ExpertFit®
Student Version 8
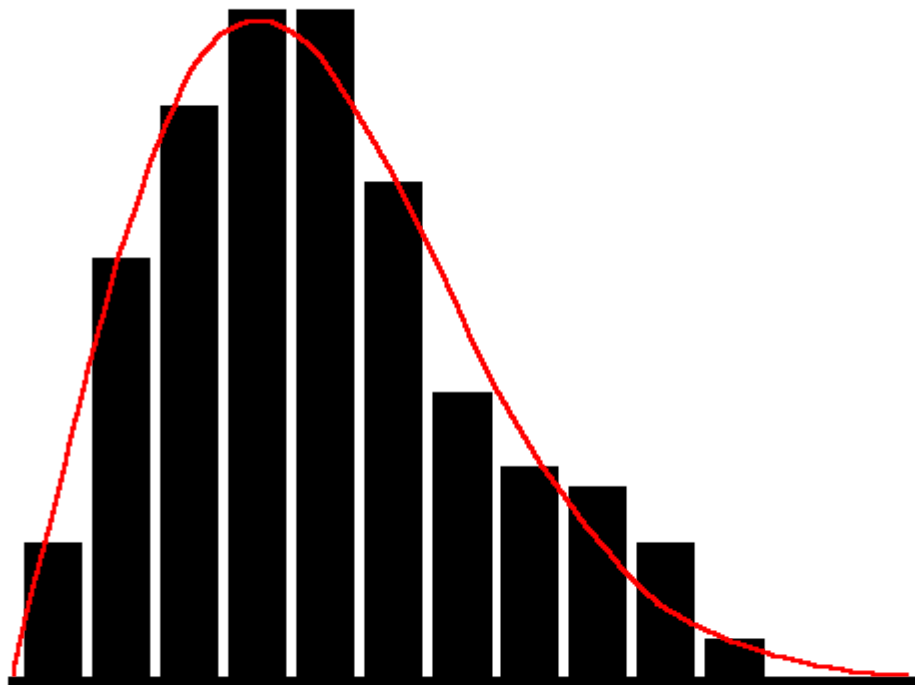
1. Introduction

ExpertFit allows one to determine automatically and accurately which probability distribution best represents a data set. In many cases a complete analysis can be done in less than 5 minutes. A secondary goal is to provide simulation analysts with assistance in modeling a source of randomness (e.g., a service time) in the absence of data (see the discussion in Section 1.1).

ExpertFit is extensively used by analysts performing discrete-event simulation studies of real-world systems in application areas such as defense, manufacturing, transportation, healthcare, call centers, and communications networks. For these users, ExpertFit will take the selected distribution and put it into the proper format for direct input to a large number of different simulation-software products. ExpertFit is also used for data analyses in such diverse disciplines as actuarial science, agriculture, chemistry, economics, environmental analysis, finance, forestry, hydrology, medicine, meteorology, mining, physics, psychology, reliability engineering, and risk analysis. ExpertFit is the result of 33 years of statistical research.

ExpertFit has extensive context-sensitive online help for every options and results screen, a **built-in User's Guide** (see the **Help** pull-down menu at the top of the screen), and a Feature Index. There is a glossary of key terms and also tutorials on a number of general topics such as the available probability distributions. All ExpertFit results can be printed or copied to the Windows Clipboard and used in other applications (e.g., Microsoft Word or Excel).

The Student Version of ExpertFit is designed to be used with the 5th edition of the book *Simulation Modeling and Analysis* by Dr. Averill M. Law. It includes 13 different data sets (see Section 3) and allows one to use most of the features available in the commercial version of ExpertFit on **these** data sets.

A large amount of additional information on ExpertFit can be found on our website www.averill-law.com (see "ExpertFit").

1.1. Types of ExpertFit Analyses

ExpertFit can perform the three main types of analyses given in Table 1.1.

Table 1.1. Main types of ExpertFit analyses.

Module	Description
Data Analysis	Used to determine what probability distribution best represents a data set. You can either have ExpertFit determine the best distribution automatically or specify the distributions for consideration manually. See Chapter 2 for further discussion and an example.
Task-Time Models	Used to specify a probability distribution for a task time when no data are available. Based on subjective estimates of the minimum task time, the most-likely task time, and, say, the 90th percentile of the task time, ExpertFit specifies a Weibull, lognormal, or triangular distribution as a model for the task time. **This analysis type is not available in the Student Version.**
Machine-Breakdown Models	Used to model the random downtimes of a machine when no downtime data are available. Based on subjective estimates of such parameters as machine efficiency (e.g., 0.90) and mean downtime, ExpertFit specifies a busy-time distribution and a downtime distribution. **This analysis type is not available in the Student Version.**

In addition, the important features given in Table 1.2 can be accessed from the Menu Bar at the top of the screen.

