9

Input Modeling

Input data provide the driving force for a simulation model. In the simulation of a queueing system, typical input data are the distributions of time between arrivals and service times. For an inventory system simulation, input data include the distributions of demand and lead time. For the simulation of a reliability system, the distribution of time-to-failure of a component is an example of input data.

In the examples and exercises in Chapters 2 and 3, the appropriate distributions were specified for you. In real-world simulation applications, however, determining appropriate distributions for input data is a major task from the standpoint of time and resource requirements. Regardless of the sophistication of the analyst, faulty models of the inputs will lead to outputs whose interpretation may give rise to misleading recommendations.

There are four steps in the development of a useful model of input data:

1. Collect data from the real system of interest. This often requires a substantial time and resource commitment. Unfortunately, in some situations it is not possible to collect data (for example, when time is extremely limited, when the input process does not yet exist, or when laws or rules prohibit the collection of data). When data are not available, expert opinion and knowledge of the process must be used to make educated guesses.
2. Identify a probability distribution to represent the input process. When data are available, this step typically begins by developing a frequency distribution, or histogram, of the data. Based on the frequency distribution and structural knowledge of the process, a family of distributions is chosen. Fortunately, as described in Chapter 5, several well-known distributions often provide good approximations in practice.

3. Choose parameters that determine a specific instance of the distribution family. When data are available, these parameters may be estimated from the data.
4. Evaluate the chosen distribution and the associated parameters for goodness-of-fit. Goodness-of-fit may be evaluated informally via graphical methods, or formally via statistical tests. The chi-square and the Kolmogorov-Smirnov tests are standard goodness-of-fit tests. If not satisfied that the chosen distribution is a good approximation of the data, then the analyst returns to the second step, chooses a different family of distributions, and repeats the procedure. If several iterations of this procedure fail to yield a fit between an assumed distributional form and the collected data, the empirical form of the distribution may be used as described in Section 8.1.5 of the previous chapter.

Each of these steps is discussed in this chapter. Although software is now widely available to accomplish steps 2, 3 and 4 — including standalone programs such as ExpertFit®and Stat::Fit™, and integrated programs such as Arena's Input Processor and @Risk's BestFit — it is still important to understand what the software does so that it can be used appropriately. Unfortunately, software is not as readily available for input modeling when there is a relationship between two or more variables of interest, or when no data are available. These two topics are discussed toward the end of the chapter.

9.1 Data Collection

Problems are found at the end of each chapter, as exercises for the reader, in mathematics, physics, chemistry, and other technical subject texts. Years and years of working these problems may give the reader the impression that data are readily available. Nothing could be further from the truth. Data collection is one of the biggest tasks in solving a real problem. It is one of the most important and difficult problems in simulation. And even when data are available, they have rarely been recorded in a form that is directly useful for simulation input modeling.

"GIGO," or "garbage-in, garbage-out," is a basic concept in computer science and it applies equally in the area of discrete system simulation. Many are fooled by a pile of computer output or a sophisticated animation, as if these were the absolute truth. Even if the model structure is valid, if the input data are inaccurately collected, inappropriately analyzed, or not representative of the environment, the simulation output data will be misleading and possibly damaging or costly when used for policy or decision making.

EXAMPLE 9.1 (The Laundromat)

As budding simulation students, the first two authors had assignments to simulate the operation of an ongoing system. One of these systems, which seemed to be a rather simple operation, was a self-service laundromat with 10 washing machines and six dryers.

However, the data-collection aspect of the problem rapidly became rather enormous. The interarrival-time distribution was not homogeneous; that is, the distribution changed by time of day and by day of week. The laundromat was open 7 days a week for 16 hours per day, or 112 hours per week. It would have been impossible to cover the operation of the laundromat with the limited resources available (two students who were also taking four other courses) and with a tight time constraint (the simulation was to be completed in a 4-week period). Additionally, the distribution of time between arrivals during one week may not have been followed during the next week. As a compromise, a sample of times was selected, and the interarrival-time distributions were determined and classified according to arrival rate (perhaps inappropriately) as "high," "medium," and "low."

Service-time distributions also presented a difficult problem from many perspectives. The proportion of customers demanding the various service combinations had to be observed and recorded. The simplest case was the customer desiring one washer followed by one dryer. However, a customer might choose two washing machines followed by one dryer, one dryer only, and so on. Since the customers used numbered machines, it was possible to follow them using that reference, rather than remembering them by personal characteristics. Because of the dependence between washer demand and dryer demand for an individual customer, it would have been inappropriate to treat the service times for washers and dryers separately as independent variables.

Some customers waited patiently for their clothes to complete the washing or drying cycle, and then they removed their clothes promptly. Others left the premises and returned after their clothes had finished their cycle on the machine being used. In a very busy period, the manager would remove a customer's clothes after the cycle and set them aside in a basket. It was decided that service termination would be measured as the point in time when the machine was emptied of its contents.

Also, machines would break down from time to time. The length of the breakdown varied from a few moments, when the manager repaired the machine, to several days (a breakdown on Friday night, requiring a part not in the laundromat storeroom, would not be fixed until the following Monday). The short-term repair times were recorded by the student team. The long-term repair completion times were estimated by the manager. Breakdowns then became part of the simulation. ◀

Many lessons can be learned from an actual experience in data collection. The first five exercises at the end of this chapter suggest some situations in which the student can gain such experience.

The following suggestions may enhance and facilitate data collection, although they are not all-inclusive.

1. A useful expenditure of time is in planning. This could begin by a practice or preobserving session. Try to collect data while preobserving. Devise forms for this purpose. It is very likely that these forms will have to be

modified several times before the actual data collection begins. Watch for unusual circumstances and consider how they will be handled. When possible, videotape the system and extract the data later by viewing the tape. Planning is important even if data will be collected automatically (e.g., via computer data collection) to insure that the appropriate data are available. When data have already been collected by someone else, be sure to allow plenty of time for converting the data into a usable format.

2. Try to analyze the data as they are being collected. Determine if the data being collected are adequate to provide the distributions needed as input to the simulation. Determine if any data being collected are useless to the simulation. There is no need to collect superfluous data.
3. Try to combine homogeneous data sets. Check data for homogeneity in successive time periods and during the same time period on successive days. For example, check for homogeneity of data from 2:00 P.M. to 3:00 P.M. and 3:00 P.M. to 4:00 P.M., and check to see if the data are homogeneous for 2:00 P.M. to 3:00 P.M. on Thursday and Friday. When checking for homogeneity, an initial test is to see if the means of the distributions (the average interarrival times, for example) are the same. The two-sample t test can be used for this purpose. A more thorough analysis would require a determination of the equivalence of the distributions using, perhaps, a quantile-quantile plot (described later).
4. Be aware of the possibility of data censoring, in which a quantity of interest is not observed in its entirety. This problem most often occurs when the analyst is interested in the time required to complete some process (for example, produce a part, treat a patient, or have a component fail), but the process begins prior to, or finishes after the completion of, the observation period. Censoring can result in especially long process times being left out of the data sample.
5. To determine whether there is a relationship between two variables, build a scatter diagram. Sometimes an eyeball scan of the scatter diagram will indicate if there is a relationship between two variables of interest. Section 9.6 describes models for statistically dependent input data.
6. Consider the possibility that a sequence of observations which appear to be independent may possess autocorrelation. Autocorrelation may exist in successive time periods or for successive customers. For example, the service time for the ith customer may be related to the service time for the $(i+n)$th customer. A brief introduction to autocorrelation was provided in Section 7.4.3, and some input models that account for autocorrelation are presented in Section 9.6.
7. Keep in mind the difference between input data and output or performance data, and be sure to collect input data. Input data typically represent the uncertain quantities that are largely beyond the control of the system and will not be altered by changes made to improve the system. Output data, on the other hand, represent the performance of the system when subjected to the inputs, performance that we may be trying to

improve. In a queueing simulation, the customer arrival times are usually inputs, while the customer delay is an output. Performance data are useful for model validation, however (see Chapter 10).

Again, these are just a few suggestions. As a rule, data collection and analysis must be approached with great care.

9.2 Identifying the Distribution with Data

In this section we discuss methods for selecting families of input distributions when data are available. The specific distribution within a family is specified by estimating its parameters, as described in Section 9.3. Section 9.5 takes up the case when no data are available.

9.2.1 Histograms

A frequency distribution or histogram is useful in identifying the shape of a distribution. A histogram is constructed as follows:

1. Divide the range of the data into intervals (intervals are usually of equal width; however, unequal widths may be used if the heights of the frequencies are adjusted).
2. Label the horizontal axis to conform to the intervals selected.
3. Determine the frequency of occurrences within each interval.
4. Label the vertical axis so that the total occurrences can be plotted for each interval.
5. Plot the frequencies on the vertical axis.

The number of class intervals depends on the number of observations and the amount of scatter or dispersion in the data. Hines and Montgomery [1990] state that choosing the number of class intervals approximately equal to the square root of the sample size often works well in practice. If the intervals are too wide, the histogram will be coarse, or blocky, and its shape and other details will not show well. If the intervals are too narrow, the histogram will be ragged and will not smooth the data. Examples of a ragged, coarse, and appropriate histogram using the same data are shown in Figure 9.1. Modern data-analysis software often allows the interval sizes to be changed easily and interactively until a good choice is found.

The histogram for continuous data corresponds to the probability density function of a theoretical distribution. If continuous, a line drawn through the center point of each class interval frequency should result in a shape like that of a pdf.

Histograms for discrete data, where there are a large number of data points, should have a cell for each value in the range of the data. However, if there are few data points, it may be necessary to combine adjacent cells to eliminate the ragged appearance of the histogram. If the histogram is associated with discrete data, it should look like a probability mass function.

