Some Uses for Distribution-Fitting Software in Teaching Statistics

Alan MADGETT

Statistics courses now make extensive use of menu-driven, interactive computer software. This article presents some insight as to how a new class of PC-based statistical software, called "distribution-fitting" software, can be used in teaching various courses in statistics.

KEY WORDS: Exploratory data analysis; Goodness-of-fit; Parameter estimation; Probability distributions; Stochastic modeling.

1. INTRODUCTION

Over the past few years, approaches to the teaching of statistics have changed dramatically. It was not long ago that students performed most numerical computations by hand or on electronic calculators and the mention of "normal scores" did not appear in many textbooks. Today, we use microcomputer software packages, such as Minitab, Systat, Mystat, and SPSS, to instantly create dot plots, stem-and-leaf displays, box plots, and histograms, for interval estimation and hypothesis testing of means, and to provide us with insight in regression or correlation problems. Modern textbooks now routinely describe Normal probability plots, with or without computer software, before proceeding to use a t or F distribution.

It is only recently that a new class of PC-based statistical software, called "distribution-fitting" software, has become available. BestFit, C-FIT, and ExpertFit are three window-based packages currently available. UniFit II is the DOS-based predecessor to ExpertFit (see Sec. 10 for more information about these packages).

In some situations, distribution-fitting software can be used in place of, or to reinforce, other statistical software. On the other hand, there are many applications in which it is the *only* computer solution available. For many students, particularly those in the applied sciences (engineering, geology, biology, life or health sciences) and mathematical statistics, familiarity with distribution-fitting software should be as important a component of their statistics education as Minitab, SPSS, and so on.

Section 2 gives a brief overview of some of the main features that might be available in a distribution-fitting package. Sections 3–7 provide some insight as to how this software can be used in various parts of statistics curricula. The ideas discussed in Sections 3 and 4 can be implemented in *any* introductory statistics course—social science, applied science, or mathematical statistics. The applications in Sections 5 and 6 are more appropriate for courses directed at students in the applied sciences or mathematical statistics. Section 7 is primarily of interest to students in mathematical statistics. Section 8 deals briefly with some software available for analyzing "mixed population" data sets.

The uses suggested in this article for this type of software are not intended to be exhaustive.

2. AN OVERVIEW OF DISTRIBUTION-FITTING SOFTWARE

Distribution-fitting packages usually include at least 25 discrete or continuous probability distributions. Most distributions can be characterized by three parameters: shape, scale, and location. Some software products will include probability laws twice—once with location parameter zero and a second time where it is estimated from the data. A good package should be able to *simultaneously* fit several probability laws to a data set and to rank them as to *relative* suitability. A preliminary indication as to which distributions seem to provide satisfactory models for the data would be helpful.

Access to the estimated parameter values for each of the fitted distributions is essential. Provision of confidence intervals for the unknown parameters might be useful to some people.

In order to investigate further the adequacy of a particular distribution as a model for a data set, one should be able to plot the probability or density function as an overplot on an empirical histogram, to produce probability plots, and to perform various goodness-of-fit tests.

Some software is capable of generating, and storing in a file, samples of data from any one of the probability laws using either specified or estimated parameter values.

For those using the software to determine an appropriate stochastic model for inclusion in a discrete-event simulation, it might be helpful if the appropriate algorithmic code could be produced for use in major simulation languages. It should be noted that distribution-fitting software is now being built into, or bundled with, some simulation languages.

Some products provide modeling tools for situations where very little or no data is available. It attempts to provide stochastic models on the basis of subjective information such as smallest, largest, most likely (modal), and average values.

Most standard statistical software packages do not include routines to perform the tasks described. Experimenters and students would find the numerical computa-

Alan Madgett is Associate Professor, Department of Mathematics and Computer Science, Laurentian University, Sudbury, Ontario, Canada P3E 2C6 (E-mail: amadgett@bethel.cs.laurentian.ca). The author thanks an associate editor and the referees for their many suggestions and, above all, for their encouragement to pursue what turned out to be a major revision.

tion associated with most of them to be overwhelming if they tried to do them by hand. Fortunately, easy-to-use distribution-fitting software has put solutions to these problems just a few keystrokes away.

Readers are reminded that some distribution-fitting products are more comprehensive than others. Consequently, not all of the features mentioned above will be available in every package.

