L A B

6 Fitting Statistical Distributions to Input Data

There are three kinds of lies: lies, damned lies, and statistics.
—Benjamin Disraeli

Input data drive our simulation models. Input data can be for interarrival times, material handling times, setup and process times, demand rates, loading and unloading times, and so forth. The determination of what data to use and where to get the appropriate data is a complicated and time-consuming task. The quality of data is also very important. We have all heard the cliché "garbage in, garbage out." In Chapter 6 we discussed various issues about input data collection and analysis. We have also described various empirical discrete and continuous distributions and their characteristics. In this lab we describe how ProModel helps in fitting empirical statistical distributions to user input data.

L6.1 An Introduction to Stat::Fit

Stat::Fit is a utility packaged with the ProModel software, which is available from the opening screen of ProModel (Figure L1.1) and also from the Windows Start → Programs menu. The Stat::Fit opening screen is shown in Figure L6.1. Stat::Fit can be used for analyzing user-input data and fitting an appropriate empirical distribution. The empirical distribution can be either continuous or discrete. If there are enough data points, say 100 or more, it may be appropriate to fit an empirical distribution to the data using conventional methods. ProModel offers built-in capability to perform input data analysis with the tool Stat::Fit. Stat::Fit fits probability distributions to empirical data. It allows comparison among various distribution functions. It performs goodness-of-fit tests using chi-square, Kolmogorov–Smirnov, and Anderson–Darling procedures. It calculates appropriate parameters for distributions. It provides distribution expressions for use in the simulation model. When the amount of data is

FIGURE L6.1
Stat::Fit opening screen.

FIGURE L6.2
Stat::Fit opening menu.

small, the goodness-of-fit tests are of little use in selecting one distribution over another because it is inappropriate to fit one distribution over another in such a situation. Also, when conventional techniques have failed to fit a distribution, the empirical distribution is used directly as a user distribution (Chapter 6, Section 6.9).

The opening menu of Stat::Fit is shown in Figure L6.2. Various options are available in the opening menu:

1. *File:* File opens a new Stat::Fit project or an existing project or data file. The File menu is also used to save a project.
2. *Edit:*
3. *Input:*
4. *Statistics:*
5. *Fit:* The Fit menu provides a Fit Setup dialog and a Distribution Graph dialog. Other options are also available when a Stat::Fit project is opened. The Fit Setup dialog lists all the distributions supported by Stat::Fit and the relevant choices for goodness-of-fit tests. At least one distribution must be chosen before the estimate, test, and graphing commands become available. The Distribution Graph command uses the distribution and parameters provided in the Distribution Graph dialog to create a graph of any analytical distribution supported by Stat::Fit. This graph is not connected to any input data or document.

6. *Utilities:* The Replications command allows the user to calculate the number of independent data points or replications of an experiment necessary to provide a given range or confidence interval for the estimate of a parameter. The confidence interval is given for the confidence level specified. The default is a 95 percent confidence interval. The resulting number of replications is calculated using the *t* distribution.
7. *View:*
8. *Window:* The Window menu is used to either cascade or tile various windows opened while working on the Stat::Fit projects.
9. *Help:*

Figure L6.3 shows the data/document input screen. Figure L6.4 shows the various data input options available in Stat::Fit. The type of distribution is also specified

FIGURE L6.3
Document input screen.

