIZE
```

```
? 10
ENTER THE DESIRED NUMBER OF SIMULATION RUNS
? 500
THE APPROXIMATE p-value IS 0.012
OK
```

Because the p-value is so low (it is extremely unlikely that the smallest of a set of 10 values from the exponential distribution with mean 100 would be as large as 66), the hypothesis would be rejected. □

11.2 Goodness of Fit Tests When Some Parameters Are Unspecified

The Discrete Data Case

We can also perform a goodness of fit test of a null hypothesis that does not completely specify the probabilities $\{p_i, i = 1, \ldots, k\}$. For example, suppose we are interested in testing whether the daily number of traffic accidents in a certain region has a Poisson distribution with some unspecified mean. To test this hypothesis, suppose that data are obtained over n days and let Y_i represent the number of accidents on day i, for $i = 1, \ldots, n$. To determine whether these data are consistent with the assumption of an underlying Poisson distribution, we must first address the difficulty that, if the Poisson assumption is correct, these data can assume an infinite number of possible values. However, this is accomplished by breaking up the set of possible values into a finite number of, say, k regions and then seeing in which of the regions the n data points lie. For instance, if the geographical area of interest is small, and so there are not too many accidents in a day, we might say that the number of accidents in a given day falls in region $i, i = 1, 2, 3, 4, 5$, when there are $i - 1$ accidents on that day, and in region 6 when there are 5 or more accidents. Hence, if the underlying distribution is indeed Poisson with mean λ, then

$$p_i = P\{Y = i - 1\} = \frac{e^{-\lambda}\lambda^{i-1}}{(i-1)!}, \quad i = 1, 2, 3, 4, 5$$

$$p_6 = 1 - \sum_{j=0}^{4} \frac{e^{-\lambda}\lambda^j}{j!} \tag{11.6}$$

Another difficulty we face in obtaining a goodness of fit test of the hypothesis that the underlying distribution is Poisson is that the mean value λ is not specified. Now, the intuitive thing to do when λ is unspecified is clearly to estimate its value from the data—call $\hat{\lambda}$ the estimate—and then compute the value of the test statistic

$$T = \sum_{i=1}^{k} \frac{(N_i - n\hat{p}_i)^2}{n\hat{p}_i}$$

where N_i is the number of the Y_j that fall in region i, and where $\hat{p}_i$ is the estimated probability, under H_0, that Y_j falls in region i, $i = 1, \dots, k$, which is obtained by substituting $\hat{\lambda}$ for λ in the expression (11.6).

The above approach can be used whenever there are unspecified parameters in the null hypothesis that are needed to compute the quantities $p_i, i = 1, \dots, k$. Suppose now that there are m such unspecified parameters. It can be proved that, for reasonable estimators of these parameters, when n is large the test quantity T has, when H_0 is true, approximately a chi-square distribution with $k-1-m$ degrees of freedom. (In other words, one degree of freedom is lost for each parameter that needs to be estimated.)

If the test quantity takes on the value, say, $T = t$, then, using the above, the p-value can be approximated by

$$p\text{-value} \approx P\left\{X^2_{k-1-m} \geqslant t\right\}$$

where X^2_{k-1-m} is a chi-square random variable with $k-1-m$ degrees of freedom.

Example 11c Suppose that over a 30-day period there are 6 days in which no accidents occurred, 2 in which 1 accident occurred, 1 in which 2 accidents occurred, 9 in which 3 occurred, 7 in which 4 occurred, 4 in which 5 occurred, and 1 in which 8 occurred. To test whether these data are consistent with the hypothesis of an underlying Poisson distribution, note first that since there were a total of 87 accidents, the estimate of the mean of the Poisson distribution is

$$\hat{\lambda} = \frac{87}{30} = 2.9$$

Since the estimate of $P\{Y = i\}$ is thus $e^{-2.9}(2.9)^i/i!$, we obtain that with the six regions as given at the beginning of this section

$$\begin{aligned} &\hat{p}_1 = 0.0500, \quad &&\hat{p}_2 = 0.1596, \quad &&\hat{p}_3 = 0.2312, \\ &\hat{p}_4 = 0.2237, \quad &&\hat{p}_5 = 0.1622, \quad &&\hat{p}_6 = 0.1682 \end{aligned}$$

Using the data values $N_1 = 6$, $N_2 = 2$, $N_3 = 1$, $N_4 = 9$, $N_5 = 7$, $N_6 = 5$, we see that the value of the test statistic is

$$T = \sum_{i=1}^{6} \frac{(N_i - 30\hat{p}_i)^2}{30\hat{p}_i} = 19.887$$

To determine the p-value we run Program 9-1, which yields

$$p\text{-value} \approx P\left\{X^2_4 > 19.887\right\} = 0.0005$$

and so the hypothesis of an underlying Poisson distribution is rejected. □

We can also use simulation to estimate the p-value. However, since the null hypothesis no longer completely specifies the probability model, the use of simulation to determine the p-value of the test statistic is somewhat trickier than before. The way it should be done is as follows.