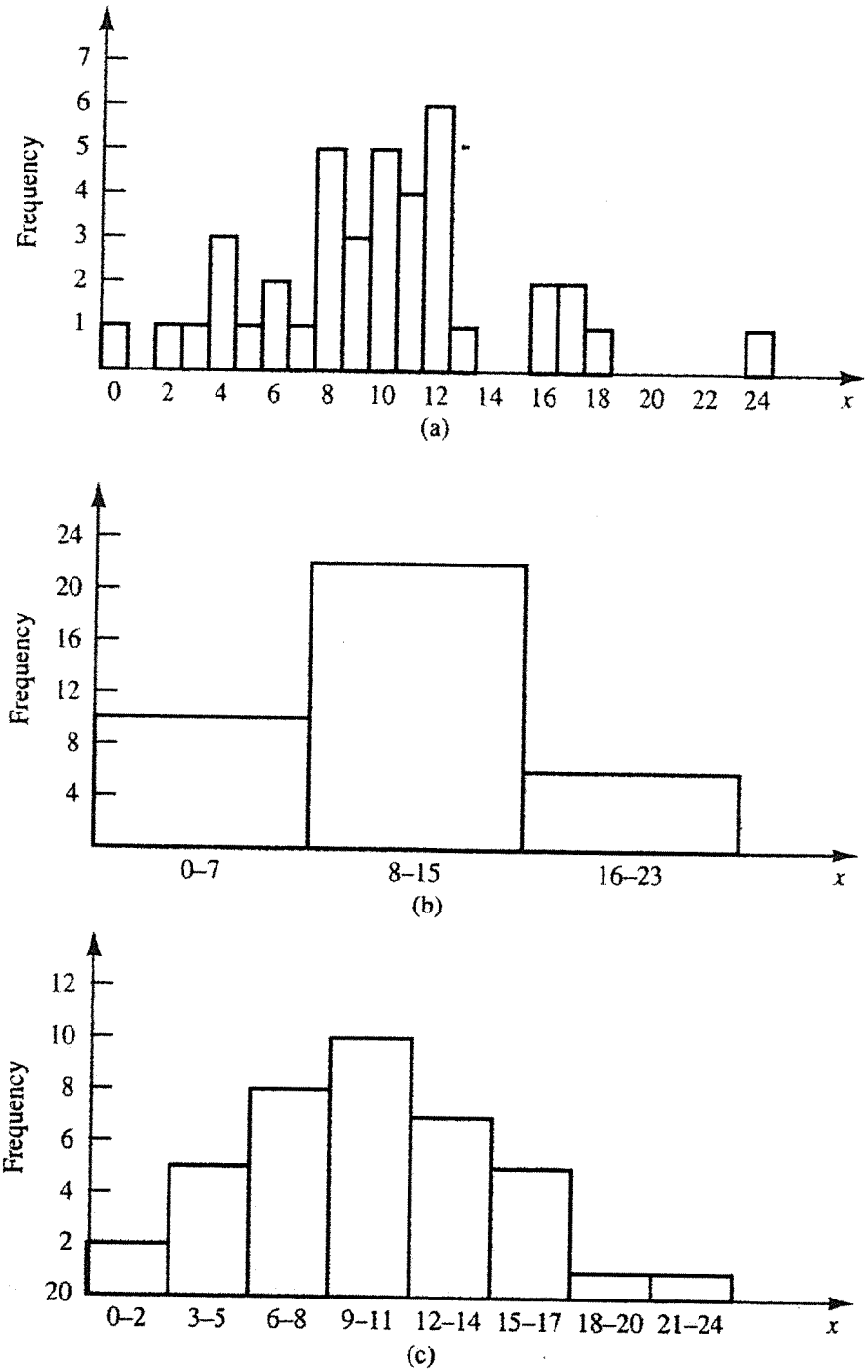

Figure 9.1. Ragged, coarse, and appropriate histograms: (a) original data — too ragged; (b) combining adjacent cells — too coarse; (c) combining adjacent cells — appropriate.

EXAMPLE 9.2 (Discrete Data)

The number of vehicles arriving at the northwest corner of an intersection in a 5-minute period between 7:00 A.M. and 7:05 A.M. was monitored for five workdays over a 20-week period. Table 9.1 shows the resulting data. The first entry in the table indicates that there were 12 5-minute periods during which zero vehicles arrived, 10 periods during which one vehicle arrived, and so on.

Since the number of automobiles is a discrete variable, and since there are ample data, the histogram can have a cell for each possible value in the range of the data. The resulting histogram is shown in Figure 9.2. ◀

Table 9.1. Number of Arrivals in a 5-Minute Period

Arrivals per Period	*Frequency*	*Arrivals per Period*	*Frequency*
0	12	6	7
1	10	7	5
2	19	8	5
3	17	9	3
4	10	10	3
5	8	11	1

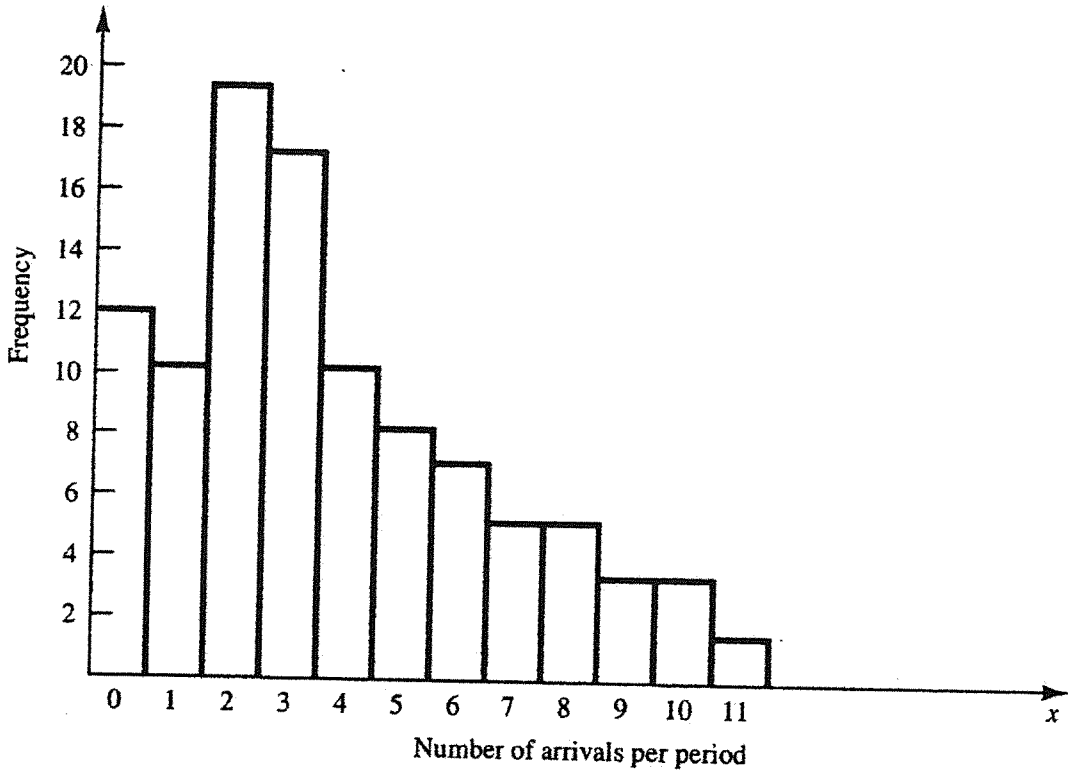

Figure 9.2. Histogram of number of arrivals per period.

EXAMPLE 9.3 (Continuous Data)

Life tests were performed on a random sample of electronic chips at 1.5 times the nominal voltage, and their lifetime (or time to failure) in days was recorded:

79.919	3.081	0.062	1.961	5.845
3.027	6.505	0.021	0.013	0.123
6.769	59.899	1.192	34.760	5.009
18.387	0.141	43.565	24.420	0.433
144.695	2.663	17.967	0.091	9.003
0.941	0.878	3.371	2.157	7.579
0.624	5.380	3.148	7.078	23.960
0.590	1.928	0.300	0.002	0.543
7.004	31.764	1.005	1.147	0.219
3.217	14.382	1.008	2.336	4.562

Lifetime, usually considered a continuous variable, is recorded here to three-decimal-place accuracy. The histogram is prepared by placing the data in class intervals. The range of the data is rather large, from 0.002 day to 144.695 days. However, most of the values (30 of 50) are in the zero-to–5-day range. Using intervals of width three results in Table 9.2. The data of Table 9.2 are then used to prepare the histogram shown in Figure 9.3. ◀

Table 9.2. Electronic Chip Data

Chip Life (Days)	*Frequency*
$0 \le x_j < 3$	23
$3 \le x_j < 6$	10
$6 \le x_j < 9$	5
$9 \le x_j < 12$	1
$12 \le x_j < 15$	1
$15 \le x_j < 18$	2
$18 \le x_j < 21$	0
$21 \le x_j < 24$	1
$24 \le x_j < 27$	1
$27 \le x_j < 30$	0
$30 \le x_j < 33$	1
$33 \le x_j < 36$	1
.	.
.	.
.	.
$42 \le x_j < 45$	1
.	.
.	.
.	.
$57 \le x_j < 60$	1
.	.
.	.
.	.
$78 \le x_j < 81$	1
.	.
.	.
.	.
$144 \le x_j < 147$	1

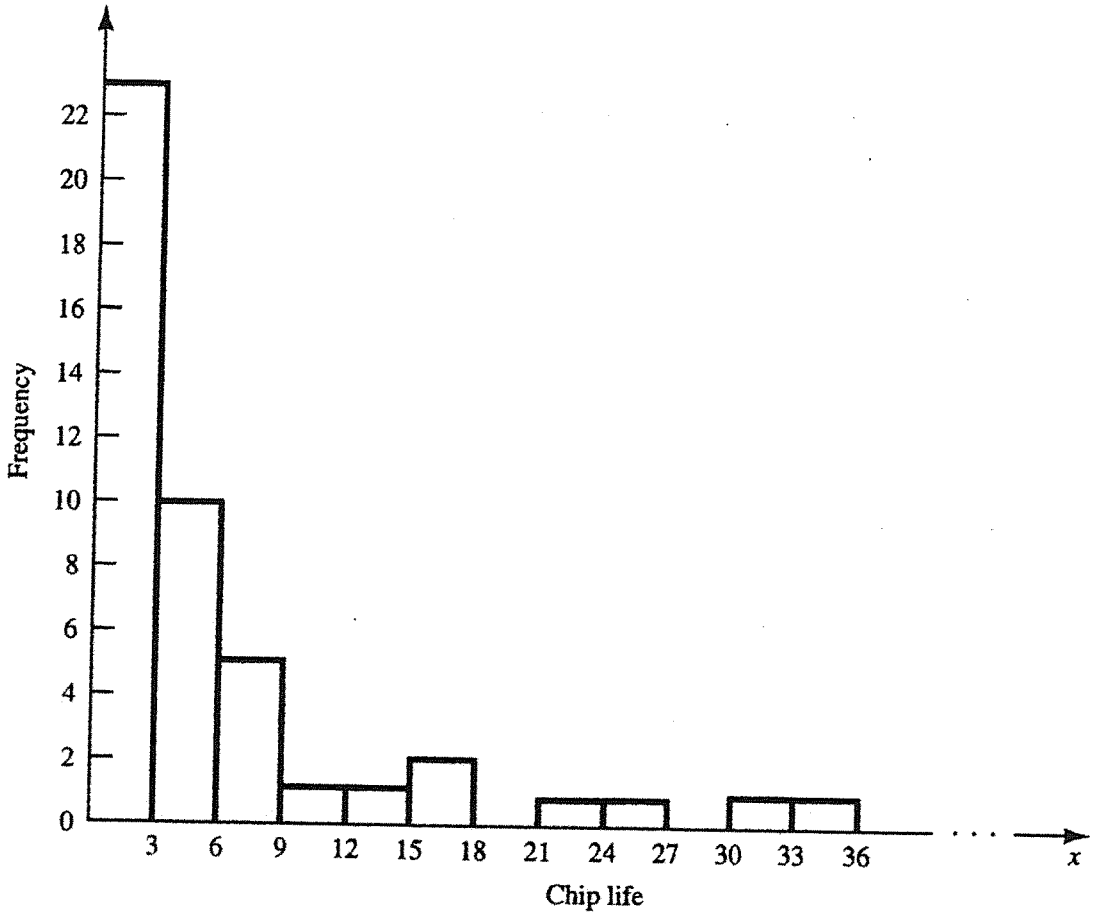

Figure 9.3. Histogram of chip life.

9.2.2 Selecting the Family of Distributions

In Chapter 5 some distributions that often arise in simulation were described. Additionally, the shapes of these distributions were displayed. The purpose of preparing a histogram is to infer a known pdf or pmf. A family of distributions is selected on the basis of what might arise in the context being investigated along with the shape of the histogram. Thus, if interarrival-time data have been collected, and the histogram has a shape similar to the pdf in Figure 5.9, the assumption of an exponential distribution would be warranted. Similarly, if measurements of the weights of pallets of freight are being made, and the histogram appears symmetric about the mean with a shape like that shown in Figure 5.12, the assumption of a normal distribution would be warranted.