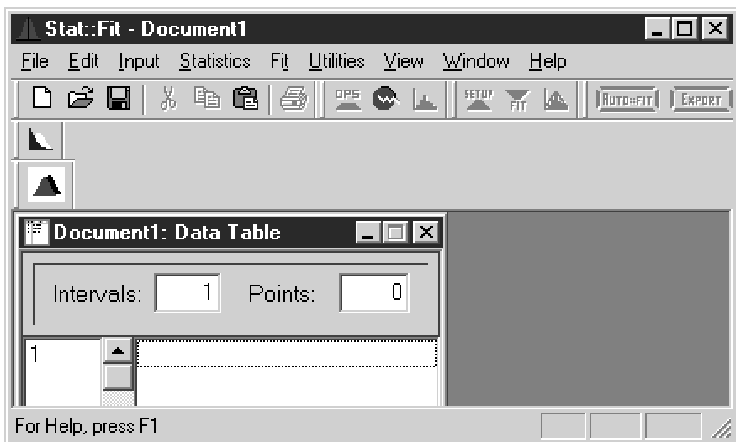
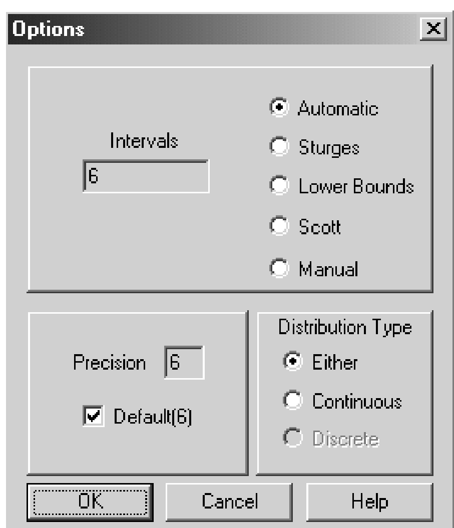


FIGURE L6.4
Stat::Fit data input options.



here. The Input Options command can be accessed from the Input menu as well as the Input Options button on the Speed Bar.

L6.2 An Example Problem

Problem Statement

The time between arrivals of cars at **San Dimas Gas Station** were collected as shown in Table L6.1. This information is also saved and available in an Excel worksheet named *L6.2_Gas Station Time Between Arrival*. Use Stat::Fit to analyze the data and fit an appropriate continuous distribution to the data. Figure L6.5 shows part of the actual data of times between arrival of cars in minutes. The sample data have 30 data points. Figure L6.6 shows the histogram of the input data, while Figure L6.7 shows some of the descriptive statistics generated by Stat::Fit.

TABLE L6.1 Times between Arrival of Cars at San Dimas Gas Station

Number	*Times between Arrival, Minutes*
1	12.36
2	5.71
3	16.79
4	18.01
5	5.12
6	7.69
7	19.41
8	8.58
9	13.42
10	15.56
11	10.
12	18.
13	16.75
14	14.13
15	17.46
16	10.72
17	11.53
18	18.03
19	13.45
20	10.54
21	12.53
22	8.91
23	6.78
24	8.54
25	11.23
26	10.1
27	9.34
28	6.53
29	14.35
30	18.45

FIGURE L6.5

Times between arrival of cars at San Dimas Gas Station.

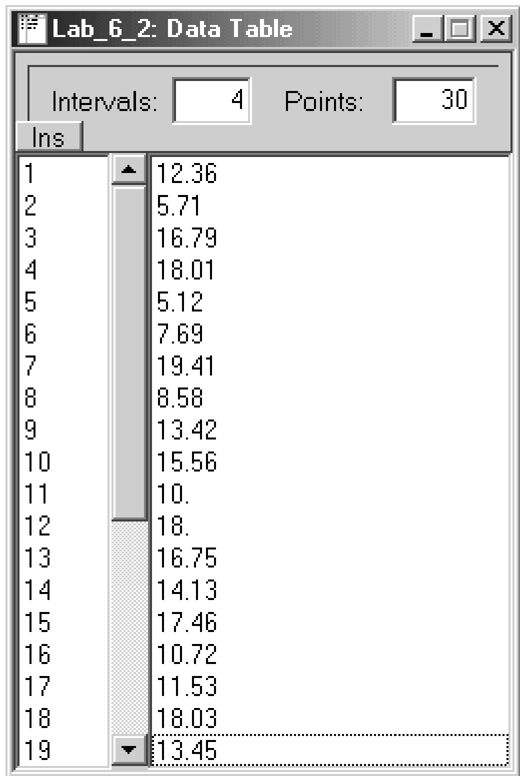


FIGURE L6.6

Histogram of the times between arrival data.