(a) *The Model*. Suppose that the null hypothesis is that the data values $Y_1, \ldots, Y_n$ constitute a random sample from a distribution that is specified up to a set of unknown parameters $\theta_1, \ldots, \theta_m$. Suppose also that when this hypothesis is true, the possible values of the Y_i are $1, \ldots, k$.

(b) *The Initial Step*. Use the data to estimate the unknown parameters. Specifically, let $\hat{\theta}_j$ denote the value of the estimator of θ_j, $j = 1, \ldots, m$. Now compute the value of the test statistic

$$T = \sum_{i=1}^{k} \frac{(N_i - n\hat{p}_i)^2}{n\hat{p}_i}$$

where N_i is the number of the data values that are equal to $i, i = 1, \ldots, k$, and $\hat{p}_i$ is the estimate of p_i that results when $\hat{\theta}_j$ is substituted for θ_j, for $j = 1, \ldots, m$. Let t denote the value of the test quantity T.

(c) *The Simulation Step*. We now do a series of simulations to estimate the p-value of the data. First note that all simulations are to be obtained by using the population distribution that results when the null hypothesis is true and θ_j is equal to its estimate $\hat{\theta}_j$, $j = 1, \ldots, m$, determined in step (b).

Simulate a sample of size n from the aforementioned population distribution and let $\hat{\theta}_j$ (sim) denote the estimate of θ_j, $j = 1, \ldots, m$, based on the simulated data. Now determine the value of

$$T_{\text{sim}} = \sum_{i=1}^{k} \frac{[N_i - n\hat{p}_i(\text{sim})]^2}{n\hat{p}_i(\text{sim})}$$

where N_i is the number of the simulated data values equal to $i, i = 1, \ldots, k$, and $\hat{p}_i$ (sim) is the value of p_i when θ_j is equal to $\hat{\theta}_j(\text{sim})$, $j = 1, \ldots, m$.

The simulation step should then be repeated many times. The estimate of the p-value is then equal to the proportion of the values of T_{sim} that are at least as large as t. □

Example 11d Let us reconsider Example 11c. The data presented in this example resulted in the estimate $\hat{\lambda} = 2.9$ and the test quantity value $T = 19.887$. The simulation step now consists of generating 30 independent Poisson random variables each having mean 2.9 and then computing the value of

$$T^* \equiv \sum_{i=1}^{6} \frac{(X_i - 30p_i^*)^2}{30p_i^*}$$

where X_i is the number of the 30 values that fall into region i, and p_i^* is the probability that a Poisson random variable with a mean equal to the average of the 30 generated values would fall into region i. This simulation step should be repeated many times, and the estimated p-value is the proportion of times it results in a T^* at least as large as 19.887. □

The Continuous Data Case

Now consider the situation where we want to test the hypothesis that the random variables $Y_1, \ldots, Y_n$ have the continuous distribution function $F_{\boldsymbol{\theta}}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ is a vector of unknown parameters. For example, we might be interested in testing that the Y_j come from a normally distributed population. To employ the Kolmogorov–Smirnov test we first use the data to estimate the parameter vector $\boldsymbol{\theta}$, say, by the vector of estimators $\hat{\boldsymbol{\theta}}$. The value of the test statistic D is now computed by

$$D = \underset{x}{\text{Maximum}} |F_e(x) - F_{\hat{\boldsymbol{\theta}}(x)}|$$

where $F_{\hat{\boldsymbol{\theta}}}$ is the distribution function obtained from $F_{\boldsymbol{\theta}}$ when $\boldsymbol{\theta}$ is estimated by $\hat{\boldsymbol{\theta}}$.

If the value of the test quantity is $D = d$, then the p-value can be *roughly* approximated by $P_{F_{\hat{\boldsymbol{\theta}}}}\{D \geqslant d\} = P_U\{D \geqslant d\}$. That is, after determining the value of D, a rough approximation, which actually overestimates the p-value, is obtained. If this does not result in a small estimate for the p-value, then, as the hypothesis is not going to be rejected, we might as well stop. However, if this estimated p-value is small, then a more accurate way of using simulation to estimate the true p-value is necessary. We now describe how this should be done.

STEP 1: Use the data to estimate $\boldsymbol{\theta}$, say, by $\hat{\boldsymbol{\theta}}$. Compute the value of D as described above.

STEP 2: All simulations are to be done using the distribution $F_{\hat{\boldsymbol{\theta}}}$. Generate a sample of size n from this distribution and let $\hat{\boldsymbol{\theta}}$ (sim) be the estimate of $\boldsymbol{\theta}$ based on this simulation run. Compute the value of

$$\underset{x}{\text{Maximum}} |F_{e,\text{sim}}(x) - F_{\hat{\boldsymbol{\theta}}(\text{sim})}(x)|$$

where $F_{e,\text{sim}}$ is the empirical distribution function of the simulated data; and note whether it is at least as large as d. Repeat this many times and use the proportion of times that this test quantity is at least as large as d as the estimate of the p-value.