Table 1.2. Additional important ExpertFit features.

Module	Description
Distribution Viewer	Used to display/calculate characteristics (e.g., the density function or moments) of a distribution without having to enter a data set.
Batch Mode	Used to fit distributions to one or more data sets with only a few mouse clicks.

2. Data Analysis Module

There are two modes of operation (see the **Mode** pull-down menu) for the **Data Analysis** module: Standard and Advanced. **Standard Mode** is sufficient for 95 percent of all analyses and is easier to use. It focuses the user on those features that are the most important at a particular point in an analysis. **Advanced Mode** contains a large number of additional features for the sophisticated user. A user can switch from one mode to another at any time. In this brochure we will only discuss the **Standard Mode**.

There are also two levels of precision when fitting distributions (see the **Precision** pull-down menu): Normal and High. **Normal Precision** provides good estimates, for many data sets, of the parameters of a distribution and has a small execution time. **High Precision** (the default) provides better parameter estimates for most data sets, but will have a larger execution time for data sets containing many observations.

The use of the **Data Analysis** module to determine what probability distribution best represents a data set is based on the sequential application of the four tabs shown in Table 2.1.

Table 2.1. Tabs for the Data Analysis module.

Tab	Overall Purpose
Data	Used to read in a data set from a file, enter a data set at the keyboard, paste in a sample from the Clipboard, or import a data set from Excel
Models	Used to "fit" probability distributions to a data set
Comparisons	Used to compare the fitted distributions to the data set
Applications	Used to determine or display characteristics of a distribution (e.g., its density function) or to represent the distribution in a simulation-software package

Although there are different ways that these four tabs could be used to determine the best distribution for a data set, the following are the explicit steps that we recommend for real-valued data:

1. Obtain a data set using the **Data** tab.

2. View the resulting **Data-Summary Table** (in **Data** tab) – provides information on the shape and range of the true underlying density function.

3. Make a histogram of your data (used in Step 5) using the **Data** tab – see the **Constructing a Histogram from Your Data** tutorial in the **Help** pull-down menu.

4. Determine the distribution that is the best representation for your data using the **Automated Fitting** option in the **Models** tab.

5. Confirm using the **Comparisons** tab that the best distribution as determined by ExpertFit is, in fact, satisfactory in an absolute sense.

6. If you are doing simulation modeling, then either represent the best distribution (if good in an absolute sense) or an empirical distribution based on your data (if the best distribution is not satisfactory) in your simulation software using the **Applications** tab.

An example of the use of the **Data Analysis** module is given next.

Example 2.1: *Customer Service Times*

This example illustrates how you can use the six-step approach given at the beginning of this chapter to find the probability distribution that best represents a data set. We discuss both the ExpertFit commands necessary to accomplish a particular part of the analysis and the actual results of the analysis. This analysis will use **High Precision.**

Steps for Action A:

At window:	*Do:*
Project 1	*Click on* **New**.
Project-Element Editing	*Select* **Fit distributions to data**.
	In the **Project-Element Name** *edit box, enter* **Example 2.1**.
	Click on **OK**.
Project 1	*Click on* **Analyze**.
Data tab	*Click on* **Enter Data**.
Enter-Data Options	*Click on* **Apply**.
Open	*In the* **File Name** *scroll list, select* **EXAM21_UsersGuide.EXP.**
	Examine the **Data-Summary Table**.
Data-Summary Table	*Click on* **Done**.

A: The **Data-Summary Table** for this set of $n = 450$ service times (read in the **Data** tab) is given in Table 2.2. The positive value of the sample skewness indicates that the underlying distribution of the data is skewed to the right (i.e., it has a longer right tail than left tail). This is supported by the sample mean being larger than the sample median.

Table 2.2. Data summary for the service-time data.