3. EXPLORATORY DATA ANALYSIS

Many introductory courses in statistics start with a section on exploratory data analysis. In this way, students become aware that real-life data can behave differently depending upon the situation under study. It is hoped that students come away with the realization that experimental data is the result of some stochastic process. To create a frequency table or histogram and then drop the matter after just pointing out that the data exhibit symmetry or a particular skewness seems rather incomplete. If we can introduce students to computer software that will create these tabular summaries, the least we can do is go one step further and demonstrate to them that other software exists that is capable of determining suitable probability models for the data.

The learning curve for this type of software compares favorably with that associated with data analysis tasks in packages such as Minitab. Students can be introduced to the "data summary" and "model selection" parts of the software in 20 to 30 minutes using a computer demonstration. I prefer to give the demonstration during a tutorial or lab period in a room where some PCs are available. In this setting, students can work along simultaneously with the instructor or follow up the presentation with some "hands on" experience under the guidance of the instructor or a teaching assistant. The inital reaction of the students is "amazement" to both the power of the software and the ease with which it can be applied. They are quite eager to try it on their own data set. To take advantage of this enthusiasm, I immediately hand out an assignment (to be submitted in a few days) in which I carefully lay out step-by-step the analysis to be performed on a data set. Each student creates his/her own (supposedly) unique data set by applying a mathematical transformation, based on their social security number, student number, or numeric day of birth, to a given set of data.

Some may argue that, at this stage of their statistics education, students do not know the names of any probability distributions, other than possibly the Normal (by reputation). Although this may be true, at the very least, students should be aware that statisticians have formulated many probability laws that are useful in describing various situations. Being able to show them a graph of what these laws look like, preferably superimposed on an empirical distribution, can be a valuable experience. Some disciplines seem to only live in a "Normal" world but this is definitely not true in the applied sciences.

Conclusions drawn from a frequency table or histogram can often be unreliable or even misleading. The "shape" exhibited by these summaries can be influenced considerably by the choice of class width or boundaries. Some data sets that illustrate this type of behavior can be found in Devore (1995); Johnson (1994); Johnson and Bhattacharyya (1996); Khazanie (1996); Mendenhall and Beaver (1994); and Samuels (1989). Because distribution-fitting software fits the raw data and not the histogram, use of this software can help prevent one from choosing an inappropriate model for their data set. One might argue that it is good practice to always do *several* frequency tables but, not only might this be inconclusive, it seems to be a waste of effort when an application of distribution-fitting software will give the most suitable probability laws the *first time*.

4. INVESTIGATING THE "NORMALITY" ASSUMPTION

Statistical procedures using t, chi-square, or F distributions require the assumption that the data come from Normal populations. In fact, the chi-square and F distributions are quite sensitive to departures from Normality.

Textbooks that simply quote the number of observations, the mean, and the standard deviation for a sample are becoming more rare. Authors are now including the raw data and stressing that, before using any of these three distributions, one should use Normal probability plots to look for any evidence of outliers or serious non-Normal behavior. These probability plots are *visual* techniques.

Distribution-fitting software is an alternative approach which uses various *quantitative* measures (based on estimation and goodness-of-fit) to rank the suitability of the Normal law among other distributions available and to provide a general indication as to whether a Normal law seems to provide a satisfactory model for the data. If there is an indication that the Normal law is probably not appropriate, then a list of more suitable models has already been provided by the software. In such cases, it is common practice to consider applying a mathematical transformation to the data in the hope that the transformed set might exhibit a more Normal-like behavior. This list of "more suitable" models might suggest an appropriate transformation. Otherwise, one can try the standard transformations. Once a transformation has been applied to the data set, the new data set can be run through the software to see if a Normal law is now satisfactory.

To obtain further information on the adequacy of the Normal, or any other probability law, for describing a set of data, one can apply one of the goodness-of-fit tests available in the software. These are discussed in Section 6.

5. PROBABILITY PROBLEMS

In statistics textbooks intended for students in mathematics or engineering, we encounter probability exercises of the form: "assuming a Weibull probability law with $\alpha = 2$ and $\beta = 3$, determine the probability that . . . ." Students wonder why this particular distribution was chosen and from where did these "nice" values for the parameters come. Would it not be much more instructive to give the students a small set of data, ask them to fit a particular probability law to it us-

ing this software, and then, on the basis of the fitted model, "determine the probability that ... "? Obtaining estimates for parameters in some of the distributions may entail rather complex computation if done by hand. Of course, this is no obstacle when distribution-fitting software is used.