11.3 The Two-Sample Problem

Suppose we have formulated a mathematical model for a service system which clears all its customers at the end of a day; moreover, suppose that our model assumes that each day is probabilistically alike in that the probability laws for successive days are identical and independent. Some of the individual assumptions of the model—such as, for example, that the service times are all independent with the common distribution G, or that the arrivals of customers constitute a Poisson process—can be individually tested by using the results of Sections 11.1 and 11.2. Suppose that none of these individual tests results in a particularly small p-value and so all the parts of the model, taken individually, do not appear to be inconsistent with the real data we have about the system. [We must be careful here in what we mean by a small p-value because, even if the model is correct, if we perform a large number of tests then, by chance, some of the resulting p-values may be small. For example, if we perform r separate tests on independent data, then the probability that at least one of the resulting p-values is as small as α is $1 - (1 - \alpha)^r$, which even for small α will become large as r increases.]

At this stage, however, we are still not justified in asserting that our model is correct and has been validated by the real data; for the totality of the model, including not only all the individual parts but also our assumptions about the ways in which these parts interact, may still be inaccurate. One way of testing the model in its entirety is to consider some random quantity that is a complicated function of the entire model. For example, we could consider the total amount of waiting time of all customers that enter the system on a given day. Suppose that we have observed the real system for m days and let $Y_i, i = 1, \ldots, m$, denote the sum of these waiting times for day i. If we now simulate the proposed mathematical model for n days, we can let $X_i, i = 1, \ldots, n$, be the sum of the waiting times of all customers arriving on the (simulated) day i. Since the mathematical model supposes that all days are probabilistically alike and independent, it follows that all the random variables $X_1, \ldots, X_m$ have some common distribution, which we denote by F. Now if the mathematical model is an accurate representation of the real system, then the real data $Y_1, \ldots, Y_m$ also have the distribution F. That is, if the mathematical model is accurate, one should not be able to tell the simulated data apart from the real data. From this it follows that one way of testing the accuracy of the model in its entirety is to test the null hypothesis H_0 that $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ are independent random variables having a common distribution. We now show how such a hypothesis can be tested.

Suppose we have two sets of data—$X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$—and we want to test the hypothesis H_0 that these $n + m$ random variables are all independent and identically distributed. This statistical hypothesis testing problem is called the two-sample problem.

To test H_0, order the $n + m$ values $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ and suppose for the time being that all $n + m$ values are distinct and so the ordering is unique. Now for $i = 1, \ldots, n$, let R_i denote the rank of X_i among the $n + m$ data values; that is, $R_i = j$ if X_i is the jth smallest among the $n + m$ values. The quantity

$$R = \sum_{i=1}^{n} R_i$$

equal to the sum of the ranks of the first data set, is used as our test quantity. (Either of the two data sets can be considered as the "first" set.)

If R is either very large (indicating that the first data set tends to be larger than the second) or very small (indicating the reverse), then this would be strong evidence against the null hypothesis. Specifically, if $R = r$, we reject the null hypothesis if either

$$P_{H_0}\{R \leqslant r\} \quad \text{or} \quad P_{H_0}\{R \geqslant r\}$$

is very low. Indeed, the p-value of the test data which results in $R = r$ is given by

$$p\text{-value} = 2\ \text{Minimum}(P_{H_0}\{R \leqslant r\}, P_{H_0}\{R \geqslant r\}) \tag{11.7}$$

[It is twice the minimum of the probabilities because we reject either if R is too small or too large. For example, suppose $r*$ and r^* were such that the probability, under H_0, of obtaining a value less (greater) than or equal to $r * (r^*)$ is 0.05. Since the probability of either event occurring is, under H_0, 0.1 it follows that if the outcome is $r*$ (or r^*) the p-value is 0.1.]

The hypothesis test resulting from the above p-value—that is, the test that calls for rejection of the null hypothesis when the p-value is sufficiently small—is called the *two-sample rank sum test*. (Other names that have also been used to designate this test are the Wilcoxon two-sample test and the Mann–Whitney two-sample test.)