Data Characteristic	Value
Source file	EXAM21_UsersGuide
Observation type	Real valued
Number of observations	450
Minimum observation	0.06438
Maximum observation	3.11115
Mean	1.18250
Median	1.15192
Variance	0.32394
Coefficient of variation	0.48131
Skewness	0.48796

B: In Figure 2.1 we present the default ExpertFit histogram for the service-time data. Note that the histogram is quite "ragged," since the interval width is too small. Using a trial-and-error approach discussed in the **Constructing a Histogram from Your Data** tutorial, we determined that a better histogram is obtained by using an interval width of 0.25. The improved "smooth" histogram is shown in Figure 2.2. In general we recommend that you construct your own histogram rather than rely on the ExpertFit default. There is no definitive prescription for choosing histogram intervals!

> **Note that the histogram interval width can also be changed by using the two buttons with arrowheads at the top of the histogram screen.** The left (right)
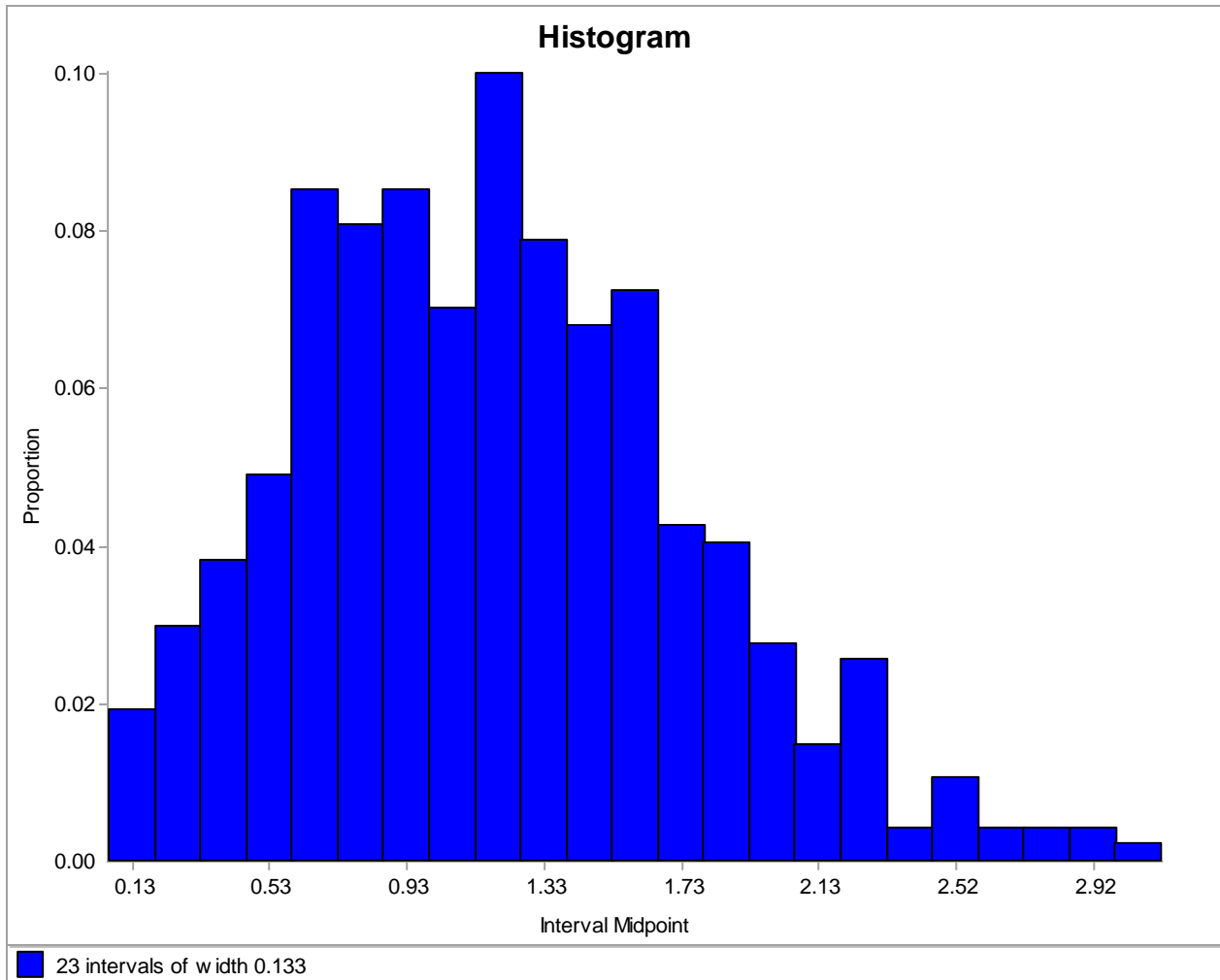
Figure 2.1. Default ExpertFit histogram of the service-time data.
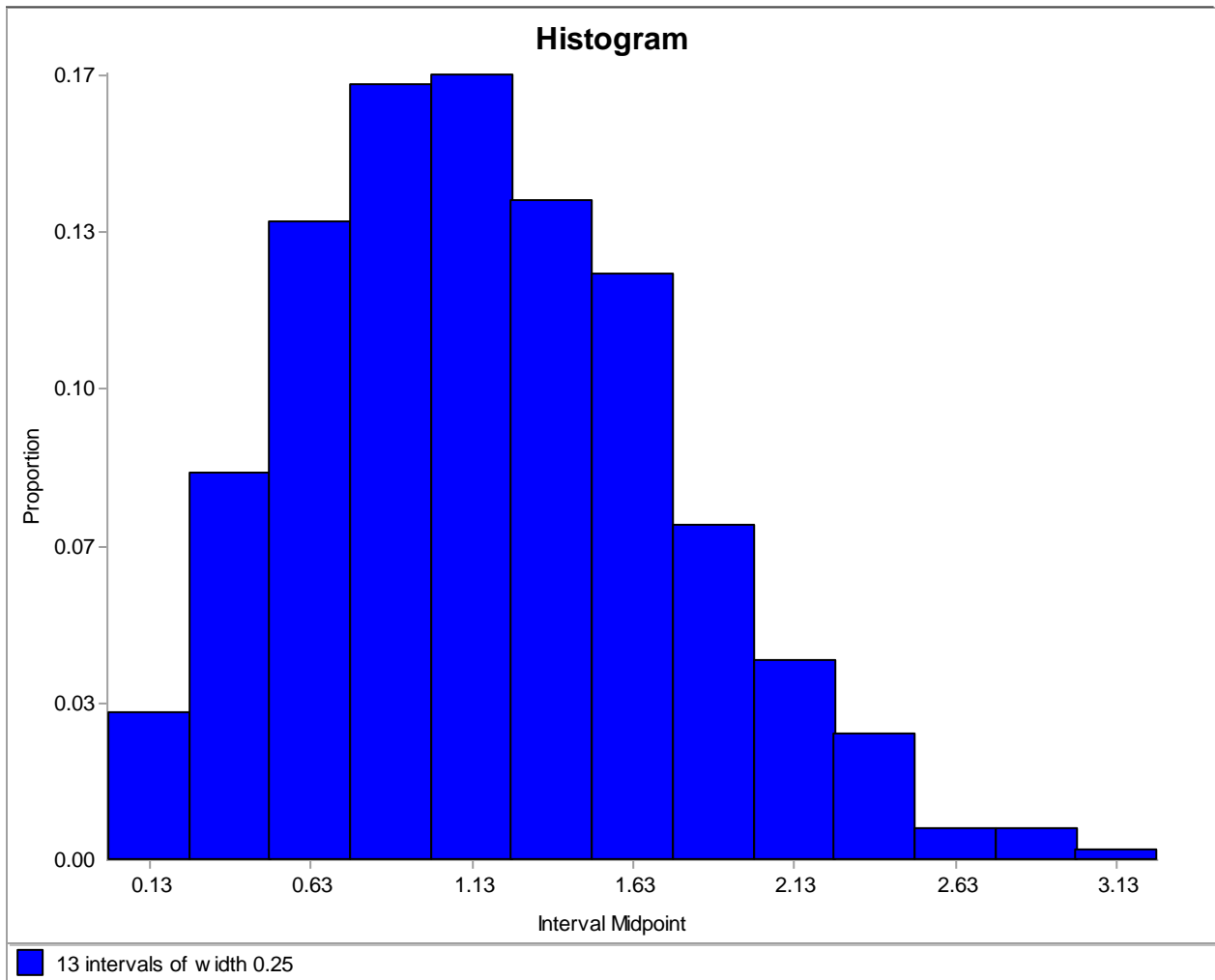
Figure 2.2. Histogram of the service-time data with an interval width of 0.25.

button decreases (increases) the interval width by 5 percent, and can be applied repeatedly.

