CLASSROOM GOODNESS OF FIT TESTING
USING MICROSOFT WINDOWS

Donald L. Byrkett

Systems Analysis Department
Miami University
Oxford, Ohio 45056

ABSTRACT

The chi-square goodness of fit test and other statistical tests for comparing observed data with a theoretical distribution are not well supported by the popular statistical analysis packages such as SAS, SPSS, BMDP, and MINITAB. To overcome this limitation, the author has written a PC based program to support goodness of fit testing and has used it in his simulation class. The original version of this program was written as a typical, menu oriented, DOS application using the language True BASIC. Recently, under a grant from IBM, this program was rewritten in the language C to run under the Microsoft Windows 3.0 operating environment. This paper discusses the importance of goodness of fit testing in simulation and other stochastic modelling classes, describes the Windows 3.0 version of this program and its limitations, and discusses how this program can be used in a simulation class. This program is available without cost to Winter Simulation Conference participants who wish to use it in their simulation classes.

1 INTRODUCTION

When teaching courses in stochastic modelling, theoretical distributions are often used to model various system activities. For example, in studying queueing systems the Exponential distribution may be used to model service time or in reliability systems the Weibull distribution may be used to model a component lifetime. Sometimes these distributions are chosen for mathematical tractability but, in principle, they are chosen to represent the activity being modelled. Unless students are asked to collect data on a particular activity and to use that data to select a theoretical distribution to model the activity, they have difficulty understanding how the distribution represents the activity. I believe that this is such an important point that it should be covered in every stochastic modelling course. This requires a discussion of data collection, coverage of goodness of fit testing, and software that allows the student to quickly try alternate distributions to model a particular activity. This paper concerns the software support.

The major statistical analysis packages (SAS, SPSS, BMDP, and MINITAB) provide very limited goodness of fit procedures. At most they provide support for two distributions, the Exponential and Normal. I am aware of only two software packages that provide the kind of capabilities needed. One is a FORTRAN program listed in Phillips (1972). Though I have used and modified this program, it contains a number of errors and is not well suited to an interactive environment. The other is UniFit (Simulation Modeling and Analysis Company, P.O. Box 40996, Tucson, AR 85717), a commercially available PC based software package for distribution fitting. Though I have not used UniFit, I expect that it is an excellent package. Unfortunately, the cost of obtaining sufficient copies prohibits its use for classroom instruction.

MenuGof and Gof are programs that I created to provide a user friendly environment for students to try fitting alternate distributions to a set of data. MenuGof is a traditional, menu oriented, DOS based program described in Byrkett (1989). This paper describes the operation and use of Gof, a Microsoft Windows 3.0 version of the same program. This program was developed under a grant from IBM to develop Windows based software for educational use.

2 PROGRAM OVERVIEW

Gof is coded in the language C using the Microsoft C 6.0 Compiler and the Microsoft Windows 3.0 Software Development Kit. The layout of the screen is illustrated in Figure 1. The program is designed to follow the standard Windows Application Style Guide (1987). There is a single main window with a menu

```
┌──────────────────────────────────────────────────────────────────────────────┐
│  ─                            Gof - PROB19.DAT                           ▼ ▲   │
├────────────────────────────────────────────────────────────────────────────────┤
│  File   Modify   Distribution   Parameters   Perform test   View      F1=Help │
│ ┌──────────────────────────────────────────┐ ┌──────┬──────────┐               │
│ │Current hypothesis:                        │ │─ Data│─ Statistics│            │
│ │H0: The data appear to be IID random variables │ 72.00 │Mean =    99.22 │      │
│ │     from the Normal distribution with     │ │74.40 │Std dev = 10.17 │        │
│ │     Shift = 0.00                           │ │80.70 │Min =     72.00 │        │
│ │     Mean = 99.22                           │ │84.80 │Max =    122.70 │        │
│ │     Std dev = 10.07                        │ │86.00 │No Obs =    50  │        │
│ │H1: The data are not distributed as above  │ │87.50 │                │        │
│ │                                           │ │88.30 │                │        │
│ │Chi-square test results [df = 5]:          │ │88.50 │                │        │
│ │     Test statistic:    1.89               │ │90.70 │                │        │
│ │     Critical value:   11.10               │ │90.90 │                │        │
```