Example 11e Suppose that direct observation of a system over 5 days has yielded that a certain quantity has taken on the successive values

$$342, 448, 504, 361, 453$$

whereas a 10-day simulation of a mathematical model proposed for the system has resulted in the following values:

$$186, 220, 225, 456, 276, 199, 371, 426, 242, 311$$

Because the five data values from the first set have ranks 8, 12, 15, 9, 13, it follows that the value of the test quantity is $R = 57$. □

We can explicitly compute the p-value given in Equation (11.7) when n and m are not too large and all the data are distinct. To do so let

$$P_{n,m}(r) = P_{H_0}\{R \leqslant r\}$$

Hence $P_{n,m}(r)$ is the probability that from two identically distributed data sets of sizes n and m, the sum of the ranks of the data values from the first set is less than or equal to r. We can obtain a recursive equation for these probabilities by conditioning on whether the largest data value comes from the first or the second set. If the largest value is indeed contained in the first data set, the sum of the ranks of this set equals $n + m$ (the rank of the largest value) plus the sum of the ranks of the other $n-1$ values from this set when considered along with the m values from the other set. Hence, when the largest is contained in the first data set, the sum of the ranks of that set is less than or equal to r if the sum of the ranks of the remaining $n-1$ elements is less than or equal to $r-n-m$, and this is true with probability $P_{n-1,m}(r-n-m)$. By a similar argument we can show that if the largest value is contained in the second set, the sum of the ranks of the first set is less than or equal to r with probability $P_{n,m-1}(r)$. Finally, since the largest value is equally likely to be any of the $n+m$ values, it follows that it is a member of the first set with probability $n/(n+m)$. Putting this together yields the following recursive equation:

$$P_{n,m}(r) = \frac{n}{n+m}P_{n-1,m}(r-n-m) + \frac{m}{n+m}P_{n,m-1}(r) \tag{11.8}$$

Starting with the boundary conditions

$$P_{1,0}(k) = \begin{cases} 0, & k \leqslant 0 \\ 1, & k > 0 \end{cases} \qquad \text{and} \qquad P_{0,1}(k) = \begin{cases} 0, & k < 0 \\ 1, & k \geqslant 0 \end{cases}$$

Equation (11.8) can be recursively solved to obtain $P_{n,m}(r) = P_{H_0}\{R \leqslant r\}$ and $P_{n,m}(r-1) = 1 - P_{H_0}\{R \geqslant r\}$.

Example 11f Five days of observation of a system yielded the following values of a certain quantity of interest:

$$132, 104, 162, 171, 129$$

A 10-day simulation of a proposed model of this system yielded the values

$$107, 94, 136, 99, 114, 122, 108, 130, 106, 88$$

Suppose the formulated model implies that these daily values should be independent and have a common distribution. To determine the p-value that results from the above data, note first that R, the sum of the ranks of the first sample, is

$$R = 12 + 4 + 14 + 15 + 10 = 55$$

A program using the recursion (11.8) yielded the following output:

```
THIS PROGRAM COMPUTES THE p-value FOR THE TWO-SAMPLE
   RANK SUM TEST
THIS PROGRAM WILL RUN FASTEST IF YOU DESIGNATE AS THE
   FIRST
  SAMPLE THE SAMPLE HAVING THE SMALLER SUM OF RANKS
ENTER THE SIZE OF THE FIRST SAMPLE
? 5
```

```
ENTER THE SIZE OF THE SECOND SAMPLE
? 10
ENTER THE SUM OF THE RANKS OF THE FIRST SAMPLE
? 55
The p-value IS 0.0752579
OK
```

□

The difficulty with employing the recursion (11.8) to compute the p-value is that the amount of computation needed grows enormously as the sample sizes increase. For example, if $n = m = 20$, even if we choose the test quantity to be the smaller sum of ranks, then since the sum of all the ranks is $1 + 2 + \cdots + 40 = 820$, it is possible that the test statistic could have a value as large as 410. Hence, there can be as many as $20 \times 20 \times 410 = 164,000$ values of $P_{n,m}(r)$ that would have to be computed to determine the p-value. Thus, for large samples, the use of the recursion provided by (11.8) may not be viable. Two different approximation methods that can be used in such cases are (a) a classical approach based on approximating the distribution of R and (b) simulation.

To use the classical approach for approximating the p-value we make use of the fact that under H_0 all possible orderings of the $n + m$ values are equally likely. Using this fact it is easy to show that

$$E_{H_0}[R] = n\frac{(n+m+1)}{2}$$
$$\text{Var}_{H_0}(R) = nm\frac{(n+m+1)}{12}$$

Now it can be shown that, under H_0, when n and m are large, R is approximately normally distributed. Hence, when H_0 is true,

$$\frac{R - n(n+m+1)/2}{\sqrt{nm(n+m+1)/12}} \text{ is approximately a standard normal.}$$

Because for a normal random variable W, the minimum of $P\{W \leqslant r\}$ and $P\{W \geqslant r\}$ is the former when $r \leqslant E[W]$, and the latter otherwise, it follows that when n and m are not too small (both being greater than 7 should suffice), we can approximate the p-value of the test result $R = r$ by

$$p\text{-value} \approx \begin{cases} 2\ P\{Z < r^*\} & \text{if } r \leqslant n\dfrac{(n+m+1)}{2} \\ 2\ P\{Z > r^*\} & \text{otherwise} \end{cases} \tag{11.9}$$

where

$$r^* = \frac{r - \dfrac{n(n+m+1)}{2}}{\sqrt{\dfrac{nm(n+m+1)}{12}}}$$

and where Z is a standard normal random variable.