The exponential, normal, and Poisson distributions are frequently encountered and are not difficult to analyze from a computational standpoint. Although more difficult to analyze, the gamma and Weibull distributions provide a wide array of shapes and should not be overlooked when modeling an underlying probabilistic process. Perhaps an exponential distribution was assumed, but it was found not to fit the data. The next step would be to examine where the lack of fit occurred. If the lack of fit was in one of the tails of the distribution, perhaps a gamma or Weibull distribution would more adequately fit the data.

Literally hundreds of probability distributions have been created, many with some specific physical process in mind. One aid to selecting distributions is to use the physical basis of the distributions as a guide. Here are some examples:

Binomial Models the number of successes in n trials, when the trials are independent with common success probability, p; for example, the number of defective computer chips found in a lot of n chips.

Negative Binomial (includes the geometric distribution) Models the number of trials required to achieve k successes; for example, the number of computer chips that we must inspect to find 4 defective chips.

Poisson Models the number of independent events that occur in a fixed amount of time or space; for example, the number of customers that arrive to a store during 1 hour, or the number of defects found in 30 square meters of sheet metal.

Normal Models the distribution of a process that can be thought of as the sum of a number of component processes; for example, the time to assemble a product which is the sum of the times required for each assembly operation. Notice that the normal distribution admits negative values, which may be impossible for process times.

Lognormal Models the distribution of a process that can be thought of as the product of (meaning to multiply together) a number of component processes; for example, the rate of return on an investment, when interest is compounded, is the product of the returns for a number of periods.

Exponential Models the time between independent events, or a process time which is memoryless (knowing how much time has passed gives no information about how much additional time will pass before the process is complete); for example, the times between the arrivals of a large number of customers who act independently of each other. The exponential is a highly variable distribution and is sometimes overused because it often leads to mathematically tractable models. Recall that, if the time between events is exponentially distributed, then the number of events in a fixed period of time is Poisson.

Gamma An extremely flexible distribution used to model nonnegative random variables. The gamma can be shifted away from 0 by adding a constant.

Beta An extremely flexible distribution used to model bounded (fixed upper and lower limits) random variables. The beta can be shifted away from 0 by adding a constant and can have a larger range than [0, 1] by multiplying by a constant.

Erlang Models processes that can be viewed as the sum of several exponentially distributed processes; for example, a computer network fails when a computer and two backup computers fail, and each has a time to failure that is exponentially distributed. The Erlang is a special case of the gamma.

Weibull Models the time to failure for components; for example, the time to failure for a disk drive. The exponential is a special case of the Weibull.

Discrete or Continuous Uniform Models complete uncertainty, since all outcomes are equally likely. This distribution is often overused when there are no data.

Triangular Models a process when only the minimum, most-likely, and maximum values of the distribution are known; for example, the minimum, most-likely, and maximum time required to test a product.

Empirical Resamples from the actual data collected; often used when no theoretical distribution seems appropriate.

Do not ignore physical characteristics of the process when selecting distributions. Is the process naturally discrete or continuous valued? Is it bounded or is there no natural bound? This knowledge, which does not depend on data, can help narrow the family of distributions from which to choose. And keep in mind that there is no "true" distribution for any stochastic input process. An input model is an approximation of reality, so the goal is to obtain an approximation that yields useful results from the simulation experiment.

The reader is encouraged to complete Exercises 6 through 11 to learn more about the shapes of the distributions mentioned in this section. Examining the variations in shape, as the parameters change will be very instructive.

9.2.3 *Quantile-Quantile Plots*

The construction of histograms, as discussed in Section 9.2.1, and the recognition of a distributional shape, as discussed in Section 9.2.2, are necessary ingredients for selecting a family of distributions to represent a sample of data. However, a histogram is not as useful for evaluating the *fit* of the chosen distribution. When there are a small number of data points, say 30 or fewer, a histogram can be rather ragged. Further, our perception of the fit depends on the widths of the histogram intervals. But even if the intervals are well chosen, grouping data into cells makes it difficult to compare a histogram to a continuous probabiliy density function. A quantile-quantile (q-q) plot is a useful tool for evaluating distribution fit that does not suffer from these problems.

If X is a random variable with cdf F, then the q-quantile of X is that value γ such that $F(\gamma) = P(X \leq \gamma) = q$, for $0 < q < 1$. When F has an inverse, we write $\gamma = F^{-1}(q)$.

Now let $\{x_i, i = 1, 2, \ldots, n\}$ be a sample of data from X. Order the observations from the smallest to the largest, and denote these as $\{y_j, j = 1, 2, \ldots, n\}$, where $y_1 \leq y_2 \leq \cdots \leq y_n$. Let j denote the ranking or order number. Therefore, $j = 1$ for the smallest and $j = n$ for the largest. The q-q plot is based on the fact that y_j is an estimate of the $(j - 1/2)/n$ quantile of X. In other words,

$$y_j \text{ is approximately } F^{-1}\left(\frac{j - \frac{1}{2}}{n}\right)$$

Now suppose that we have chosen a distribution with cdf F as a possible representation of the distribution of X. If F is a member of an appropriate family of distributions, then a plot of y_j versus $F^{-1}((j-1/2)/n)$ will be *approximately a straight line*. If F is from an appropriate family of distributions and also has appropriate parameter values, then the line will have slope 1. On the other hand, if the assumed distribution is inappropriate, the points will deviate from a straight line, usually in a systematic manner. The decision of whether or not to reject some hypothesized model is subjective.

EXAMPLE 9.4 (Normal Q-Q Plot)

A robot is used to install the doors on automobiles along an assembly line. It was thought that the installation times followed a normal distribution. The robot is capable of accurately measuring installation times. A sample of 20 installation times was automatically taken by the robot with the following results, where the values are in seconds:

99.79	99.56	100.17	100.33
100.26	100.41	99.98	99.83
100.23	100.27	100.02	100.47
99.55	99.62	99.65	99.82
99.96	99.90	100.06	99.85

The sample mean is 99.99 seconds, and the sample variance is $(0.2832)^2$ seconds^2. These values can serve as the parameter estimates for the mean and variance of the normal distribution. The observations are now ordered from smallest to largest as follows:

j	*Value*	j	*Value*	j	*Value*	j	*Value*
1	99.55	6	99.82	11	99.98	16	100.26
2	99.56	7	99.83	12	100.02	17	100.27
3	99.62	8	99.85	13	100.06	18	100.33
4	99.65	9	99.90	14	100.17	19	100.41
5	99.79	10	99.96	15	100.23	20	100.47

The ordered observations are then plotted versus $F^{-1}((j-1/2)/20)$, for $j = 1, 2, \ldots, 20$, where F is the cdf of the normal distribution with mean 99.99 and variance $(0.2832)^2$ to obtain a q-q plot. The plotted values are shown in Figure 9.4, along with a histogram of the data that has the density function of the normal distribution superimposed. Notice that it is difficult to tell if the data are well represented by a normal distribution by looking at the histogram, but the general perception of a straight line is quite clear in the q-q plot, supporting the hypothesis of a normal distribution. ◀

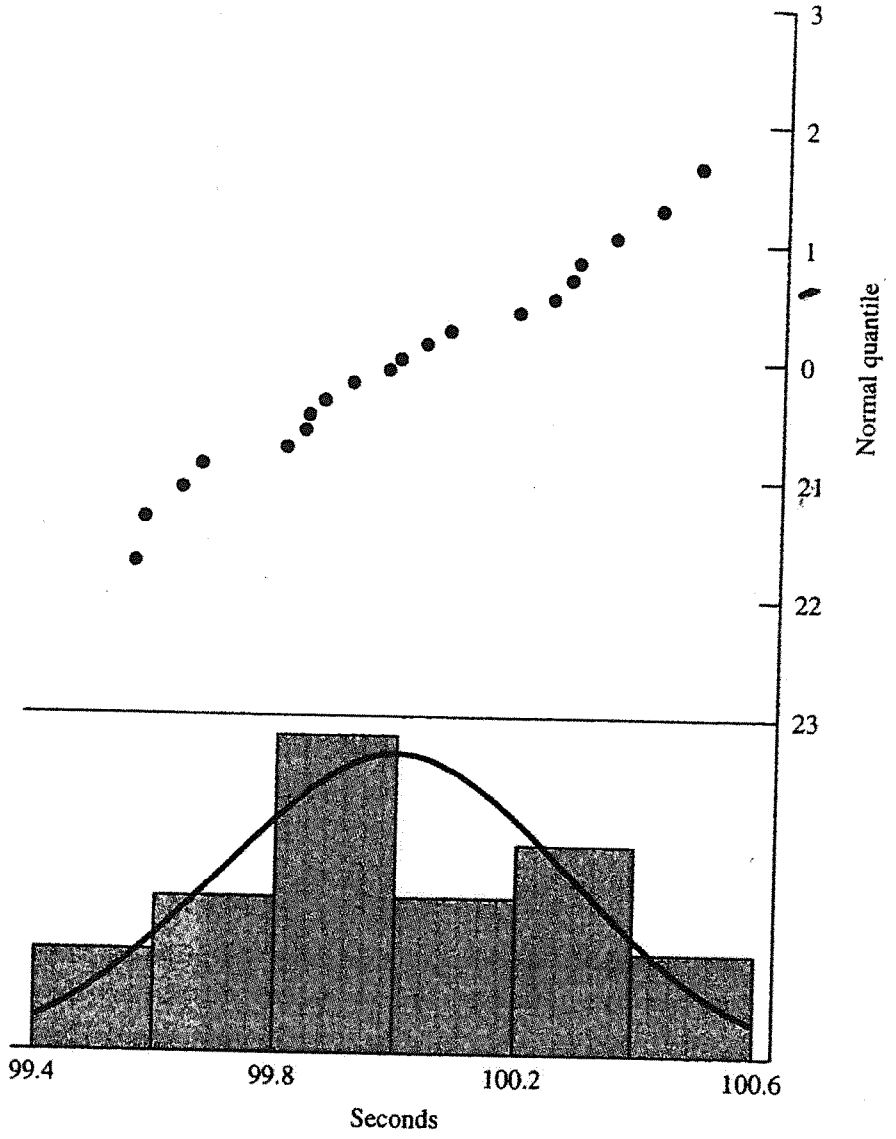

Figure 9.4. Histogram and q-q plot of the installation times.

In the evaluation of the linearity of a q-q plot, the following should be considered:

1. The observed values will never fall exactly on a straight line.
2. The ordered values are not independent, since they have been ranked. Hence, if one point is above a straight line, it is likely that the next point will also lie above the line. And it is unlikely that the points will be scattered about the line.
3. The variances of the extremes (largest and smallest values) are much higher than the variances in the middle of the plot. Greater discrepancies can be accepted at the extremes. The linearity of the points in the middle of the plot is more important than the linearity at the extremes.

Modern data-analysis software often includes tools for generating q-q plots, especially for the normal distribution. The q-q plot can also be used to compare two samples of data to see if they can be represented by the same distribution

(that is, they are homogeneous). If $x_1, x_2, \ldots, x_n$ are a sample of the random variable X, and $z_1, z_2, \ldots, z_n$ are a sample of the random variable Z, then plotting the ordered values of X versus the ordered values of Z will reveal approximately a straight line if both samples are well represented by the same distribution (Chambers, Cleveland, Kleiner, and Tukey [1983]).