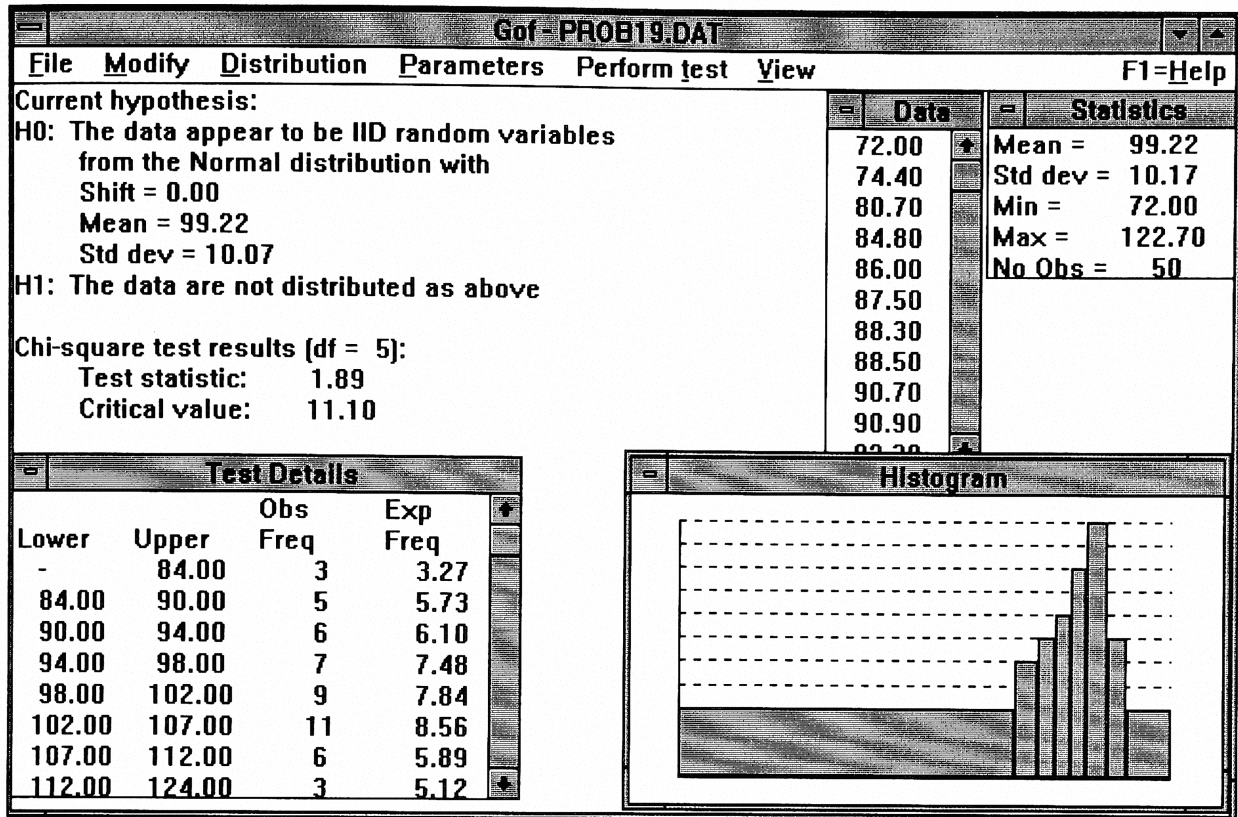| Lower	Upper	Obs Freq	Exp Freq
-	84.00	3	3.27
84.00	90.00	5	5.73
90.00	94.00	6	6.10
94.00	98.00	7	7.48
98.00	102.00	9	7.84
102.00	107.00	11	8.56
107.00	112.00	6	5.89
112.00	124.00	3	5.12