Example 11g Let us see how well the classical approximation works for the data of Example 11g. In this case, since $n = 5$ and $m = 10$, we have that

$$\begin{aligned} p\text{-value} &= 2\ P_{H_0}\{R \geqslant 55\} \\ &\approx 2\ P\left\{Z \geqslant \frac{55 - 40}{\sqrt{\frac{50\times 16}{12}}}\right\} \\ &= 2\ P\{Z \geqslant 1.8371\} \\ &= 0.066 \end{aligned}$$

which should be compared with the exact answer 0.075. □

The p-value of the two-sample rank test can also be approximated by simulation. To see how this is accomplished, recall that if the observed value of the test quantity R is $R = r$, then the p-value is given by

$$p\text{-value} = 2\ \text{Minimum}(P_{H_0}\{R \geqslant r\}, P_{H_0}\{R \leqslant r\})$$

Now, under H_0, provided that all the $n + m$ data values are distinct, it follows that all orderings among these data values are equally likely, and thus the ranks of the first data set of size n have the same distribution as a random selection of n of the values $1, 2, \ldots, n + m$. Thus, under H_0, the probability distribution of R can be approximated by continually simulating a random subset of n of the integers $1, 2, \ldots, n + m$ and determining the sum of the elements in the subset. The value of $P_{H_0}\{R \leqslant r\}$ can be approximated by the proportion of simulations that result in a sum less than or equal to r, and the value of $P_{H_0}\{R \geqslant r\}$ by the proportion of simulations that result in a sum greater than or equal to r.

The above analysis supposes that all the $n + m$ data values are distinct. When certain of the values have a common value, one should take as the rank of a datum value the average of the ranks of the values equal to it. For example, if the first data set is 2, 3, 4 and the second 3, 5, 7, then the sum of the ranks of the first set is $1 + 2.5 + 4 = 7.5$. The p-value should be approximated by using the normal approximation via Equation (11.9).

A generalization of the two-sample problem is the multisample problem, where one has the following m data sets:

$$\begin{matrix} X_{1,1}, & X_{1,2}, & \ldots, & X_{1,n_1} \\ X_{2,1}, & X_{2,2}, & \ldots, & X_{2,n_2} \\ \vdots & \vdots & \vdots & \vdots \\ X_{m,1}, & X_{m,2}, & \ldots, & X_{m,n_m} \end{matrix}$$

and we are interested in testing the null hypothesis H_0 that all the $n = \sum_{i=1}^{m} n_i$ random variables are independent and have a common distribution.

A generalization of the two-sample rank test, called the multisample rank test (or often referred to as the Kruskal–Wallis test), is obtained by first ranking all the n data values. Then let $R_i, i = 1, \dots, m$, denote the sum of the ranks of all the n_i data values from the ith set. (Note that with this notation R_i is a sum of ranks and not an individual rank as previously.) Since, under H_0, all orderings are equally likely (provided all the data values are distinct), it follows exactly as before that

$$E[R_i] = n_i \frac{(n+1)}{2}$$

Using the above, the multisample rank sum test is based on the test quantity

$$R = \frac{12}{n(n+1)} \sum_{i=1}^{m} \frac{[R_i - n_i(n+1)/2]^2}{n_i}$$

Since small values of R indicate a good fit to H_0, the test based on the quantity R rejects H_0 for sufficiently large values of R. Indeed, if the observed value of R is $R = y$, the p-value of this result is given by

$$p\text{-value} = P_{H_0}\{R \geqslant y\}$$

This value can be approximated by using the result that for large values of $n_1, \dots, n_m$, R has approximately a chi-square distribution with $m - 1$ degrees of freedom [this latter result being the reason why we include the term $12/n(n+1)$ in the definition of R]. Hence, if $R = y$,

$$p\text{-value} \approx P\left\{\chi^2_{m-1} \geqslant y\right\}$$

Simulation can also be used to evaluate the p-value (see Exercise 14).

Even when the data values are not all distinct, the above approximation for the p-value should be used. In computing the value of R the rank of an individual datum value should be, as before, the average of all the ranks of the data equal to it.

11.4 Validating the Assumption of a Nonhomogeneous Poisson Process

Consider a mathematical model which supposes that the daily arrivals to a system occur in accordance with a nonhomogeneous Poisson process, with the arrival process from day to day being independent and having a common, but unspecified, intensity function.