9.3 Parameter Estimation

After a family of distributions has been selected, the next step is to estimate the parameters of the distribution. Estimators for many useful distributions are described in this section. In addition, many software packages—some of them integrated into simulation languages—are now available to compute these estimates.

9.3.1 Preliminary Statistics: Sample Mean and Sample Variance

In a number of instances the sample mean, or the sample mean and sample variance, are used to estimate the parameters of a hypothesized distribution; see Example 9.4. In the following paragraphs, three sets of equations are given for computing the sample mean and sample variance. Equations (9.1) and (9.2) can be used when discrete or continuous raw data are available. Equations (9.3) and (9.4) are used when the data are discrete and have been grouped in a frequency distribution. Equations (9.5) and (9.6) are used when the data are discrete or continuous and have been placed in class intervals. Equations (9.5) and (9.6) are approximations and should be used only when the raw data are unavailable.

If the observations in a sample of size n are $X_1, X_2, \ldots, X_n$, the sample mean ($\bar{X}$) is defined by

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{9.1}$$

and the sample variance, S^2, is defined by

$$S^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{n - 1} \tag{9.2}$$

If the data are discrete and grouped in a frequency distribution, Equations (9.1) and (9.2) can be modified to provide for much greater computational efficiency. The sample mean can be computed by

$$\bar{X} = \frac{\sum_{j=1}^{k} f_j X_j}{n} \tag{9.3}$$

and the sample variance by

$$S^2 = \frac{\sum_{j=1}^{k} f_j X_j^2 - n\bar{X}^2}{n - 1} \tag{9.4}$$

where k is the number of distinct values of X and f_j is the observed frequency of the value X_j of X.

EXAMPLE 9.5 (Grouped Data)

The data in Table 9.1 can be analyzed to obtain $n = 100$, $f_1 = 12$, $X_1 = 0$, $f_2 = 10$, $X_2 = 1, \ldots$, $\sum_{j=1}^{k} f_j X_j = 364$, and $\sum_{j=1}^{k} = f_j X_j^2 = 2080$. From Equation (9.3),

$$\bar{X} = \frac{364}{100} = 3.64$$

and from Equation (9.4),

$$S^2 = \frac{2080 - 100(3.64)^2}{99} = 7.63$$

The sample standard deviation, S, is just the square root of the sample variance. In this case $S = \sqrt{7.63} = 2.76$. Equations (9.1) and (9.2) would have yielded exactly the same results for $\bar{X}$ and S^2. ◀

It is preferable to use the raw data, if possible, when the values are continuous. However, the data may have been received after they have been placed in class intervals. Then it is no longer possible to obtain the exact sample mean and variance. In such cases, the sample mean and sample variance are approximated from the following equations:

$$\bar{X} \doteq \frac{\sum_{j=1}^{c} f_j m_j}{n} \tag{9.5}$$

and

$$S^2 \doteq \frac{\sum_{j=1}^{c} f_j m_j^2 - n\bar{X}^2}{n - 1} \tag{9.6}$$

where f_j is the observed frequency in the jth class interval, m_j is the midpoint of the jth interval, and c is the number of class intervals.

EXAMPLE 9.6 (Continuous Data in Class Intervals)

Assume that the raw data on chip life shown in Example 9.3 were either discarded or lost. However, the data shown in Table 9.2 are still available. To determine approximate values of $\bar{X}$ and S^2 Equations (9.5) and (9.6) are used. The following values are determined: $f_1 = 23$, $m_1 = 1.5$, $f_2 = 10$, $m_2 = 4.5, \ldots$, $\sum_{j=1}^{49} f_j m_j = 614$, and $\sum_{j=1}^{49} f_j m_j^2 = 37{,}226.5$. With $n = 50$, $\bar{X}$ is approximated from Equation (9.5) as

$$\bar{X} \doteq \frac{614}{50} = 12.28$$

Then, S^2 is approximated from Equation (9.6) as

$$S^2 \doteq \frac{37{,}226.5 - 50(12.28)^2}{49} = 605.849$$

and

$$S \doteq 24.614$$

Applying Equations (9.1) and (9.2) to the original data in Example 9.3 results in $\bar{X} = 11.894$ and $S = 24.953$. Thus, when the raw data are either discarded or lost, some inaccuracies may result. ◀

9.3.2 *Suggested Estimators*

Numerical estimates of the distribution parameters are needed to reduce the family of distributions to a specific distribution and to test the resulting hypothesis. Table 9.3 contains suggested estimators for distributions often used in simulation, all of which were described in Chapter 5. Except for an adjustment to remove bias in the estimate of σ^2 for the normal distribution, these estimators are the maximum-likelihood estimators based on the raw data. (If the data are in class intervals, these estimators must be modified.) The reader is referred to Fishman [1973] and Law and Kelton [2000] for parameter estimates for the beta, uniform, binomial, and negative binomial distributions. The triangular distribution is usually employed when no data are available, with the parameters obtained from educated guesses for the minimum, most likely, and maximum possible values; the uniform distribution may also be used in this way if only minimum and maximum values are available.

Examples of the use of the estimators are given in the following paragraphs. The reader should keep in mind that a parameter is an unknown constant, but the estimator is a statistic or random variable because it depends on the sample values. To distinguish the two clearly, if, say, a parameter is denoted by α, the estimator will be denoted by $\widehat{\alpha}$.

EXAMPLE 9.7 (Poisson Distribution)

Assume that the arrival data in Table 9.1 require analysis. By comparison to Figure 5.7, an examination of Figure 9.2 suggests a Poisson distributional assumption with unknown parameter α. From Table 9.3, the estimator of α is $\bar{X}$, which was determined in Example 9.5. Thus, $\widehat{\alpha} = 3.64$. Recall that the true mean and variance are equal for the Poisson distribution. In Example 9.5, the sample variance was estimated by $S^2 = 7.63$. However, it should never be expected that the sample mean and the sample variance will be precisely equal, since both are random variables. ◀

EXAMPLE 9.8 (Lognormal Distribution)

The percentage rates of return on 10 investments in a portfolio are 18.8, 27.9, 21.0, 6.1, 37.4, 5.0, 22.9, 1.0, 3.1, and 8.3. To estimate the parameters of a lognormal model of this data, we first take the natural log of the data to obtain 2.9, 3.3, 3.0, 1.8, 3.6, 1.6, 3.1, 0, 1.1, and 2.1. Then set $\widehat{\mu} = \bar{X} = 2.3$ and $\widehat{\sigma}^2 = S^2 = 1.3$. ◀

EXAMPLE 9.9 (Normal Distribution)

The parameters of the normal distribution, μ and σ^2, are estimated by $\bar{X}$ and S^2, as shown in Table 9.3. The q-q plot in Example 9.4 leads to a distributional assumption that the installation times are normal. Using Equations (9.1) and (9.2), the data in Example 9.4 yield $\widehat{\mu} = \bar{X} = 99.9865$ and $\widehat{\sigma}^2 = S^2 = (0.2832)^2$ second^2. ◀

EXAMPLE 9.10 (Gamma Distribution)

The estimator, $\widehat{\beta}$, for the gamma distribution is determined by the use of Table A.9 from Choi and Wette [1969]. Table A.9 requires the computation of the quantity $1/M$, where

$$M = \ell n \bar{X} - \frac{1}{n} \sum_{i=1}^{n} \ell n X_i \tag{9.7}$$

Also, it can be seen in Table 9.3 that $\widehat{\theta}$ is given by

$$\widehat{\theta} = \frac{1}{\bar{X}} \tag{9.8}$$

In Chapter 5 it was stated that lead time is often gamma distributed. Suppose that the lead times (in days) associated with 20 orders have been

Table 9.3. Suggested Estimators for Distributions Often Used in Simulation

Distribution	*Parameter(s)*	*Suggested Estimator(s)*
Poisson	α	$\hat{\alpha} = \bar{X}$
Exponential	λ	$\hat{\lambda} = \frac{1}{\bar{X}}$
Gamma	β, θ	$\hat{\beta}$ (see Table A.9)
		$\hat{\theta} = \frac{1}{\bar{X}}$
Normal	μ, σ^2	$\hat{\mu} = \bar{X}$
		$\hat{\sigma}^2 = S^2$ (unbiased)
Lognormal	μ, σ^2	$\hat{\mu} = \bar{X}$ (after taking ℓn of the data)
		$\hat{\sigma}^2 = S^2$ (after taking ℓn of the data)
Weibull with $\nu = 0$	α, β	$\hat{\beta}_0 = \frac{\bar{X}}{S}$
		$\hat{\beta}_j = \hat{\beta}_{j-1} - \frac{f(\hat{\beta}_{j-1})}{f'(\hat{\beta}_{j-1})}$
		See Equations (9.12) and (9.15) for $f(\hat{\beta})$ and $f'(\hat{\beta})$.
		Iterate until convergence:
		$\hat{\alpha} = \left(\frac{1}{n} \sum_{i=1}^{n} X_i^{\widehat{\beta}}\right)^{1/\widehat{\beta}}$

accurately measured as follows:

Order	Lead Time (Days)	Order	Lead Time (Days)
1	70.292	11	30.215
2	10.107	12	17.137
3	48.386	13	44.024
4	20.480	14	10.552
5	13.053	15	37.298
6	25.292	16	16.314
7	14.713	17	28.073
8	39.166	18	39.019
9	17.421	19	32.330
10	13.905	20	36.547

To determine $\widehat{\beta}$ and $\widehat{\theta}$, it is first necessary to determine M using Equation (9.7). Here, $\bar{X}$ is determined from Equation (9.1) to be

$$\bar{X} = \frac{564.32}{20} = 28.22$$

Then,

$$\ell n \bar{X} = 3.34$$

Next,

$$\sum_{i=1}^{20} \ell n X_i = 63.99$$

Then,

$$M = 3.34 - \frac{63.99}{20} = 0.14$$

and

$$1/M = 7.14$$

By interpolation in Table A.9, $\widehat{\beta} = 3.728$. Finally, Equation (9.8) results in

$$\widehat{\theta} = \frac{1}{28.22} = 0.035$$

◀

EXAMPLE 9.11 (Exponential Distribution)

Assuming that the data in Example 9.3 come from an exponential distribution, the parameter estimate, $\widehat{\lambda}$, can be determined. In Table 9.3, $\widehat{\lambda}$ is obtained using $\bar{X}$ as follows:

$$\widehat{\lambda} = \frac{1}{\bar{X}} = \frac{1}{11.894} = 0.084 \text{ per day}$$

◀

EXAMPLE 9.12 (Weibull Distribution)

Suppose that a random sample of size n, $X_1, X_2, \ldots, X_n$, has been taken, and the observations are assumed to come from a Weibull distribution. The likelihood function derived using the pdf given by Equation (5.46) can be shown to be

$$L(\alpha, \beta) = \frac{\beta^n}{\alpha^{\beta n}} \left[\prod_{i=1}^{n} X_i^{(\beta-1)} \right] \exp\left[-\sum_{i=1}^{n} \left(\frac{X_i}{\alpha} \right)^{\beta} \right] \tag{9.9}$$

The maximum-likelihood estimates are those values of $\hat{\alpha}$ and $\hat{\beta}$ that maximize $L(\alpha, \beta)$, or equivalently maximize $\ell n L(\alpha, \beta)$, denoted by $l(\alpha, \beta)$. The maximum value of $l(\alpha, \beta)$ is obtained by taking the partial derivatives $\partial l(\alpha, \beta)/\partial \alpha$ and $\partial l(\alpha, \beta)/\partial \beta$, setting each to zero, and solving the resulting equations, which after substitution become

$$f(\beta) = 0 \tag{9.10}$$

and

$$\alpha = \left(\frac{1}{n} \sum_{i=1}^{n} X_i^{\beta} \right)^{1/\beta} \tag{9.11}$$

where

$$f(\beta) = \frac{n}{\beta} + \sum_{i=1}^{n} \ell n X_i - \frac{n \sum_{i=1}^{n} X_i^{\beta} \ell n X_i}{\sum_{i=1}^{n} X_i^{\beta}} \tag{9.12}$$

The maximum-likelihood estimates, $\hat{\alpha}$ and $\hat{\beta}$, are the solutions of Equations (9.10) and (9.11). First $\hat{\beta}$ is determined through the iterative procedure explained below. Then $\hat{\alpha}$ is determined using Equation (9.11) with $\beta = \hat{\beta}$.