Figure 1: Gof Screen Layout

bar (line 1) indicating the data file being analyzed (PROB19.DAT), a set of six pull-down menus (line 2) plus the Help menu, and a display containing the current hypothesis and the results of any goodness of fit test performed. Other windows may be opened or closed to display a histogram, sample statistics, the observations, the cell frequencies, or the test details. The illustration displays the main window and four display windows. Windows may be opened, closed, resized, moved, scrolled and are accessible using either a keyboard or mouse. The observations to be analyzed are stored in a standard text (ASCII) file which may be created outside Gof using any text editor or within the Gof program using a built in editor. The set of data is read from this file and analyzed by Gof.

Figure 2 illustrates the menu options provided under each of the pull-down menus. Gof is used by making selections from these menus. These menus are designed following standard Microsoft Windows practices. Selections are made by pointing with a mouse or by pressing the ALT key in combinations with the underlined character in the menu. Certain menu options may be selected directly by pressing the indicated functions keys (for example F2 selects the Histogram option directly). Menu options followed by

three periods (...) indicate that when this option is selected a dialog box will appear to obtain additional information.

The File menu is almost standard on all Windows based applications. It provides options for creating a new data file or for reading an existing data file (Open), for printing the test\Xresults (Print), ending the program (Exit), and learning about the Gof program (About Gof). The remaining pull-down menus are unique to the Gof application.

The Modify menu provides options for editing the observation data (Data) and/or the cell bounds (Cells) used for the chi-square test and the histogram. The Modify option Data allows the user to edit the current file of observations. To create a new data file of observations, the user selects the File option Open to name the file and is then automatically placed in the editor to enter the data. Once a data file is opened or created, Gof automatically forms cells and creates a frequency distribution. These automatic cell bounds may be modified as desired using the Modify option Cells.

The Distribution menu is used to select the desired theoretical distribution to be tested. Selecting an option from this menu places a check beside the desired
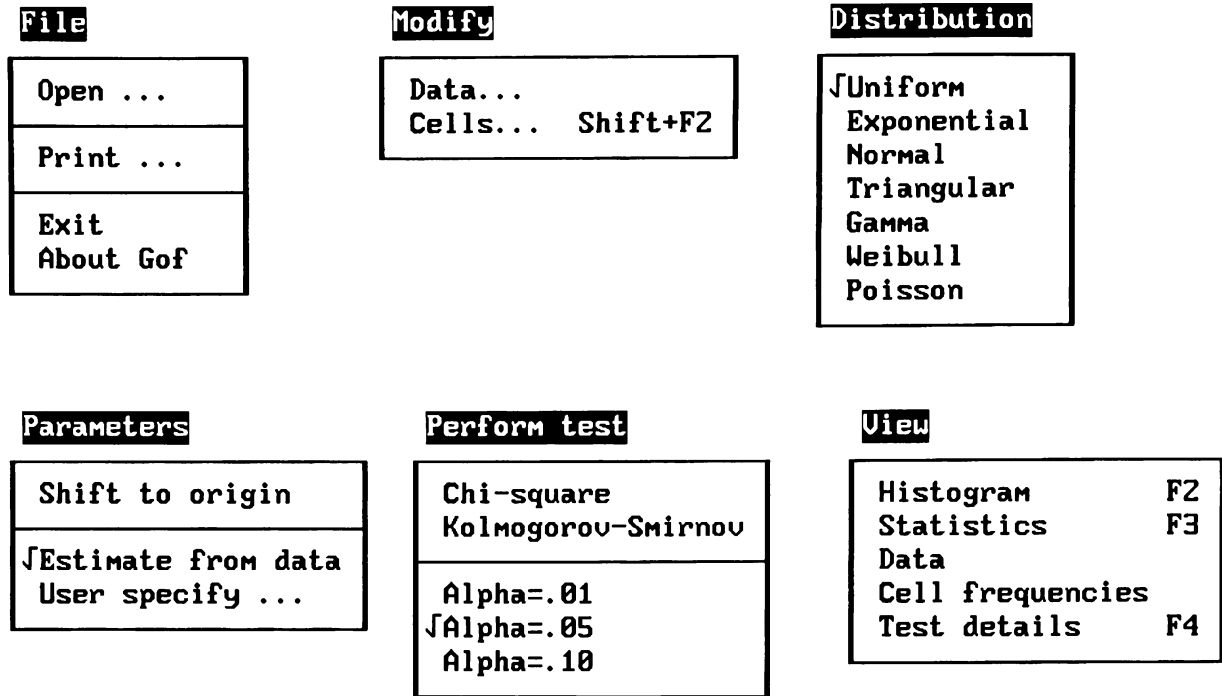
```
█File█
┌─────────────────┐
│ Open ...        │
├─────────────────┤
│ Print ...       │
├─────────────────┤
│ Exit            │
│ About Gof       │
└─────────────────┘
```

```
█Modify█
┌──────────────────────────┐
│ Data...                  │
│ Cells...   Shift+F2      │
└──────────────────────────┘
```

```
█Distribution█
┌──────────────────┐
│ √Uniform         │
│ Exponential      │
│ Normal           │
│ Triangular       │
│ Gamma            │
│ Weibull          │
│ Poisson          │
└──────────────────┘
```

```
█Parameters█
┌──────────────────────┐
│ Shift to origin      │
├──────────────────────┤
│ √Estimate from data  │
│ User specify ...     │
└──────────────────────┘
```

```
█Perform test█
┌──────────────────────────┐
│ Chi-square               │
│ Kolmogorov-Smirnov       │
├──────────────────────────┤
│ Alpha=.01                │
│ √Alpha=.05               │
│ Alpha=.10                │
└──────────────────────────┘
```

```
█View█
┌──────────────────────────┐
│ Histogram          F2    │
│ Statistics         F3    │
│ Data                     │
│ Cell frequencies         │
│ Test details       F4    │
└──────────────────────────┘
```