To validate such an assumption, suppose that we observe the system over r days, noting the arrival times. Let $N_i, i = 1, \dots, r$, denote the number of arrivals on day i, and note that if the arrival process is indeed a nonhomogeneous Poisson process, then these quantities are independent Poisson random variables with the same mean. Now whereas this consequence could be tested by using the goodness of fit approach, as is done in Example 11a, we present an alternative approach that is sometimes more efficient. This alternative approach is based on the fact that the mean and variance of a Poisson random variable are equal. Hence, if the N_i are indeed a sample from a Poisson distribution, the sample mean

$$\overline{N} = \sum_{i=1}^{r} \frac{N_i}{r}$$

and the sample variance

$$S^2 = \sum_{i=1}^{r} \frac{(N_i - \overline{N})^2}{r - 1}$$

should be roughly equal. Motivated by this, we base our test of the hypothesis

$H_0 : N_i$ are independent Poisson random variables with a common mean on the test quantity

$$T = \frac{S^2}{\overline{N}} \tag{11.10}$$

Because either a very small or very large value of T would be inconsistent with H_0, the p-value for the outcome $T = t$ would be

$$p\text{-value} = 2 \text{ Minimum}(P_{H_0}\{T \leqslant t\}, P_{H_0}\{T \geqslant t\})$$

However, since H_0 does not specify the mean of the Poisson distribution, we cannot immediately compute the above probabilities; rather, we must first use the observed data to estimate the mean. By using the estimator $\overline{N}$, it follows that if the observed value of $\overline{N}$ is $\overline{N} = m$, the p-value can be approximated by

$$p\text{-value} \approx 2 \text{ Minimum}(P_m\{T \leqslant t\}, P_m\{T \geqslant t\})$$

where T is defined by Equation (11.10) with $N_1, \ldots, N_r$ being independent Poisson random variables each with mean m. We can now approximate $P_m\{T \leqslant t\}$ and $P_m\{T \geqslant t\}$ via a simulation. That is, we continually generate r independent Poisson random variables with mean m and compute the resulting value of T. The proportion of these for which $T \leqslant t$ is our estimate of $P\{T \leqslant t\}$, and the proportion for which $T \geqslant t$ is our estimate of $P\{T \geqslant t\}$.

If the above p-value is quite small, we reject the null hypothesis that the daily arrivals constitute a nonhomogeneous Poisson process. However, if the p-value is not small, this only implies that the assumption that the number of arrivals each day has a Poisson distribution is a viable assumption and does not by itself validate the stronger assumption that the actual arrival pattern (as determined by the nonhomogeneous intensity function) is the same from day to day. To complete our validation we must now consider the actual arrival times for each of the r days observed. Suppose that the arrival times on day j, $j = 1, \ldots, r$, are $X_{j,1}, X_{j,2}, \ldots, X_{j,N_j}$. Now if the arrival process is indeed a nonhomogeneous Poisson process, it can be shown that each of these r sets of arrival times constitutes a sample from a common distribution. That is, under the null hypothesis, the r sets of data $X_{j,1}, \ldots, X_{j,N_j}$, $j = 1, \ldots, r$, are all independent random variables from a common distribution.

The above consequence, however, can be tested by the multisample rank test given in Section 11.3. That is, first rank all the $N \equiv \sum_{j=1}^{r} N_j$ data values, and then let R_j denote the sum of the ranks of all the N_j data values from the jth set. The test quantity

$$R = \frac{12}{N(N+1)} \sum_{j=1}^{r} \frac{\left(R_j - N_j \dfrac{(N+1)}{2}\right)^2}{N_j}$$

can now be employed by using the fact that, when H_0 is true, R has approximately a chi-square distribution with $r - 1$ degrees of freedom. Hence, if the observed value of R is $R = y$, the resulting p-value can be approximated by

$$\begin{aligned} p\text{-value} &= 2\ \text{Minimum}(P_{H_0}\{R \leqslant y\}, P_{H_0}\{R \geqslant y\}) \\ &\approx 2\ \text{Minimum}\left(P\left\{X_{r-1}^2 \leqslant y\right\}, 1 - P\left\{X_{r-1}^2 \leqslant y\right\}\right) \end{aligned}$$

where X_{r-1}^2 is a chi-square random variable with $r - 1$ degrees of freedom. (Of course, we could also approximate the p-value by a simulation.) If the above p-value, along with the previous p-value considered, is not too small, we may conclude that the data are not inconsistent with our assumption that daily arrivals constitute a nonhomogeneous Poisson process.

A Technical Remark Many readers may wonder why we used a two-sided region to calculate the p-value in (11.11), rather than the one-sided region used in the multisample rank sum test. It is because a multisample rank sum test *assumes* that the data come from m distributions, and, because R is small when these distributions are equal, a p-value based on a one-sided probability is appropriate. However, in testing for a periodic nonhomogeneous Poisson process, we want to test both that the arrival times on day i come from some distribution and that this distribution is the same for all i. That is, we do not start by assuming, as is done in the rank sum test, that we have data from a fixed number of separate distributions. Consequently, a two-sided test is appropriate, because a very small value of R might be indicative of some pattern of arrivals during a day, i.e., even though the number of arrivals each day might have the same Poisson distribution, the daily arrival times might not be independent and identically distributed. □

Example 11h Suppose that the daily times at which deliveries are made at a certain plant are noted over 5 days. During this time the numbers of deliveries during each of the days are as follows:

$$18, 24, 16, 19, 25$$

Suppose also that when the 102 delivery times are ranked according to the time of day they arrived, the sums of the ranks of the deliveries from each day are

$$1010, 960, 1180, 985, 1118$$

Using the above data, let us test the hypothesis that the daily arrival process of deliveries is a nonhomogeneous Poisson process.