Since Equation (9.10) is nonlinear, it is necessary to use a numerical analysis technique to solve it. In Table 9.3 an iterative method for determining $\hat{\beta}$ is given as

$$\hat{\beta}_j = \hat{\beta}_{j-1} - \frac{f(\hat{\beta}_{j-1})}{f'(\hat{\beta}_{j-1})} \tag{9.13}$$

Equation (9.13) employs Newton's method in reaching $\hat{\beta}$, where $\hat{\beta}_j$ is the jth iteration beginning with an initial estimate for $\hat{\beta}_0$, given in Table 9.3, as follows:

$$\hat{\beta}_0 = \frac{\bar{X}}{S} \tag{9.14}$$

If the initial estimate, $\hat{\beta}_0$, is sufficiently close to the solution $\hat{\beta}$, then $\hat{\beta}_j$ approaches $\hat{\beta}$ as $j \longrightarrow \infty$. When using Newton's method, $\hat{\beta}$ is approached through increments of size $f(\hat{\beta}_{j-1})/f'(\hat{\beta}_{j-1})$. Equation (9.12) is used to compute $f(\hat{\beta}_{j-1})$ and Equation (9.15) is used to compute $f'(\hat{\beta}_{j-1})$ as follows:

$$f'(\beta) = -\frac{n}{\beta^2} - \frac{n \sum_{i=1}^{n} X_i^{\beta} (\ell n X_i)^2}{\sum_{i=1}^{n} X_i^{\beta}} + \frac{n \left(\sum_{i=1}^{n} X_i^{\beta} \ell n X_i \right)^2}{\left(\sum_{i=1}^{n} X_i^{\beta} \right)^2} \tag{9.15}$$

Equation (9.15) can be derived from Equation (9.12) by differentiating $f(\beta)$ with respect to β. The iterative process continues until $f(\hat{\beta}_j) \doteq 0$, for example, until $|f(\hat{\beta}_j)| \leq 0.001$.

Consider the data given in Example 9.3. These data concern the failure of electronic components and may come from an exponential distribution. In Example 9.11, the parameter $\hat{\lambda}$ was estimated on the hypothesis that the data were from an exponential distribution. If the hypothesis that the data came from an exponential distribution is rejected, an alternative hypothesis is that the data come from a Weibull distribution. The Weibull distribution is suspected, since the data pertain to electronic component failures which occur suddenly.

Equation (9.14) is used to determine $\hat{\beta}_0$. For the data in Example 9.3, $n = 50$, $\bar{X} = 11.894$, $\bar{X}^2 = 141.467$, and $\sum_{i=1}^{50} X_1^2 = 37{,}575.850$, so that S^2 is found by Equation (9.2) to be

$$S^2 = \frac{37{,}578.850 - 50(141.467)}{49} = 622.650$$

and $S = 24.953$. Thus,

$$\hat{\beta}_0 = \frac{11.894}{24.953} = 0.477$$

To compute $\hat{\beta}_1$ using Equation (9.13) requires the determination of $f(\hat{\beta}_0)$ and $f'(\hat{\beta}_0)$ using Equations (9.12) and (9.15). The following additional values are needed: $\sum_{i=1}^{50} X_i^{\hat{\beta}_0} = 115.125$, $\sum_{i=1}^{50} \ell n X_i = 38.294$, $\sum_{i=1}^{50} X_i^{\hat{\beta}_0} \ell n X_i = 292.629$, and $\sum_{i=1}^{50} X_i^{\hat{\beta}_0} (\ell n X_i)^2 = 1057.781$. Thus,

$$f(\hat{\beta}_0) = \frac{50}{0.477} + 38.294 - \frac{50(292.629)}{115.125} = 16.024$$

and

$$f'(\hat{\beta}_0) = \frac{-50}{(0.477)^2} - \frac{50(1057.781)}{115.125} + \frac{50(292.629)^2}{(115.125)^2} = -356.110$$

Then, by Equation (9.13),

$$\hat{\beta}_1 = 0.477 - \frac{16.024}{-356.110} = 0.522$$

After four iterations, $|f(\hat{\beta}_3)| \leq 0.001$, at which point $\hat{\beta} \doteq \hat{\beta}_4 = 0.525$ is the approximate solution to Equation (9.10). Table 9.4 contains the values needed to complete each iteration.

Now, $\hat{\alpha}$ can be determined using Equation (9.11) with $\beta = \hat{\beta} = 0.525$ as follows:

$$\hat{\alpha} = \left[\frac{130.608}{50}\right]^{1/0.525} = 6.227$$

If $\hat{\beta}_0$ is sufficiently close to $\hat{\beta}$, the procedure converges quickly, usually in four to five iterations. However, if the procedure appears to be diverging,

Table 9.4. Iterative Estimation of Parameters of the Weibull Distribution

j	$\hat{\beta}_j$	$\sum_{i=1}^{50} X_i^{\hat{\beta}_j}$	$\sum_{i=1}^{50} X_i^{\hat{\beta}_j} \ln X_i$	$\sum_{i=1}^{50} X_i^{\hat{\beta}_j} (\ln X_i)^2$	$f(\hat{\beta}_j)$	$f'(\hat{\beta}_j)$	$\hat{\beta}_{j+1}$
0	0.477	115.125	292.629	1057.781	16.024	−356.110	0.522
1	0.522	129.489	344.713	1254.111	1.008	−313.540	0.525
2	0.525	130.603	348.769	1269.547	0.004	−310.853	0.525
3	0.525	130.608	348.786	1269.614	0.000	−310.841	0.525

try other initial guesses for $\widehat{\beta}_0$ — for example, one-half the initial estimate or twice the initial estimate.

The difficult task of determining parameters for the Weibull distribution by hand emphasizes the value of having software support for input modeling. ◀

9.4 Goodness-of-Fit Tests

Hypothesis testing was discussed in Section 7.4 with respect to testing random numbers. In Section 7.4.1 the Kolmogorov-Smirnov test and the chi-square test were introduced. These two tests are applied in this section to hypotheses about distributional forms of input data.

Goodness-of-fit tests provide helpful guidance for evaluating the suitability of a potential input model. However, since there is no single correct distribution in a real application, you should not be a slave to the verdict of such tests. It is especially important to understand the effect of sample size. If very little data are available, then a goodness-of-fit test is unlikely to reject *any* candidate distribution; but if a lot of data are available, then a goodness-of-fit test will likely reject *all* candidate distributions. Therefore, failing to reject a candidate distribution should be taken as one piece of evidence in favor of that choice, while rejecting an input model is only one piece of evidence against the choice.

9.4.1 Chi-Square Test

One procedure for testing the hypothesis that a random sample of size n of the random variable X follows a specific distributional form is the chi-square goodness-of-fit test. This test formalizes the intuitive idea of comparing the histogram of the data to the shape of the candidate density or mass function. The test is valid for large sample sizes, for both discrete and continuous distributional assumptions, when parameters are estimated by maximum likelihood. The test procedure begins by arranging the n observations into a set of k class intervals or cells. The test statistic is given by

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{9.16}$$

where O_i is the observed frequency in the ith class interval and E_i is the expected frequency in that class interval. The expected frequency for each class interval is computed as $E_i = np_i$, where p_i is the theoretical, hypothesized probability associated with the ith class interval.

It can be shown that χ_0^2 approximately follows the chi-square distribution with $k - s - 1$ degrees of freedom, where s represents the number of parameters of the hypothesized distribution estimated by the sample statistics. The hypotheses are:

H_0: the random variable, X, conforms to the distributional assumption with the parameter(s) given by the parameter estimate(s)

H_1: the random variable X does not conform

The critical value $\chi^2_{\alpha,k-s-1}$ is found in Table A.6. The null hypothesis, H_0, is rejected if $\chi_0^2 > \chi^2_{\alpha,k-s-1}$.

In applying the test, if expected frequencies are too small, χ_0^2 will reflect not only the departure of the observed from the expected frequency but the smallness of the expected frequency as well. Although there is no general agreement regarding the minimum size of E_i, values of 3, 4, and 5 have been widely used. In Section 7.4.1, when the chi-square test was discussed, the minimum expected frequency of five was suggested. If an E_i value is too small, it can be combined with expected frequencies in adjacent class intervals. The corresponding O_i values should also be combined, and k should be reduced by one for each cell that is combined.

If the distribution being tested is discrete, each value of the random variable should be a class interval, unless it is necessary to combine adjacent class intervals to meet the minimum expected cell-frequency requirement. For the discrete case, if combining adjacent cells is not required,

$$p_i = p(x_i) = P(X = x_i)$$

Otherwise, p_i is determined by summing the probabilities of appropriate adjacent cells.

If the distribution being tested is continuous, the class intervals are given by $[a_{i-1}, a_i)$, where a_{i-1} and a_i are the endpoints of the ith class interval. For the continuous case with assumed pdf $f(x)$, or assumed cdf $F(x)$, p_i can be computed by:

$$p_i = \int_{a_{i-1}}^{a_i} f(x)\,dx = F(a_i) - F(a_{i-1})$$

For the discrete case, the number of class intervals is determined by the number of cells resulting after combining adjacent cells as necessary. However, for the continuous case the number of class intervals must be specified. Although there are no general rules to be followed, the recommendations in Table 9.5 are made to aid in determining the number of class intervals for continuous data.

Table 9.5. Recommendations for Number of Class Intervals for Continuous Data

Sample Size, n	*Number of Class Intervals, k*
20	Do not use the chi-square test
50	5 to 10
100	10 to 20
>100	$\sqrt{n}$ to $n/5$

EXAMPLE 9.13 (Chi-Square Test Applied to Poisson Assumption)

In Example 9.7, the vehicle-arrival data presented in Example 9.2 were analyzed. Since the histogram of the data, shown in Figure 9.2, appeared to follow a Poisson distribution, the parameter, $\widehat{\alpha} = 3.64$, was determined. Thus, the following hypotheses are formed:

H_0: the random variable is Poisson distributed

H_1: the random variable is not Poisson distributed

The pmf for the Poisson distribution was given in Equation (5.18) as follows:

$$p(x) = \begin{cases} \dfrac{e^{-\alpha}\alpha^x}{x!}, & x = 0, 1, 2, \ldots \\ 0, & \text{otherwise} \end{cases} \tag{9.17}$$

For $\alpha = 3.64$, the probabilities associated with various values of x are obtained using Equation (9.17) with the following results:

$$\begin{aligned} p(0) &= 0.026, & p(6) &= 0.085 \\ p(1) &= 0.096, & p(7) &= 0.044 \\ p(2) &= 0.174, & p(8) &= 0.020 \\ p(3) &= 0.211, & p(9) &= 0.008 \\ p(4) &= 0.192, & p(10) &= 0.003 \\ p(5) &= 0.140, & p(11) &= 0.001 \end{aligned}$$

With this information, Table 9.6 is constructed. The value of E_1 is given by $np_0 = 100(0.026) = 2.6$. In a similar manner, the remaining E_i values are determined. Since $E_1 = 2.6 < 5$, E_1 and E_2 are combined. In that case O_1 and O_2 are also combined and k is reduced by one. The last five class intervals are also combined for the same reason, and k is further reduced by four.