C: We begin the actual process of finding a distribution that is a good representation for our data by selecting the **Automated Fitting** option at the **Models** tab. Based on certain heuristics, ExpertFit determined that the "best" representation for the data is provided by a Weibull distribution (see Table 2.3) with location, scale, and shape parameters of 0, 1.334, and 2.183, respectively. This best model received a **Relative Score** of 100.00 and its **Absolute Evaluation** message is "Good," indicating no reason for concern. (See the context-dependent online help for a discussion of the terms in boldface.) Furthermore, the model mean and the sample mean are almost identical.

Note that the third-best fitting model is a Rayleigh distribution with an estimated location parameter (denoted by "E") of 0.062. (**If we were to click on View/Delete Models at the Models tab, we would see that the normal distribution was not automatically fit to our non-negative service-time data.** This is because the normal distribution can take on negative values. However, the normal distribution could, if desired, be fit to our data using the **Fit Individual Models** option at the **Models** tab.)

Table 2.3. Evaluation of the candidate models.

Relative Evaluation of Candidate Models

Model	Relative Score	Parameters	
1 - Weibull	100.00	Location Scale Shape	0.00000 1.33367 2.18304
2 - Beta	95.00	Lower endpoint Upper endpoint Shape #1 Shape #2	$6.46210\,e^{-4}$ 3.72935 2.53832 5.47937
3 - Rayleigh(E)	88.75	Location Scale	0.06196 1.25652

21 models are defined with scores between 3.75 and 100.00

Absolute Evaluation of Model 1 - Weibull

Evaluation: Good

Suggestion: Additional evaluations using Comparisons Tab might be informative.
              See Help for more information.

 Additional Information about Model 1 - Weibull

"Error in" the model mean
relative to the sample mean   0.00139 = 0.12%

D: We now do some additional confirmation of the best-fitting Weibull distribution using the **Comparison** tab, as suggested by the latter part of the **Absolute Evaluation** message. The **Density-Histogram Plot** based on the final histogram is shown in Figure 2.3. The closeness of the density function to the histogram visually confirms the quality of the Weibull representation. **Note that we could have simultaneously plotted the density functions of several distributions in Figure 2.3.**
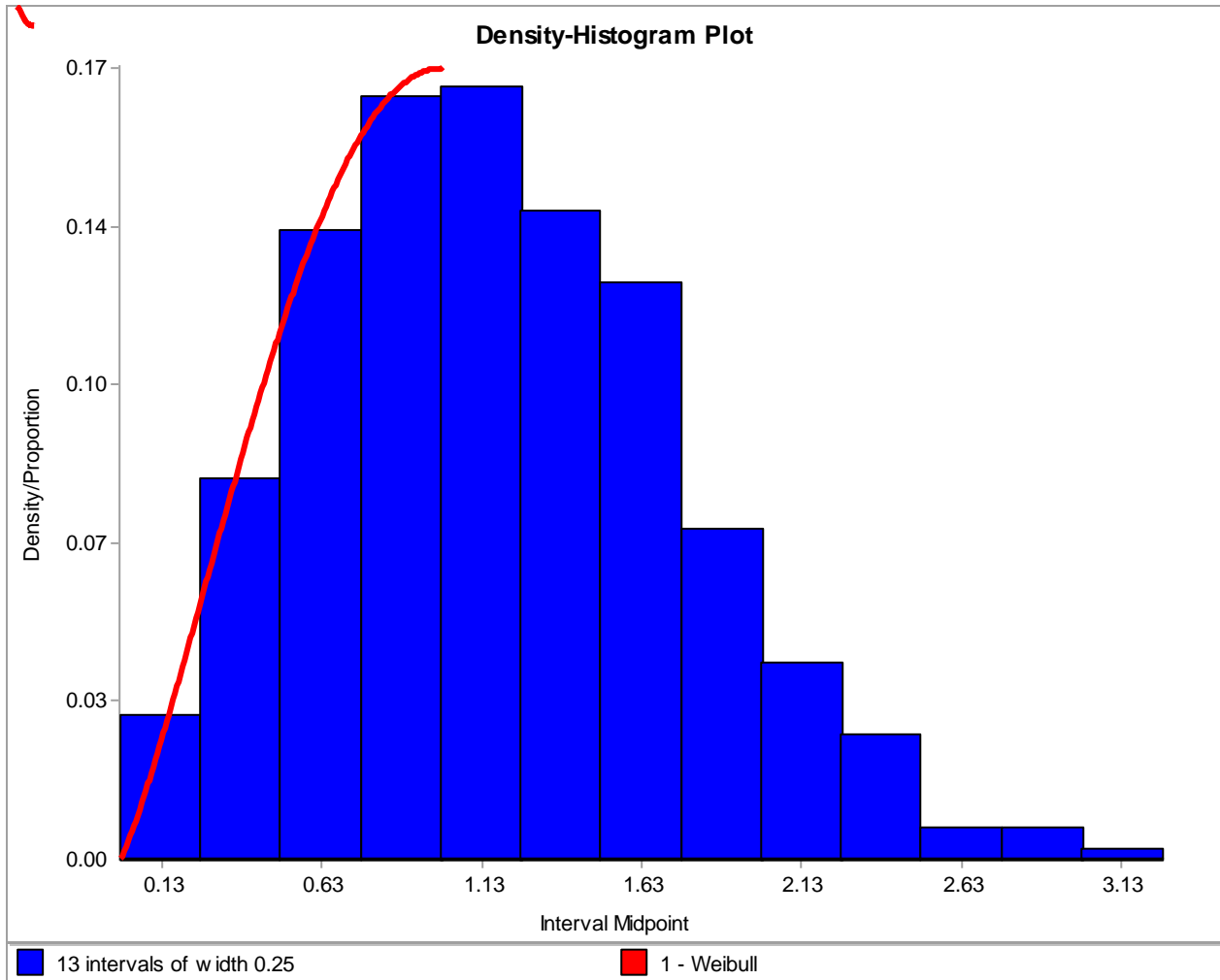
Figure 2.3. Density-Histogram Plot for the fitted Weibull distribution and the service-time data.