There is no longer any justification for avoiding probability laws, such as the Gamma, on the basis that they require numerical integration to compute probabilities. On most campuses, students have access to a "mathematical" computer package (such as Maple, Mathematica, or Matlab) or have their own programmable calculator. Adding distribution-fitting software to these tools seems like a natural step forward.

6. STOCHASTIC MODELING

For students in the applied sciences, stochastic modeling is an important part of their education. It is usually included in their first course in statistics disguised as "goodness-of-fit" tests. For many students, it may later become an important analysis tool for their laboratory work or thesis.

From time to time, statisticians are visited by students seeking advice on the analysis of their experimental data. Often, they have been told by someone to use a particular distribution or test but they now have reservations about proceeding in that direction. In many instances, just running their data set through a distribution-fitting package and doing a probability plot using the proposed distribution will allay their fears. Others may want to apply a goodness-of-fit test to the data but they remember only too well this procedure from their introductory statistics course—lots of numerical computation! With distribution-fitting software, goodness-of-fit tests can be applied quickly, and painlessly, for any one of the 25 or more distributions. Students who are taught how to do this in their statistics course will have acquired a valuable tool for future data analysis.

The Anderson–Darling test, the Kolmogorov–Smirnov test, and a chi-square test are available in BestFit, ExpertFit, and UniFit. C-FIT only includes the latter two. In ExpertFit and UniFit, one can apply these tests using *prescribed* (hypothesized) values for the parameters as well as values estimated from the data.

Because the chi-square goodness-of-fit test is the test that is most frequently taught in a first course in statistics, the flexibility available in ExpertFit and UniFit for applying this test will be appreciated by most instructors. Two chi-square tests are provided. One uses intervals, or classes, of "equal width" (the method presented in most textbooks); the other uses intervals having "equal probability." In both cases, the user can determine the number of intervals to be used. If the intervals of "equal width" method is chosen, there is an easy-to-use routine that permits one to group together adjacent classes having very small expected frequencies and, if necessary, to undo this grouping later. The intervals of "equal probability" method permits the user to specify the expected frequency desired in each class thereby avoiding the necessity of grouping some adjacent classes. This latter method is seldom included in textbooks because of the intricate computation required in some situations.

With the assistance of ExpertFit or UniFit, students seem to exhibit a much more positive disposition towards carrying out chi-square goodness-of-fit tests. They are no longer overwhelmed by the many computations and are not restricted to simple models such as the uniform, Poisson, binomial, and Normal. Furthermore, if there is any doubt as to the outcome of the test, they can quickly repeat the analysis using different classes or data groupings.

I often use this software to impress upon students the need to be realistic when applying goodness-of-fit tests to a data set. To illustrate this, we take a fairly large set of data, say 300 to 500 values, for which a histogram exhibits some "generally recognizable" stochastic pattern. When we perform a chi-square goodness-of-fit test with 30 classes, many times it turns out that all of the probability laws provided by the software are rejected as suitable models. Although it is true that "more data gives more information", it may be unrealistic to expect a data set to exhibit good agreement with a model over this many classes!

7. ESTIMATION OF DISTRIBUTION PARAMETERS

The backbone of distribution-fitting software is the estimation of the unknown parameters that define the various probability laws. We cannot create visual plots of density functions or carry out goodness-of-fit tests without these estimates. They are also essential if one wishes to determine a particular probability model for use in a simulation. In some practical situations, the parameters defining a distribution might have some special physical significance. In that case, the distribution of the estimators may also be required.

For students in mathematical statistics programs, the theory associated with deriving formulas for estimators of parameters is part of the "principles of estimation" topic in the upper year curriculum. In practice, these principles are illustrated using only a few probability laws, say three or four. Distribution-fitting software implements this theory for a large number of distributions and, when applied to appropriate data sets, provides a *practical* component to the study of this topic.

Students in the applied sciences are not usually interested in the mathematical derivation of the estimators. Some do, however, have need for software that will give them numeric estimates, let them set up interval estimates and permit them to test hypotheses for parameters in a variety of probability laws.