We first test that the first data set of the daily number of deliveries consists of a set of five independent and identically distributed Poisson random variables. Now the sample mean and sample variance are equal to

$$\overline{N} = 20.4 \quad \text{and} \quad S^2 = 15.3$$

and so the value of the test quantity is $T = 0.75$. To determine the approximate p-value of the test that the N_i are independent Poisson random variables, we then simulated 500 sets of five Poisson random variables with mean 20.4 and then computed the resulting value of $T = S^2/\overline{N}$. The output of this simulation indicated a p-value of approximately 0.84, and so it is clear that the assumption that the numbers of daily deliveries are independent Poisson random variables having a common mean is consistent with the data.

To continue our test of the null hypothesis of a nonhomogeneous Poisson process, we compute the value of the test quantity R, which is seen to be equal to 14.425. Because the probability that a chi-square random variable with four degrees of freedom is as large as 14.425 is 0.006, it follows that the p-value is 0.012, For such a small p-value we must reject the null hypothesis. □

If we wanted to test the assumption that a daily arrival process constituted a *homogeneous* Poisson process, we would proceed as before and first test the hypothesis that the numbers of arrivals each day are independent and identically distributed Poisson random variables. If the hypothesis remains plausible after we perform this test, we again continue as in the nonhomogeneous case by considering the actual set of $N = \sum_{j=1}^{r} N_j$ arrival times. However, we now use the result that under a homogeneous Poisson process, given the number of arrivals in a day, the arrival times are independently and uniformly distributed over $(0,T)$, where T is the length of a day. This consequence, however, can be tested by the Kolmogorov–Smirnov goodness of fit test presented in Section 11.1. That is, if the arrivals constitute a homogeneous Poisson process, the N random variables $X_{j,i}$, $i = 1, \ldots, N_j$, $j = 1, \ldots, r$, where $X_{j,i}$ represents the ith arrival time on day j, can be regarded as constituting a set of N independent and uniformly distributed random variables over $(0,T)$. Hence, if we define the empirical distribution function F_e by letting $F_e(x)$ be the proportion of the N data values that are less than or equal to x—that is,

$$F_e(x) = \sum_{j=1}^{r}\sum_{i=1}^{N_j} \frac{I_{j,i}}{N}$$

where

$$I_{j,i} = \begin{cases} 1 \text{ if } X_{j,i} \leqslant x \\ 0 \text{ otherwise} \end{cases}$$

then the value of the test quantity is

$$D = \underset{0 \leqslant x \leqslant T}{\text{Maximum}} \left| F_e(x) - \frac{x}{T} \right|$$

Once the value of the test statistic D is determined, we can then find the resulting p-value by simulation, as is shown in Section 11.1.

If the hypothesis of a nonhomogeneous Poisson process is shown to be consistent with the data, we face the problem of estimating the intensity function $\lambda(t)$, $0 \leqslant t \leqslant T$, of this process. [In the homogeneous case the obvious estimator is $\lambda(t) = \hat{\lambda}/T$, where $\hat{\lambda}$ is the estimate of the mean number of arrivals in a day of length T.] To estimate the intensity function, order the $N = \sum_{j=1}^{r} N_j$ daily arrival times. Let $y_0 = 0$, and for $k = 1, \ldots, N$, let y_k denote the kth smallest of these N arrival times. Because there has been a total of 1 arrival over r days within the time interval (y_{k-1}, y_k), $k = 1, \ldots, N$, a reasonable estimate of $\lambda(t)$ would be

$$\hat{\lambda}(t) = \frac{1}{r(y_k - y_{k-1})} \quad \text{for } y_{k-1} < t < y_k$$

[To understand the above estimator, note that if $\hat{\lambda}(t)$ were the intensity function, the expected number of daily arrivals that occur at a time point t such that $y_{k-1} < t \leqslant y_k$ would be given by

$$E[N(y_k) - N(y_{k-1})] = \int_{y_{k-1}}^{y_k} \hat{\lambda}(t)\, dt = \frac{1}{r}$$

and hence the expected number of arrivals within that interval over r days would be 1, which coincides with the actual observed number of arrivals in that interval.]

Exercises

1. According to the Mendelian theory of genetics, a certain garden pea plant should produce white, pink, or red flowers, with respective probabilities $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$. To test this theory a sample of 564 peas was studied with the result that 141 produced white, 291 produced pink, and 132 produced red flowers. Approximate the p-value of this data set

(a) by using the chi-square approximation, and
(b) by using a simulation.

2. To ascertain whether a certain die was fair, 1000 rolls of the die were recorded, with the result that the numbers of times the die landed $i, i = 1, 2, 3, 4, 5, 6$ were, respectively, 158, 172, 164, 181, 160, 165. Approximate the p-value of the test that the die was fair

(a) by using the chi-square approximation, and
(b) by using a simulation.