The calculated χ_0^2 is 27.68. The degrees of freedom for the tabulated value of χ^2 is $k - s - 1 = 7 - 1 - 1 = 5$. Here, $s = 1$, since one parameter, $\widehat{\alpha}$, was estimated from the data. At the 0.05 level of significance, the critical value $\chi^2_{0.05,5}$ is 11.1. Thus, H_0 would be rejected at level of significance 0.05. The analyst may therefore want to search for a better-fitting model or use the empirical distribution of the data. ◀

Table 9.6. Chi-Square Goodness-of-Fit Test for Example 9.2

x_i	Observed Frequency, O_i		Expected Frequency, E_i		$\frac{(O_i - E_i)^2}{E_i}$
0	12	22	2.6	12.2	7.87
1	10		9.6		
2	19		17.4		0.15
3	17		21.1		0.80
4	10		19.2		4.41
5	8		14.0		2.57
6	7		8.5		0.26
7	5	17	4.4	7.6	11.62
8	5		2.0		
9	3		0.8		
10	3		0.3		
11	1		0.1		
	100		100.0		27.68

9.4.2 *Chi-Square Test with Equal Probabilities*

If a continuous distributional assumption is being tested, class intervals that are equal in probability rather than equal in width of interval should be used. This has been recommended by a number of authors (Mann and Wald [1942]; Gumbel [1943]; Law and Kelton [2000]; Stuart, Ord, and Arnold [1998]). It should be noted that the procedure is not applicable to data collected in class intervals, where the raw data have been discarded or lost.

Unfortunately, there is as yet no method for determining the probability associated with each interval that maximizes the power for a test of a given size. (The power of a test is defined as the probability of rejecting a false hypothesis.) However, if using equal probabilities, then $p_i = 1/k$. Since we recommend

$$E_i = np_i \geq 5$$

substituting for p_i yields

$$\frac{n}{k} \geq 5$$

and solving for k yields

$$k \leq \frac{n}{5} \tag{9.18}$$

Equation (9.18) was used in determining the recommended maximum number of class intervals in Table 9.5.

If the assumed distribution is normal, exponential, or Weibull, the method described in this section is straightforward. Example 9.14 indicates how the procedure is accomplished for the exponential distribution. If the assumed

distribution is gamma (but not Erlang), or certain other distributions, then the computation of endpoints for class intervals is complex and may require numerical integration of the density function. Statistical-analysis software is very helpful in such cases.

EXAMPLE 9.14 (Chi-Square Test for Exponential Distribution)

In Example 9.11, the failure data presented in Example 9.3 were analyzed. Since the histogram of the data, shown in Figure 9.3, appeared to follow an exponential distribution, the parameter $\hat{\lambda} = 1/\bar{X} = 0.084$ was determined. Thus, the following hypotheses are formed:

H_0: the random variable is exponentially distributed

H_1: the random variable is not exponentially distributed

In order to perform the chi-square test with intervals of equal probability, the endpoints of the class intervals must be determined. Equation (9.18) indicates that the number of intervals should be less than or equal to $n/5$. Here, $n = 50$, so that $k \leq 10$. In Table 9.5, it is recommended that 7 to 10 class intervals be used. Let $k = 8$, then each interval will have probability $p = 0.125$. The endpoints for each interval are computed from the cdf for the exponential distribution, given in Equation (5.27), as follows:

$$F(a_i) = 1 - e^{-\lambda a_i} \tag{9.19}$$

where a_i represents the endpoint of the ith interval, $i = 1, 2, \ldots, k$. Since $F(a_i)$ is the cumulative area from zero to a_i, $F(a_i) = ip$, so Equation (9.19) can be written as

$$ip = 1 - e^{-\lambda a_i}$$

or

$$e^{-\lambda a_i} = 1 - ip$$

Taking the logarithm of both sides and solving for a_i gives a general result for the endpoints of k equiprobable intervals for the exponential distribution, namely

$$a_i = -\frac{1}{\lambda}\ell n(1 - ip), \quad i = 0, 1, \ldots, k \tag{9.20}$$

Regardless of the value of λ, Equation (9.20) will always result in $a_0 = 0$ and $a_k = \infty$. With $\hat{\lambda} = 0.084$ and $k = 8$, a_1 is determined from Equation (9.20) as

$$a_1 = -\frac{1}{0.084}\ell n(1 - 0.125) = 1.590$$

Continued application of Equation (9.20) for $i = 2, 3, \ldots, 7$ results in $a_2, \ldots, a_7$ as 3.425, 5.595, 8.252, 11.677, 16.503, and 24.755. Since $k = 8$, $a_8 = \infty$. The first interval is [0, 1.590), the second interval is [1.590, 3.425), and so on. The expectation is that 0.125 of the observations will fall in each interval. The observations, expectations, and the contributions to the calculated value of χ_0^2 are shown in Table 9.7. The calculated value of χ_0^2 is 39.6. The degrees of

Table 9.7. Chi-Square Goodness-of-Fit Test for Example 9.14

Class Interval	*Observed Frequency,* O_i	*Expected Frequency,* E_i	$\frac{(O_i - E_i)^2}{E_i}$
[0, 1.590)	19	6.25	26.01
[1.590, 3.425)	10	6.25	2.25
[3.425, 5.595)	3	6.25	0.81
[5.595, 8.252)	6	6.25	0.01
[8.252, 11.677)	1	6.25	4.41
[11.677, 16.503)	1	6.25	4.41
[16.503, 24.755)	4	6.25	0.81
[24.755, ∞)	6	6.25	0.01
	50	50	39.6

freedom are given by $k - s - 1 = 8 - 1 - 1 = 6$. At $\alpha = 0.05$, the tabulated value of $\chi^2_{0.05,6}$ is 12.6. Since $\chi^2_0 > \chi^2_{0.05,6}$, the null hypothesis is rejected. (The value of $\chi^2_{0.01,6}$ is 16.8, so the null hypothesis would also be rejected at level of significance $\alpha = 0.01$.) ◀

9.4.3 *Kolmogorov-Smirnov Goodness-of-Fit Test*

The chi-square goodness-of-fit test can accommodate the estimation of parameters from the data with a resultant decrease in the degrees of freedom (one for each parameter estimated). The chi-square test requires that the data be placed in class intervals, and in the case of a continuous distributional assumption, this grouping is arbitrary. Changing the number of classes and the interval width affects the value of the calculated and tabulated chi-square. A hypothesis may be accepted when the data are grouped one way but rejected when grouped another way. Also, the distribution of the chi-square test statistic is known only approximately, and the power of the test is sometimes rather low. As a result of these considerations, goodness-of-fit tests, other than the chi-square, are desired. The Kolmogorov-Smirnov test formalizes the idea behind examining a q-q plot.

The Kolmogorov-Smirnov test was presented in Section 7.4.1 to test for the uniformity of numbers and again in Section 7.4.4 to perform the gap test. Both of these uses fall into the category of testing for goodness-of-fit. Any continuous distributional assumption can be tested for goodness-of-fit using the method of Section 7.4.1, while discrete distributional assumptions can be tested using the method of Section 7.4.4.

The Kolmogorov-Smirnov test is particularly useful when sample sizes are small and when no parameters have been estimated from the data. When parameter estimates have been made, the critical values in Table A.8 are biased; in particular, they are too conservative. In this context "conservative" means that the critical values will be too large, resulting in smaller Type I (α) errors

than those specified. The exact value of α can be determined in some instances as discussed at the end of this section.

The Kolmogorov-Smirnov test does not take any special tables when an exponential distribution is assumed. The following example indicates how the test is applied in this instance. (Notice that it is not necessary to estimate the parameter of the distribution in this example, permitting the use of Table A.8.)

EXAMPLE 9.15 (Kolmogorov-Smirnov Test for Exponential Distribution)

Suppose that 50 interarrival times (in minutes) are collected over the following 100-minute interval (arranged in order of occurrence):

0.44	0.53	2.04	2.74	2.00	0.30	2.54	0.52	2.02	1.89	1.53	0.21
2.80	0.04	1.35	8.32	2.34	1.95	0.10	1.42	0.46	0.07	1.09	0.76
5.55	3.93	1.07	2.26	2.88	0.67	1.12	0.26	4.57	5.37	0.12	3.19
1.63	1.46	1.08	2.06	0.85	0.83	2.44	2.11	3.15	2.90	6.58	0.64

The null hypothesis and its alternate are formed as follows:

H_0: the interarrival times are exponentially distributed

H_1: the interarrival times are not exponentially distributed

The data were collected over the interval 0 to $T = 100$ minutes. It can be shown that if the underlying distribution of interarrival times $\{T_1, T_2, \ldots\}$ is exponential, the arrival times are uniformly distributed on the interval $(0, T)$. The arrival times T_1, $T_1 + T_2$, $T_1 + T_2 + T_3, \ldots,$ $T_1 + \cdots + T_{50}$ are obtained by adding interarrival times. The arrival times are then normalized to a (0, 1) interval so that the Kolmogorov-Smirnov test, as presented in Section 7.4.1, can be applied. On a (0, 1) interval, the points will be $[T_1/T, (T_1+T_2)/T, \ldots, (T_1+\cdots+T_{50})/T]$. The resulting 50 data points are as follows:

0.0044	0.0097	0.0301	0.0575	0.0775	0.0805	0.1059	0.1111	0.1313	0.1502
0.1655	0.1676	0.1956	0.1960	0.2095	0.2927	0.3161	0.3356	0.3366	0.3508
0.3553	0.3561	0.3670	0.3746	0.4300	0.4694	0.4796	0.5027	0.5315	0.5382
0.5494	0.5520	0.5977	0.6514	0.6526	0.6845	0.7008	0.7154	0.7262	0.7468
0.7553	0.7636	0.7880	0.7982	0.8206	0.8417	0.8732	0.9022	0.9680	0.9744

Following the procedure in Example 7.6 yields a D^+ of 0.1054 and a D^- of 0.0080. Therefore, the Kolmogorov-Smirnov statistic is $D = \max(0.1054, 0.0080) = 0.1054$. The critical value of D obtained from Table A.8 for a level of significance of $\alpha = 0.05$ and $n = 50$ is $D_{0.05} = 1.36/\sqrt{n} = 0.1923$. Since $D = 0.1054$, the hypothesis that the interarrival times are exponentially distributed cannot be rejected. ◀

The Kolmogorov-Smirnov test has been modified so that it can be used in several situations where the parameters are estimated from the data. The computation of the test statistic is the same, but different tables of critical values are used. Different tables of critical values are required for different distributional assumptions. Lilliefors [1967] developed a test for normality. The null hypothesis states that the population is one of the family of normal distributions without specifying the parameters of the distribution. The interested reader may wish to study Lilliefors' original work, as he describes how simulation was used to develop the critical values.