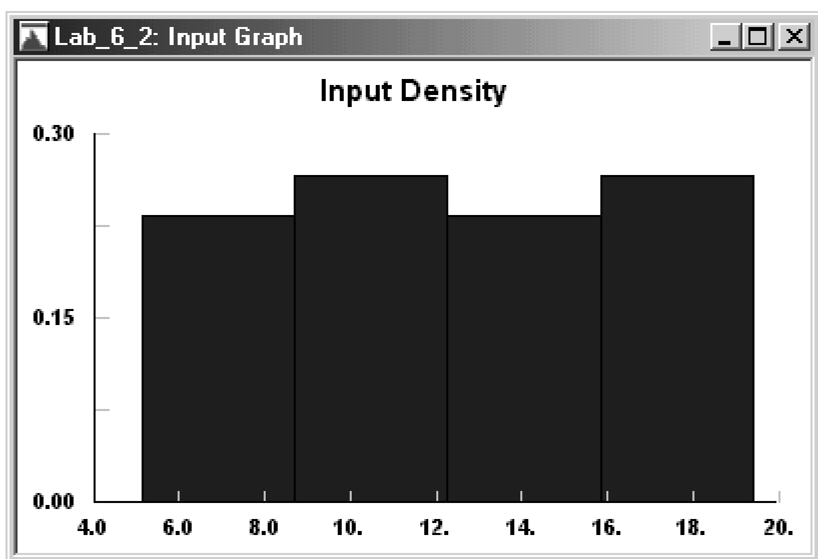
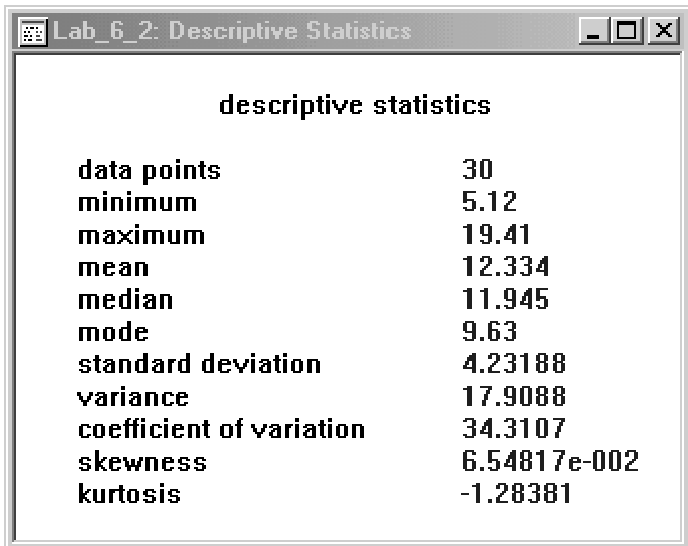


FIGURE L6.7
Descriptive statistics for the input data.



L6.3 Auto::Fit Input Data

Automatic fitting of continuous distributions can be performed by clicking on the Auto::Fit icon or by selecting Fit from the Menu bar and then Auto::Fit from the Submenu (Figure L6.8).

This command follows the same procedure as discussed in Chapter 6 for manual fitting. Auto::Fit will automatically choose appropriate continuous distributions to fit to the input data, calculate Maximum Likelihood Estimates for those distributions, test the results for Goodness of Fit, and display the distributions in order of their relative rank. The relative rank is determined by an empirical method, which uses effective goodness-of-fit calculations. While a good rank usually indicates that the fitted distribution is a good representation of the input data, an absolute indication of the goodness of fit is also given.

The Auto::Fit dialog allows the number of continuous distributions to be limited by choosing only those distributions with a lower bound or by forcing a lower bound to a specific value as in Fit Setup. Also, the number of distributions will be limited if the skewness of the input data is negative; many continuous distributions with lower bounds do not have good parameter estimates in this situation.