E: We present a **Distribution-Function-Differences Plot** for the Weibull distribution in Figure 2.4. The plot shows the differences between the Weibull distribution function and the sample distribution function, over the range of the data. (The sample distribution function, which is an estimate of the true underlying distribution function of the data, is defined at a particular value of x as the proportion of observations in the sample that is less than or equal to x.) Since the vertical differences in the plot are close to 0, this is further indication that the Weibull distribution is a good model for the data.
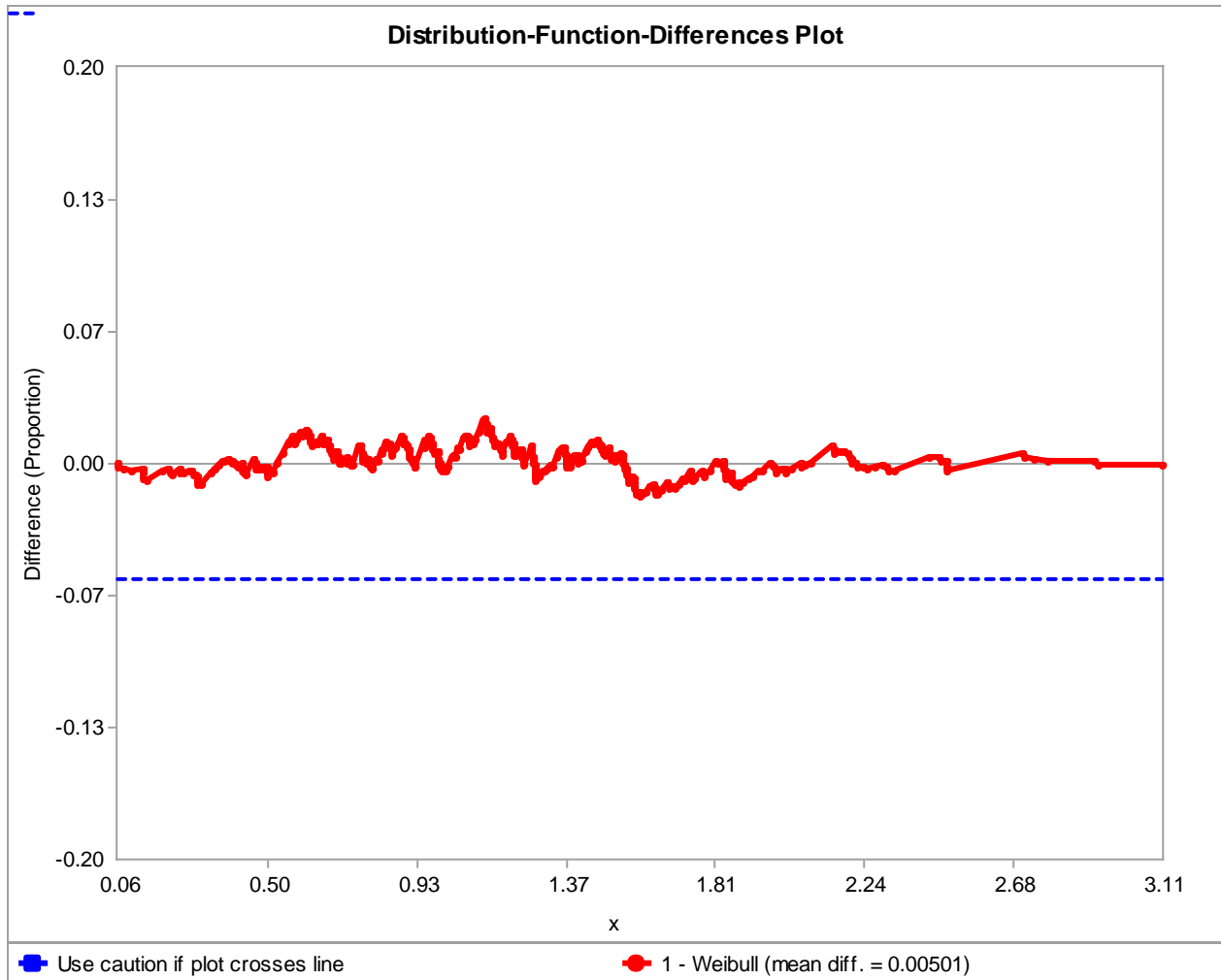
Figure 2.4. Distribution-Function-Differences Plot for the fitted Weibull distribution and the service-time data.