Lilliefors [1969] also modified the critical values of the Kolmogorov-Smirnov test for the exponential distribution. Lilliefors again used random sampling to obtain approximate critical values, but Durbin [1975] subsequently obtained the exact distribution. Connover [1980] gives examples of Kolmogorov-Smirnov tests for the normal and exponential distributions. He also refers to several other Kolmogorov-Smirnov-type tests which may be of interest to the reader.

A test that is similar in spirit to the Kolmogorov-Smirnov test is the *Anderson-Darling test.* Like the Kolmogorov-Smirnov test, the Anderson-Darling test is based on the difference between the empirical cdf and the fitted cdf; unlike the Kolmogorov-Smirnov test, the Anderson-Darling test is based on a more comprehensive measure of difference (not just the maximum difference) and is more sensitive to discrepancies in the tails of the distributions. The critical values for the Anderson-Darling test also depend on the candidate distribution and on whether or not parameters have been estimated. Fortunately, this test and the Kolmogorov-Smirnov test have been implemented in a number of software packages that support simulation input modeling.

9.4.4 *p-Values and "Best Fits"*

To apply a goodness-of-fit test a significance level must be chosen. Recall that the significance level is the probability of falsely rejecting H_0: the random variable conforms to the distributional assumption. The traditional significance levels are 0.1, 0.05, and 0.01. Prior to the availability of high-speed computing, having a small set of standard values made it possible to produce tables of useful critical values. Now most statistical software computes critical values as needed, rather than storing them in tables. Thus, if the analyst prefers a level of significance of, say, 0.07, then he or she can choose it.

However, rather than require a prespecified significance level, many software packages compute a *p-value* for the test statistic. The p-value is the significance level at which one would *just reject* H_0 for the given value of the test statistic. Therefore, a large p-value tends to indicate a good fit (we would have to accept a large chance of error in order to reject), while a small p-value suggests a poor fit (to accept we would have to insist on almost no risk).

Recall Example 9.13, in which a chi-square test was used to check the Poisson assumption for the vehicle arrival data. The value of the test statistic

was $\chi_0^2 = 27.68$ with 5 degrees of freedom. The p-value for this test statistic is 0.00004, meaning that we would reject the hypothesis that the data are Poisson at the 0.00004 significance level (recall that we rejected the hypothesis at the 0.05 level; now we know that we would also reject it at even lower levels).

The p-value can be viewed as a measure of fit, with larger values being better. This suggests that we could fit every distribution at our disposal, compute a test statistic for each fit, and then choose the distribution that yields the largest p-value. While we know of no input modeling software that implements this specific algorithm, many such packages do include a "best-fit" option in which the software recommends an input model to the user based on evaluating all feasible models. While the software may also take into account other factors — such as whether the data are discrete or continuous, bounded or unbounded — in the end some summary measure of fit, like the p-value, is used to rank the distributions. There is nothing wrong with this, but there are several things to keep in mind:

1. The software may know nothing about the physical basis of the data, and that information can suggest distribution families that are appropriate (see the list in Section 9.2.2). Remember that the goal of input modeling is often to fill in gaps or smooth the data, rather than find an input model that conforms as closely as possible to the given sample.
2. Recall that both the Erlang and the exponential distributions are special cases of the gamma, while the exponential is also a special case of the more flexible Weibull. Automated best-fit procedures tend to choose the more flexible distributions (gamma and Weibull over Erlang and exponential) because the extra flexibility allows closer conformance to the data and a better summary measure of fit. But again, close conformance to the data may not always lead to the most appropriate input model.
3. A summary statistic, like the p-value, is just that, a summary measure. It says little or nothing about where the lack of fit occurs (in the body of the distribution, in the right tail or in the left tail). A human, using graphical tools, can see where the lack of fit occurs and decide whether or not it is important for the application at hand.

Our recommendation is that automated distribution selection be used as one of several ways to suggest candidate distributions. Always inspect the automatic selection using graphical methods, and remember that the final choice is yours.

9.5 Selecting Input Models without Data

Unfortunately, it is often necessary in practice to develop a simulation model —perhaps for demonstration purposes or a preliminary study—before any process data are available. In this case the modeler must be resourceful in choosing input models and must carefully check the sensitivity of results to the chosen models.

There are a number of ways to obtain information about a process even if data are not available:

Engineering data Often a product or process has performance ratings provided by the manufacturer (for example, the mean time to failure of a disk drive is 5000 hours; a laser printer can produce 4 pages/minute; the cutting speed of a tool is 1 inch/second, etc.). Company rules may specify time or production standards. These values provide a starting point for input modeling by fixing a central value.

Expert option Talk to people who are experienced with the process or similar processes. Often they can provide optimistic, pessimistic, and most likely times. They may also be able to say if the process is nearly constant or highly variable, and they may be able to define the source of variability.

Physical or conventional limitations Most real processes have physical limits on performance (for example, computer data entry cannot be faster than a person can type). Because of company policies, there may be upper limits on how long a process may take. Do not ignore obvious limits or bounds that narrow the range of the input process.

The nature of the process The description of the distributions in Section 9.2.2 can be used to justify a particular choice even when no data are available.

When data are not available, the uniform, triangular, and beta distributions are often used as input models. The uniform can be a poor choice, because the upper and lower bounds are rarely just as likely as the central values in real processes. If, in addition to upper and lower bounds, a most-likely value can be given, then the triangular distribution can be used. The triangular distribution places much of its probability near the most-likely value and much less near the extremes (see Section 5.4). If a beta distribution is used, then be sure to plot the density function of the selected distribution, since the beta can take unusual shapes.

A useful refinement is obtained when a minimum, maximum, and one or more "breakpoints" can be given. A breakpoint is an intermediate value and a probability of being less than or equal to that value. The following example illustrates how breakpoints are used.

EXAMPLE 9.16

For a production planning simulation, the sales volume of various products is required. The salesperson responsible for product XYZ–123 says that no fewer than 1000 units will be sold because of existing contracts, and no more than 5000 units will be sold because that is the entire market for the product. Based on her experience, she believes that there is a 90% chance of selling more than 2000 units, a 25% chance of selling more than 3500 units, and only a 1% chance of selling more than 4500 units.

Table 9.8 summarizes this information. Notice that the chances of exceeding certain sales goals have been translated into the cumulative probability of being less than or equal to those goals. With the information in this form the method of Section 8.1.5 can be employed to generate simulation input data. ◀

Table 9.8. Summary of Sales Information

i	*Interval (Hours)*	*Cumulative Frequency,* c_i
1	$1000 \leq x \leq 2000$	0.10
2	$2000 < x \leq 3500$	0.75
3	$3500 < x \leq 4500$	0.99
4	$4500 < x \leq 5000$	1.00

When input models have been selected without data, it is especially important to test the sensitivity of simulation results to the distribution chosen. Check sensitivity not only to the center of the distribution but also to the variability or limits. Extreme sensitivity of output results to the input model provides a convincing argument against making critical decisions based on the results, and in favor of undertaking data collection.

For additional discussion of input modeling in the absence of data, see Pegden, Shannon, and Sadowski [1995].

9.6 Multivariate and Time-Series Input Models

In Sections 9.1–9.4, the random variables presented were considered to be independent of any other variables within the context of the problem. However, variables may be related, and if the variables appear in a simulation model as inputs, the relationship should be determined and taken into consideration.

Example 9.17

An inventory simulation includes the lead time and annual demand for industrial robots. An increase in demand results in an increase in lead time, since the final assembly of the robots must be made according to the specifications of the purchaser. Therefore, rather than treat lead time and demand as independent random variables, a multivariate input model should be developed. ◀

Example 9.18

A simulation of the web-based trading site of a stock broker includes the time between arrivals of orders to buy and sell. Since investors tend to react to what other investors are doing, these buy and sell orders arrive in bursts. Therefore, rather than treat the time between arrivals as independent random variables, a time-series model should be developed. ◀

We distinguish between *multivariate input models* of a fixed, finite number of random variables (such as the two random variables lead time and annual demand in Example 9.17), and *time-series input models* of a (conceptually infinite) sequence of related random variables (such as the successive times between orders in Example 9.18). We describe input models appropriate for these examples after reviewing two measures of dependence, the covariance and correlation.

DURBIN, J. [1975], "Kolmogorov-Smirnov Tests When Parameters Are Estimated with Applications to Tests of Exponentiality and Tests on Spacings," *Biometrika*, Vol. 65, pp. 5–22.

FISHMAN, G. S. [1973], *Concepts and Methods in Discrete Event Digital Simulation*, John Wiley, New York.

GUMBEL, E. J. [1943], "On the Reliability of the Classical Chi-squared Test," *Annals of Mathematical Statistics*, Vol. 14, pp. 253*ff*.

HINES, W. W., AND D. C. MONTGOMERY [1990], *Probability and Statistics in Engineering and Management Science*, 3d ed., John Wiley, New York.

JOHNSON, M. E. [1987], *Multivariate Statistical Simulation*, John Wiley, New York.

LAW, A. M., AND W. D. KELTON [2000], *Simulation Modeling & Analysis*, 3d ed., McGraw-Hill, New York.

LILLIEFORS, H. W. [1967], "On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown," *Journal of the American Statistical Association*, Vol. 62, pp. 339–402.

LILLIEFORS, H. W. [1969], "On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown," *Journal of the American Statistical Association*, Vol. 64, pp. 387–389.

MANN, H. B., AND A. WALD [1942], "On the Choice of the Number of Intervals in the Application of the Chi-squared Test," *Annals of Mathematical Statistics*, Vol. 18, p. 50*ff*.

NELSON, B. L., AND M. YAMNITSKY [1998], "Input Modeling Tools for Complex Problems," in *Proceedings of the 1998 Winter Simulation Conference*, eds. D. Medeiros, E. Watson, J. Carson and M. Manivannan, pp. 105–112, The Institute for Electrical and Electronics Engineers, Piscataway, NJ.

PEGDEN, C. D., R. E. SHANNON, AND R. P. SADOWSKI [1995], *Introduction to Simulation using SIMAN*, 2d ed. McGraw-Hill, New York.

STUART, A., J. K. ORD AND E. ARNOLD [1998], *Kendall's Advanced Theory of Statistics*, 6th ed., Vol. 2, Oxford University Press, Oxford.

EXERCISES

1. Go to a small appliance store and determine the interarrival and service-time distributions. If there are several workers, how do the service-time distributions compare to each other? Do service-time distributions need to be constructed for each type of appliance? (Make sure that the management gives permission to perform this study.)

2. Go to a cafeteria and collect data on the distributions of interarrival and service times. The distribution of interarrival times is probably different for each of the three daily meals and may also vary during the meal; that is, the interarrival time distribution for 11:00 A.M. to 12:00 noon may be different than from 12:00 noon to 1:00 P.M. Define service time as the time from when the customer reaches the point at which the first selection could be made until he or she exits from the cafeteria line. (Any reasonable modification of this definition is acceptable.) The service-time distribution probably changes for each meal. Can times of the day or days

of the week for either distribution be grouped due to homogeniety of the data? (Make sure that the management gives permission to perform this study.)