The acceptance of fit usually reflects the results of the goodness-of-fit tests at the level of significance chosen by the user. However, the acceptance may be modified if the fitted distribution would generate significantly more data points in the tails of the distribution than are indicated by the input data.

The Auto::Fit function forces the setup of the document so that *only continuous distributions will be used*. Figure L6.9 shows various continuous distributions fitted to the input data of the San Dimas Gas Station (Section L6.2) and their

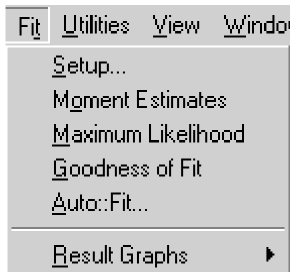
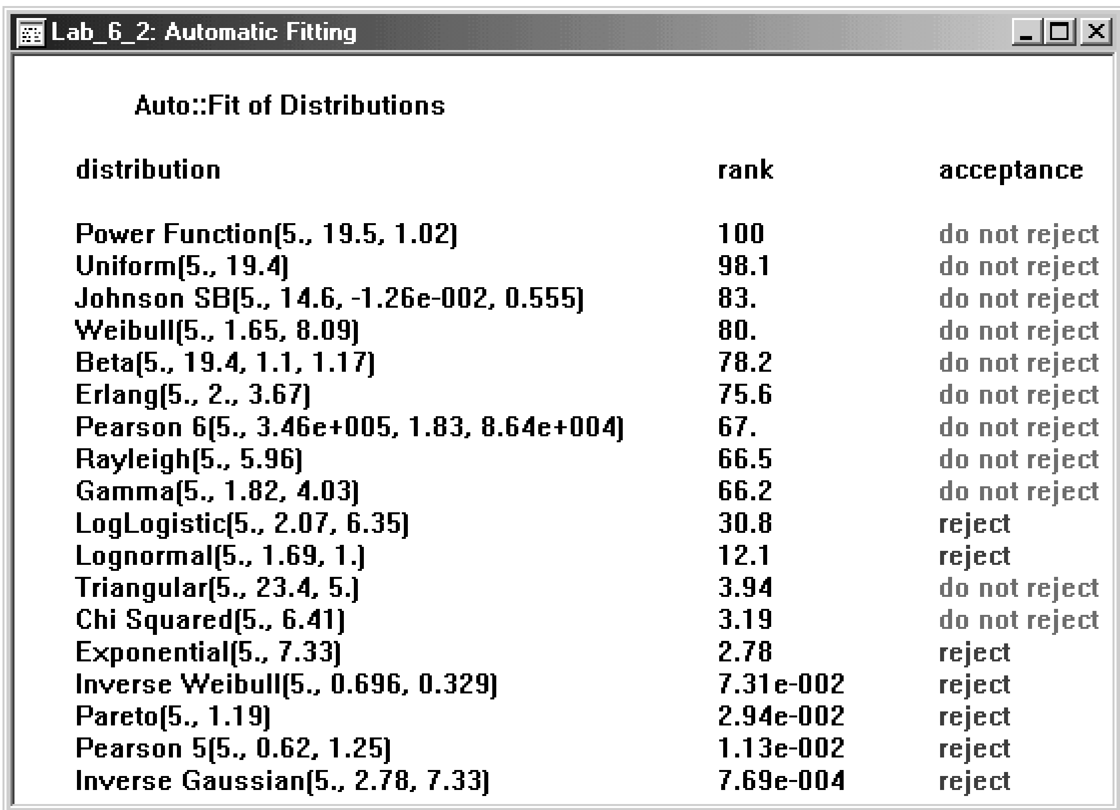


FIGURE L6.8
The Auto::Fit submenu.