F: We conclude the confirmation process by performing an **Anderson-Darling Test** (the most-powerful test available in ExpertFit) to see formally whether our data could have been generated from the specified Weibull distribution. (You may want to read the discussion of goodness-of-fit tests in the **Goodness-of-Fit Tests and Their Interpretation** tutorial in the software before proceeding.) We will perform the test at a level (alpha) of 0.05. Since the Anderson-Darling statistic, 0.205, is less than critical value, 0.750, we do not reject the Weibull distribution. You should keep in mind that failure to reject by this test does not necessarily mean that the Weibull distribution is exactly the distribution that produced the data; this test tends to have low power for small to moderate sample sizes. (We also performed the **Kolmogorov-Smirnov Test** and **Chi-Square Test** and they did not reject the Weibull distribution.)

In summary, there is no reason to believe based on the above heuristics and tests that the Weibull distribution does not provide a good model for the service-time data.

G: If you are using ExpertFit in the context of simulation modeling, we now see how to put the selected Weibull distribution into the proper format for several different simulation-software products using the **Applications** tab. In particular, the above actions show how to represent the Weibull distribution in the software product of your choice; the actual representations for selected products are shown in Table 2.4.

Table 2.4. Simulation-software representations for the Weibull distribution.

Software Product	Representation
AnyLogic	weibull(2.183039, 1.333669, 0.000000)
Arena	WEIB(1.333669, 2.183039, <stream>)
AutoMod	weibull 2.183039, 1.333669
ExtendSim	Distribution Weibull Scale 1.333669 Shape 2.183039 Location 0.000000
FlexSim	Distribution Weibull Location 0.000000 Scale 1.333669 Shape 2.183039
ProModel	W(2.183039, 1.333669, <stream>)
Simio	Random.Weibull(2.183039, 1.333669, <stream>)
WITNESS	WEIBULL(2.183039, 1.333669, <stream>)

3. Example Data Sets

Data-Set Name	Description
EXAM21_UsersGuide.EXP	Example 2.1 in the User's Guide (service times of customers)
EXAM64.EXP	219 interarrival times of cars to a drive-up bank
EXAM65.EXP	156 demand sizes from an inventory
PROB620a.EXP	91 laboratory processing times
PROB620b.EXP	1000 paper-roll yardages
PROB620c.EXP	218 post-office service times
PROB620d.EXP	1592 times at an ATM machine
PROB620e.EXP	694 machine-repair times
PROB620f.EXP	3035 post-anesthesia recovery times
PROB621a.EXP	369 university test scores
PROB621b.EXP	200 deaths per corps per year by horse kick in the Prussian army in the late 1800s
PROB622a.EXP	122 machine-repair times
PROB622b.EXP	155 machine-repair times