3. Go to a major traffic intersection and determine the interarrival-time distributions from each direction. Some arrivals want to go straight, some turn left, some turn right. The interarrival-time distribution varies during the day and by day of the week. Every now and then an accident occurs.
4. Go to a grocery store and determine the interarrival and service distributions at the checkout counters. These distributions may vary by time of day and by day of week. Record, also, the number of service channels available at all times. (Make sure that the management gives permission to perform this study.)
5. Go to a laundromat and "relive" the authors' data-collection experience as discussed in Example 9.1. (Make sure that the management gives permission to perform this study.)
6. Prepare four theoretical normal density functions, all on the same figure, each distribution having mean zero, but let the standard deviations be 1/4, 1/2, 1, and 2.
7. On one figure, draw the pdfs of the Erlang distribution where $\theta = 1/2$ and $k = 1, 2, 4$, and 8.
8. On one figure, draw the pdfs of the Erlang distribution where $\theta = 2$ and $k = 1, 2, 4$, and 8.
9. Draw the pmf of the Poisson distribution that results when the parameter α is equal to the following:
   (a) $\alpha = 1/2$
   (b) $\alpha = 1$
   (c) $\alpha = 2$
   (d) $\alpha = 4$
10. On one figure draw the two exponential pdf's that result when the parameter, λ, equals 0.6 and 1.2.
11. On one figure draw the three Weibull pdf's which result when $\nu = 0$, $\alpha = 1/2$, and $\beta = 1, 2$, and 4
12. The following data are randomly generated from a gamma distribution:

1.691	1.437	8.221	5.976
1.116	4.435	2.345	1.782
3.810	4.589	5.313	10.90
2.649	2.432	1.581	2.432
1.843	2.466	2.833	2.361

Determine the maximum-likelihood estimators $\widehat{\beta}$ and $\widehat{\theta}$.

13. The following data are randomly generated from a Weibull distribution where $\nu = 0$:

7.936	5.224	3.937	6.513
4.599	7.563	7.172	5.132
5.259	2.759	4.278	2.696
6.212	2.407	1.857	5.002
4.612	2.003	6.908	3.326

Determine the maximum-likelihood estimators $\hat{\alpha}$ and $\hat{\beta}$. (This exercise requires a programmable calculator, a computer, or a lot of patience.)

14. The highway between Atlanta, Georgia, and Athens, Georgia, has a high incidence of accidents along its 100 kilometers. Public safety officers say that the occurrence of accidents along the highway is randomly (uniformly) distributed, but the news media say otherwise. The Georgia Department of Public Safety published records for the month of September. These records indicated the point at which 30 accidents involving an injury or death occurred, as follows (the data points represent the distance from the city limits of Atlanta):

88.3	40.7	36.3	27.3	36.8
91.7	67.3	7.0	45.2	23.3
98.8	90.1	17.2	23.7	97.4
32.4	87.8	69.8	62.6	99.7
20.6	73.1	21.6	6.0	45.3
76.6	73.2	27.3	87.6	87.2

Use the Kolmogorov-Smirnov test to determine whether the distribution of location of accidents is uniformly distributed for the month of September.

15. Show that the Kolmogorov-Smirnov test statistic for Example 9.15 is $D = 0.1054$.

16. Records pertaining to the monthly number of job-related injuries at an underground coal mine were being studied by a federal agency. The values for the past 100 months were as follows:

Injuries per Month	*Frequency of Occurrence*
0	35
1	40
2	13
3	6
4	4
5	1
6	1

(a) Apply the chi-square test to these data to test the hypothesis that the underlying distribution is Poisson. Use a level of significance of $\alpha = 0.05$.

(b) Apply the chi-square test to these data to test the hypothesis that the distribution is Poisson with mean 1.0. Again let $\alpha = 0.05$.

(c) What are the differences in parts (a) and (b), and when might each case arise?

17. The time required for 50 different employees to compute and record the number of hours worked during the week was measured with the following results in minutes:

Employee	Time (Minutes)	Employee	Time (Minutes)
1	1.88	26	0.04
2	0.54	27	1.49
3	1.90	28	0.66
4	0.15	29	2.03
5	0.02	30	1.00
6	2.81	31	0.39
7	1.50	32	0.34
8	0.53	33	0.01
9	2.62	34	0.10
10	2.67	35	1.10
11	3.53	36	0.24
12	0.53	37	0.26
13	1.80	38	0.45
14	0.79	39	0.17
15	0.21	40	4.29
16	0.80	41	0.80
17	0.26	42	5.50
18	0.63	43	4.91
19	0.36	44	0.35
20	2.03	45	0.36
21	1.42	46	0.90
22	1.28	47	1.03
23	0.82	48	1.73
24	2.16	49	0.38
25	0.05	50	0.48

Use the chi-square test (as in Example 9.14) to test the hypothesis that these service times are exponentially distributed. Let the number of class intervals be $k = 6$. Use a level of significance of $\alpha = 0.05$.

18. Studentwiser Beer Company is trying to determine the distribution of the breaking strength of their glass bottles. Fifty bottles are selected at random and tested for breaking strength, with the following results (in pounds per square inch):

218.95	232.75	212.80	231.10	215.95
237.55	235.45	228.25	218.65	212.80
230.35	228.55	216.10	229.75	229.00
199.75	225.10	208.15	213.85	205.45
219.40	208.15	198.40	238.60	219.55
243.10	198.85	224.95	212.20	222.90
218.80	203.35	223.45	213.40	206.05
229.30	239.20	201.25	216.85	207.25
204.85	219.85	226.15	230.35	211.45
227.95	229.30	225.25	201.25	216.10

Using input modeling sofware, apply as many tests for normality as are available in the software. If the chi-square test is available, apply it with at least two different choices for the number of intervals. Do all of the tests reach the same conclusion?

19. The Crosstowner is a bus that cuts a diagonal path from northeast Atlanta to southwest Atlanta. The time required to complete the route is maintained by the bus operator. The bus runs Monday through Friday. The times of the last fifty 8:00 A.M. runs, in minutes, are as follows:

92.3	92.8	106.8	108.9	106.6
115.2	94.8	106.4	110.0	90.9
104.6	72.0	86.0	102.4	99.8
87.5	111.4	105.9	90.7	99.2
97.8	88.3	97.5	97.4	93.7
99.7	122.7	100.2	106.5	105.5
80.7	107.9	103.2	116.4	101.7
84.8	101.9	99.1	102.2	102.5
111.7	101.5	95.1	92.8	88.5
74.4	98.9	111.9	96.5	95.9

How are these run times distributed? Develop and test a suitable model.

20. The time required for the transmission of a message (in minutes) is sampled electronically at a communications center. The last 50 values in the sample are as follows:

7.936	4.612	2.407	4.278	5.132
4.599	5.224	2.003	1.857	2.696
5.259	7.563	3.937	6.908	5.002
6.212	2.759	7.172	6.513	3.326
8.761	4.502	6.188	2.566	5.515
3.785	3.742	4.682	4.346	5.359
3.535	5.061	4.629	5.298	6.492
3.502	4.266	3.129	1.298	3.454
5.289	6.805	3.827	3.912	2.969
4.646	5.963	3.829	4.404	4.924

How are the transmission times distributed? Develop and test an appropriate model.

21. The time (in minutes) between requests for the hookup of electric service was accurately maintained at the Gotwatts Flash and Flicker Company with the following results for the last 50 requests:

0.661	4.910	8.989	12.801	20.249
5.124	15.033	58.091	1.543	3.624
13.509	5.745	0.651	0.965	62.146
15.512	2.758	17.602	6.675	11.209
2.731	6.892	16.713	5.692	6.636
2.420	2.984	10.613	3.827	10.244
6.255	27.969	12.107	4.636	7.093
6.892	13.243	12.711	3.411	7.897
12.413	2.169	0.921	1.900	0.315
4.370	0.377	9.063	1.875	0.790

How are the times between requests for service distributed? Develop and test a suitable model.

22. Daily demands for transmission overhaul kits for the D–3 dragline were maintained by Earth Moving Tractor Company with the following results:

0	2	0	0	0
1	0	1	1	1
0	1	0	0	0
2	0	1	0	1
0	1	0	0	2
1	0	1	0	0
0	0	0	0	0
1	0	1	0	1
0	0	3	0	1
1	0	0	0	0

How are the daily demands distributed? Develop and test an appropriate model.

23. A simulation is to be conducted of a job shop that performs two operations, milling and planing, in that order. It would be possible to collect data about processing times for each operation, then generate random occurrences from each distribution. However, the shop manager says that the times might be related; large milling jobs take lots of planing. Data are collected for the next 25 orders with the following results in minutes:

Order	*Milling Time (Minutes)*	*Planing Time (Minutes)*	*Order*	*Milling Time (Minutes)*	*Planing Time (Minutes)*
1	12.3	10.6	14	24.6	16.6
2	20.4	13.9	15	28.5	21.2
3	18.9	14.1	16	11.3	9.9
4	16.5	10.1	17	13.3	10.7
5	8.3	8.4	18	21.0	14.0
6	6.5	8.1	19	19.5	13.0
7	25.2	16.9	20	15.0	11.5
8	17.7	13.7	21	12.6	9.9
9	10.6	10.2	22	14.3	13.2
10	13.7	12.1	23	17.0	12.5
11	26.2	16.0	24	21.2	14.2
12	30.4	18.9	25	28.4	19.1
13	9.9	7.7			

(a) Plot milling time on the horizontal axis and planing time on the vertical axis. Do these data seem dependent?

(b) Compute the sample correlation between milling time and planing time.

(c) Fit a bivariate normal distribution to this data.

24. Write a computer program to compute the maximum-likelihood estimates $(\hat{\alpha}, \hat{\beta})$ of the Weibull distribution. Inputs to the program should include the sample size, n; the observations, $x_1, x_2, \ldots, x_n$; a stopping criterion, ϵ (stop when $|f(\hat{\beta}_j)| \leq \epsilon$); and a print option, OPT (usually set = 0). Output would be the estimates $\hat{\alpha}$ and $\hat{\beta}$. If OPT = 1, additional output would be printed as in Table 9.4 showing convergence. Make the program as "user friendly" as possible.
25. Examine a computer software library or simulation support environment to which you have access. Obtain documentation on data-analysis software that would be useful in solving Exercises 7 through 24. Use the software as an aid in solving selected problems.
26. The numbers of patrons staying at a small hotel on 20 successive nights were observed to be 20, 14, 21, 19, 14, 18, 21, 25, 27, 26, 22, 18, 13, 18, 18, 18, 25, 23, 20, 21. Fit both an AR(1) and an EAR(1) model to this data. Decide which model provides a better fit by looking at a histogram of the data.
27. The following data represent the time to perform transactions in a bank, measured in minutes: 0.740, 1.28, 1.46, 2.36, 0.354, 0.750, 0.912, 4.44, 0.114, 3.08, 3.24, 1.10, 1.59, 1.47, 1.17, 1.27, 9.12, 11.5, 2.42, 1.77. Develop an input model for this data.
28. Two types of jobs (A and B) are released to the input buffer of a job shop as orders arrive, and the arrival of orders is uncertain. The following data are available from the last week of production:

Day	*Number of Jobs*	*Number of A's*
1	83	53
2	93	62
3	112	66
4	65	41
5	78	55

Develop an input model for the number of new arrivals of each type each day.

29. The following data are available on the processing time at a machine (in minutes): 0.64, 0.59, 1.1, 3.3, 0.54, 0.04, 0.45, 0.25, 4.4, 2.7, 2.4, 1.1, 3.6, 0.61, 0.20, 1.0, 0.27, 1.7, 0.04, 0.34. Develop an input model for the processing time.