Figure 2: Gof Menus

distribution, updates the current hypothesis displayed on the main menu, erases any previous test results, and uses standard procedures to estimate the parameters of the desired distribution. The Uniform selection is made by default when the program is initiated. The Poisson distribution is only available if the sample mean is less than 20.

The Parameters option Shift to origin is used to shift the data observations to the origin by subtracting the minimum observation from each observation. This is useful for testing theoretical distributions that are fixed to the origin such as the Exponential, Gamma, Weibull, and Poisson. Normally, the parameters of the distribution are estimated from the data and the Parameters menu option Estimate from data is checked. However, the user can specify the parameters to use by selecting the Parameters option User specify. After selecting this option, it will be checked. When any Parameters option is selected, the current hypothesis is updated and the previous test results are erased.

The Perform test menu signals the execution of either the Chi-square test or the Kolmogorov-Smirnov test on the current hypothesis. A brief summary of the test results are displayed in the main window and remain there until a menu selection is made that will invalidate those results. The Perform test menu also allows the user to specify the Alpha level (type I error) to be used in selecting the critical region for the desired

test. A check is placed by the currently selected level. Alpha = .05 is the default selection.

Finally, the View menu provides five options for opening additional windows to display detailed information. Option Histogram displays a histogram of the current data set using the current cell boundaries (see Figure 1). Option Statistics displays the mean, standard deviation, minimum observation, maximum observation, and number of observations in the current data set (see Figure 1). Option Data displays a scrollable window containing a sorted list of observations (see Figure 1). Option Cell Frequencies displays a table of cell frequencies (not shown in Figure 1). Option Test details provides complete cell details of the chi-square test (see Figure 1) and observation details of the Kolmogorov-Smirnov test.