FIGURE L6.9
Various distributions fitted to the input data.



distribution	rank	acceptance
Power Function(5., 19.5, 1.02)	100	do not reject
Uniform(5., 19.4)	98.1	do not reject
Johnson SB(5., 14.6, -1.26e-002, 0.555)	83.	do not reject
Weibull(5., 1.65, 8.09)	80.	do not reject
Beta(5., 19.4, 1.1, 1.17)	78.2	do not reject
Erlang(5., 2., 3.67)	75.6	do not reject
Pearson 6(5., 3.46e+005, 1.83, 8.64e+004)	67.	do not reject
Rayleigh(5., 5.96)	66.5	do not reject
Gamma(5., 1.82, 4.03)	66.2	do not reject
LogLogistic(5., 2.07, 6.35)	30.8	reject
Lognormal(5., 1.69, 1.)	12.1	reject
Triangular(5., 23.4, 5.)	3.94	do not reject
Chi Squared(5., 6.41)	3.19	do not reject
Exponential(5., 7.33)	2.78	reject
Inverse Weibull(5., 0.696, 0.329)	7.31e-002	reject
Pareto(5., 1.19)	2.94e-002	reject
Pearson 5(5., 0.62, 1.25)	1.13e-002	reject
Inverse Gaussian(5., 2.78, 7.33)	7.69e-004	reject

rank in terms of the amount of fit. Both the Kolmogorov–Smirnov and the Anderson–Darling goodness-of-fit tests will be performed on the input data as shown in Figure L6.10. The Maximum Likelihood Estimates will be used with an accuracy of at least 0.00003. The actual data and the fitted uniform distribution are compared and shown in Figure L6.11.

FIGURE L6.10

Goodness-of-fit tests performed on the input data.

goodness of fit

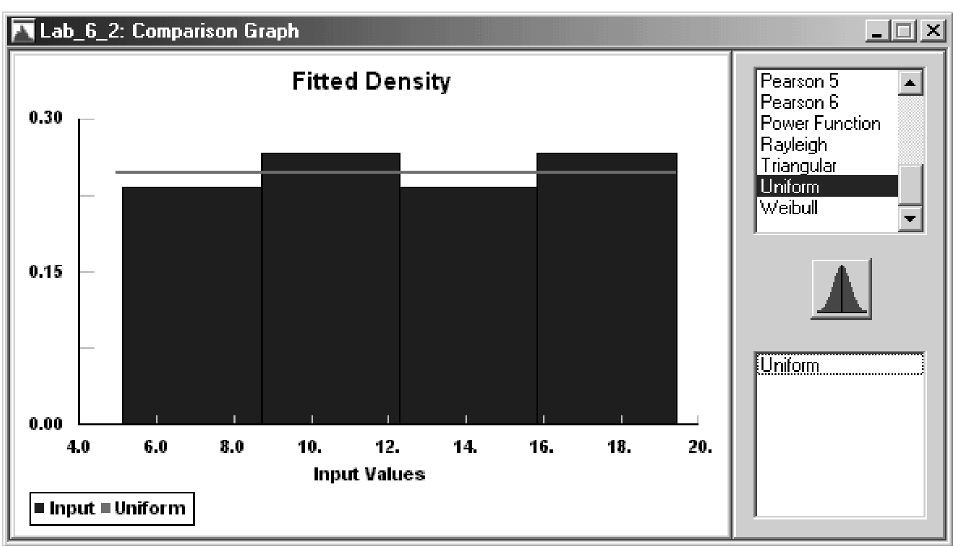
data points	30
estimates	maximum likelihood estimates
accuracy of fit	3.e-004
level of significance	5.e-002