In most distribution-fitting software packages, the method of maximum likelihood is the *primary* technique used for estimating the parameters of a distribution. For some distributions, "nice" mathematical expressions are available for these estimators. For others, such as the gamma and Weibull laws, numerical (iterative) techniques are required. In either case, distribution-fitting software handles the numerical computations quite speedily. If an instructor includes the Levenberg–Marquardt algorithm for

improving the maximum likelihood fit, the students can choose this option in BestFit.

C-FIT is unique in that it provides students with three procedures for estimating the distribution parameters—maximum likelihood, method of moments, and a least squares procedure based on regression fits for probability plots. One can select all three and compare the results visually (with overplots on a histogram) or statistically (with significance levels from goodness-of-fit tests).

In ExpertFit and UniFit, the theory for the distribution of maximum likelihood estimators is used to produce confidence intervals for the unknown parameters (at a level specified by the user). If the exact distribution of the estimator is available, it is employed. In most other situations, this software uses the "asymptotically Normal" property of the maximum likelihood estimator. Whenever an interval is displayed, it is annotated with the technique used. For those who need it, an estimate of the asymptotic variance-covariance matrix for the parameters is also outputted.

8. MODELING MIXED DISTRIBUTIONS

If an application of the distribution-fitting software presented here indicates difficulty in finding a distribution that provides a good model, it may be because the data set comes from a "mixed population." In this case, an application of the MIX software package should be considered (see Sec. 10 for more information on MIX).

The MIX software package fits grouped data (a histogram) using a mixture of statistical distributions. It determines the set of *overlapping* component distributions that gives the best fit to the histogram under the conditions or constraints imposed by the user. The process can be reapplied using different data groupings until one is satisfied with the fit. The component densities can be Normal, lognormal, gamma, exponential, or Weibull distributions.

9. CONCLUDING REMARKS

Distribution-fitting software provides an *alternative* tool for exploratory data analysis and for checking for Normality. The strengths of other software products—such as Minitab, Mystat, and Systat—lie in performing standard statistical tests and regression analysis. The forte of distribution-fitting packages is in stochastic modeling and its associated tasks.

With distribution-fitting software, instructors can expand considerably the scope of their course at very little cost in terms of additional time. Students are able to model a wide variety of probability distributions, estimate their unknown parameters, and conduct goodness-of-fit tests—tasks which would be "computationally challenging" without this software.

The increased use of menu-driven, interactive software has been recognized as a way of improving the delivery of statistics courses for engineers (Hogg 1994). The same can be said for students in other applied areas of science. Exposure to distribution-fitting software in a mathematical statistics program should result in a more versatile and better trained statistician.

10. DISTRIBUTION-FITTING SOFTWARE SOURCES

a.	BestFit:	Palisade Corporation, 31 Decker Road, Newfield, NY 14867
b.	C-FIT:	C-FER Technologies, Inc. 200 Karl Clark Road, Edmonton, Alberta, Canada T6N 1H2
c.	ExpertFit, UniFit II:	Averill M. Law & Associates, P.O. Box 40996, Tucson, AZ 85717
d.	MIX:	Ichthus Data Systems, 59 Arkell Street, Hamilton, Ontario, Canada, L8S 1N6

[Received October 1995. Revised September 1996.]

REFERENCES

Devore, J.L. (1995), *Probability and Statistics for Engineering and the Sciences* (4th ed.), Belmont, CA: Duxbury Press, p. 12.

Hogg, R.V. (1994), "A Core in Statistics for Engineering Students," *The American Statistician*, 48, 285–288.

Johnson, R.A. (1994), *Miller and Freund's Probability and Statistics for Engineers* (5th ed.), Englewood Cliffs, NJ: Prentice-Hall, pp. 14 and 157.

Johnson, R.A., and Bhattacharyya, G.K (1996), *Statistics: Principles and Methods* (3rd ed.), New York, NY: Wiley, p. 274.

Khazanie, R. (1996), *Statistics in a World of Applications* (4th ed.), New York: HarperCollins, p. 22.

Mendenhall, W., and Beaver, R.J. (1994), *Introduction to Probability and Statistics* (9th ed.), Belmont, CA: Duxbury Press, p. 37–39.

Samuels, M.L. (1989), *Statistics for the Life Sciences*, San Francisco, CA: Dellen Publishing, pp. 34–35, 54.