3. Approximate the p-value of the hypothesis that the following 10 values are random numbers: 0.12, 0.18, 0.06, 0.33, 0.72, 0.83, 0.36, 0.27, 0.77, 0.74.

4. Approximate the p-value of the hypothesis that the following data set of 14 points is a sample from a uniform distribution over (50, 200):

$$164, 142, 110, 153, 103, 52, 174, 88, 178, 184, 58, 62, 132, 128$$

5. Approximate the p-value of the hypothesis that the following 13 data values come from an exponential distribution with mean 50:

$$86, 133, 75, 22, 11, 144, 78, 122, 8, 146, 33, 41, 99$$

6. Approximate the p-value of the test that the following data come from a binomial distribution with parameters (8, p), where p is unknown:

$$6, 7, 3, 4, 7, 3, 7, 2, 6, 3, 7, 8, 2\ 1, 3, 5, 8, 7$$

7. Approximate the p-value of the test that the following data set comes from an exponentially distributed population: 122, 133, 106, 128, 135, 126.

8. To generate the ordered values of n random numbers we could generate n random numbers and then order, or sort, them. Another approach makes use of the result that given that the $(n + 1)$st event of a Poisson process occurs at time t, the first n event times are distributed as the set of ordered values of n uniform $(0, t)$ random variables. Using this result, explain why, in the following algorithm, $y_1, \ldots, y_n$ denote the ordered values of n random numbers.

$$\text{Generate } n + 1 \text{ random numbers } U_1, \ldots, U_{n+1}$$
$$X_i = -\log U_i, \qquad i = 1, \ldots, n + 1$$
$$t = \sum_{i=1}^{n+1} X_i, \qquad c = \frac{1}{t}$$
$$y_i = y_{i-1} + cX_i, \qquad i = 1, \ldots, n \text{ (with } y_0 = 0)$$

9. Let $N_1, \ldots, N_k$ have a multinomial distribution with parameters n, $p_1, \ldots, p_k, \sum_{i=1}^{k} p_i = 1$. With

$$T = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$$

suppose we want to use simulation to estimate $P(T > t)$. To reduce the variance of the estimator what might be used as a control variable?

10. Suggest a variance reduction technique when using simulation to estimate $P(D > d)$ where D is the Kolmogorov-Smirnov statistic.

11. In Exercise 10, compute the approximate p-value based on

(a) the normal approximation, and
(b) a simulation.

12. Fourteen cities, of roughly equal size, are chosen for a traffic safety study. Seven of them are randomly chosen, and in these cities a series of newspaper articles dealing with traffic safety are run over a 1-month period. The numbers of traffic accidents reported in the month following this campaign are as follows:

Treatment group:	19	31	39	45	47	66	75
Control group:	28	36	44	49	52	72	72

Determine the exact p-value when testing the hypothesis that the articles have not had any effect.

13. Approximate the p-value in Exercise 12

(a) by using the normal approximation, and
(b) by using a simulation.

14. Explain how simulation can be employed to approximate the p-value in the multisample problem—that is, when testing that a set of m samples all come from the same probability distribution.

15. Consider the following data resulting from three samples:
Compute the approximate p-value of the test that all the data come from a single probability distribution

Sample 1:	121	144	158	169	194	211	242
Sample 2:	99	128	165	193	242	265	302
Sample 3:	129	134	137	143	152	159	170

(a) by using the chi-square approximation, and
(b) by using a simulation.

16. The number of daily arrivals over an 8-day interval are as follows:

122, 118, 120, 116, 125, 119, 124, 130

Do you think the daily arrivals could be independent and identically distributed as nonhomogeneous Poisson processes?

17. Over an interval of length 100 there have been 18 arrivals at the following times:

12, 20, 33, 44, 55, 56, 61, 63, 66, 70, 73, 75, 78, 80, 82, 85, 87, 90

Approximate the p-value of the test that the arrival process is a (homogeneous) Poisson process.

Bibliography

Diaconis, P., and B. Efron, “Computer Intensive Methods in Statistics,” *Sci. Am.*, **248**(5), 96–109, 1983.

Fishman, G. S., *Concepts and Methods in Discrete Event Digital Simulations*. Wiley, New York, 1973.

Kendall, M., and A. Stuart, *The Advanced Theory of Statistics*, 4th ed. MacMillan, New York, 1979.

Mihram, G. A., *Simulation—Statistical Foundations and Methodology*. Academic Press, New York, 1972.

Sargent, R. G., “A Tutorial on Validation and Verification of Simulation Models,” *Proc. 1988 Winter Simulation Conf.*, San Diego, pp. 33–39, 1988.

Schruben, L. W., “Establishing the Credibility of Simulations,” *Simulation*, **34**, 101–105, 1980.