summary

distribution	Kolmogorov Smirnov	Anderson Darling
Beta(5., 19.4, 1.1, 1.17)	0.114	2.23
Chi Squared(5., 6.41)	0.19	2.41
Erlang(5., 2., 3.67)	0.102	0.706
Exponential(5., 7.33)	0.216	1.97
Gamma(5., 1.82, 4.03)	0.109	0.75
Inverse Gaussian(5., 2.78, 7.33)	0.358	5.76
Inverse Weibull(5., 0.696, 0.329)	0.287	3.4
Johnson SB(5., 14.6, -1.26e-002, 0.555)	9.34e-002	0.26
LogLogistic(5., 2.07, 6.35)	0.155	0.875
Lognormal(5., 1.69, 1.)	0.168	1.61
Pareto(5., 1.19)	0.305	3.96
Pearson 5(5., 0.62, 1.25)	0.319	4.51
Pearson 6(5., 3.46e+005, 1.83, 8.64e+00	0.108	0.747
Power Function(5., 19.5, 1.02)	7.44e-002	0.199
Rayleigh(5., 5.96)	0.123	0.827
Triangular(5., 23.4, 5.)	0.182	1.62
Uniform(5., 19.4)	8.21e-002	2.07
Weibull(5., 1.65, 8.09)	0.109	0.522

FIGURE L6.11

Comparison of actual data and fitted uniform distribution.



Because the Auto::Fit function requires a specific setup, the Auto::Fit view can be printed only as the active window or part of the active document, not as part of a report. The Auto::Fit function will not fit discrete distributions. The manual method, previously described, should be used instead.

L6.4 Exercises

1. Consider the operation of a fast-food restaurant where customers arrive for ordering lunch. The following is a log of the time (minutes) between arrivals of 40 successive customers. Use Stat::Fit to analyze the data and fit an appropriate continuous distribution. What are the parameters of this distribution?

11	11	12	8	15	14	15	13
9	13	14	9	14	9	13	7
12	12	7	13	12	16	7	10
8	8	17	15	10	7	16	11
11	10	16	10	11	12	14	15

2. The servers at the restaurant in Question 1 took the following time (minutes) to serve food to these 40 customers. Use Stat::Fit to analyze the data and fit an appropriate continuous distribution. What are the parameters of this distribution?

11	11	12	8	15	14	15	13
9	13	14	10	14	9	13	12
12	12	11	13	12	16	11	10
10	8	17	12	10	7	13	11
11	10	13	10	11	12	14	15

3. The following are the numbers of incoming calls (each hour for 80 successive hours) to a call center set up for serving customers of a certain Internet service provider. Use Stat::Fit to analyze the data and fit an appropriate discrete distribution. What are the parameters of this distribution?

12	12	11	13	12	16	11	10
9	13	14	10	14	9	13	12
12	12	11	13	12	16	11	10
10	8	17	12	10	7	13	11
11	11	12	8	15	14	15	13
9	13	14	10	14	9	13	12
12	12	11	13	12	16	11	10
10	8	17	12	10	7	13	11
11	10	13	10	11	12	14	15
10	8	17	12	10	7	13	11

4. Observations were taken on the times to serve online customers at a stockbroker's site (STOCK.com) on the Internet. The times (in seconds) are shown here, sorted in ascending order. Use Stat::Fit and fit an appropriate distribution to the data. What are the parameters of this distribution?

1.39	21.47	39.49	58.78	82.10
3.59	22.55	39.99	60.61	83.52
7.11	28.04	41.42	63.38	85.90
8.34	28.97	42.53	65.99	88.04
11.14	29.05	47.08	66.00	88.40
11.97	35.26	51.53	73.55	88.47
13.53	37.65	55.11	73.81	92.63
16.87	38.21	55.75	74.14	93.11
17.63	38.32	55.85	79.79	93.74
19.44	39.17	56.96	81.66	98.82

5. Forty observations for a bagging operation was shown in Chapter 6, Table 6.5. Use Stat::Fit to find out what distribution best fits the data. Compare your results with the results obtained in Chapter 6 (Suggestion: In Chapter 6 five equal probability cells were used for the chi-square goodness of fit test. Use same number of cells in Stat::Fit to match the results).