A typical use of Gof would proceed as follows:

1. Read a data file of observations (File)
2. View histogram to form hypothesis (View)
3. Modify cell bounds to improve histogram (Modify)
4. Repeat step 2.
5. Select distribution (Distribution)
6. Specify parameters or shift data (Parameters)
7. Select alpha level (Perform test)
8. Select test to perform (Perform test)
9. View details of test (View)
10. Repeat steps 5-9 with other distributions
11. Print results (File)

3 STATISTICAL DETAILS

Generally, the statistical background for this program is based on the material in Chapter 6 of Law and Kelton's (1991) simulation text. Exceptions to this are noted.

3.1 Distribution Selection

Gof is capable of comparing a set of observations to seven theoretical distributions as listed in the Distribution menu options. For computational simplicity, the Gamma distribution is limited to the case in which the shape parameter, alpha, is integer. The program is written in a modular fashion so that additional distributions can be added if desired though I have found this selection to be adequate for simulation classes.

3.2 Parameter Estimation

When a distribution is selected, Gof automatically calculates parameter estimates of the selected distribution. The program uses the procedures specified in Law and Kelton (1991) to compute the parameter estimates for all distributions except the Gamma distribution and the Triangular distribution.

Parameter estimates of the Triangular and Gamma distribution are calculated using the method of moments. Alpha, in the Gamma distribution, is rounded to the nearest integer.

$$\underline{\text{Triangular}} \qquad \underline{\text{Gamma}}$$
$$\hat{a} = \min\{X_i\} \qquad \widehat{\text{alpha}} = \text{round}(\bar{X}(n)^2/S^2(n))$$
$$\hat{b} = \max\{X_i\} \qquad \hat{\text{beta}} = \bar{X}(n)/\widehat{\text{alpha}}$$
$$\hat{c} = 3*\bar{X}(n) - \hat{a} - \hat{b}$$

3.3 Chi-square Test

When the chi-square test is performed, the expected frequency in each cell of the defined histogram is calculated and compared with the observed frequency via the standard chi-square test statistic. No attempt is made to form cells of equal expected frequency. The user is warned if the expected frequency in any cell is less than 2.

3.4 Kolmogorov-Smirnov Test

When the Kolmogorov-Smirnov test is performed, the expected cumulative distribution value is calculated for each observation and compared with the observed cumulative distribution values on each side of each observation. The largest difference between the observed and expected values is computed as the test statistic.

This test has two limitations. First, it is only appropriate in situations where the parameters of the hypothesized distribution are known. Second, it is not applicable to discrete distributions such as the Poisson. However, the test is performed in all cases by the program. The critical value is approximated as $1.22/\text{sqrt}(n)$, $1.36/\text{sqrt}(n)$, and $1.63/\text{sqrt}(n)$ for the three levels of alpha.

4 PROGRAM LIMITATIONS

Gof has a number of inherent limitations that make it only appropriate as a classroom instruction tool. First, there is a rather limited set of distributions available. There are plenty to choose from for classroom use but a professional tool would provide a greater variety. Second, no computation efficiencies are built into the program. The goal was simply to get a working program available as quickly as possible. Third, the correct interpretation of the test statistics is up to the user. Critical values are provided but a conclusion is not stated. Fourth, the formation of cells for the chi-square test is up to the user. Cells are not formed so that the expected frequency in each cell is the same as suggested by most statisticians. Though the user is warned if the expected frequency in any cell is less that 2, the test is performed using the cell boundaries specified.

Despite these limitations, Gof is invaluable in the classroom as a quick and dirty instructional tool. This is illustrated in the next section.

5 USE OF PROGRAM IN CLASSROOM

When teaching goodness of fit testing, I do not introduce MenuGof or Gof until I have covered how to model stochastic activities and asked the students to perform some goodness of fit tests by hand. Generally, I begin the discussion with the idea that a particular activity may be represented as a constant (deterministic model) or as a random variable (stochastic model). This decision is based on the inherent variability of the activity. To measure this variability, sample data is collected and the sample variance is calculated. If it is relatively large, a stochastic model is appropriate.

Once it is determined that a stochastic model is appropriate, a distribution needs to be specified. Histograms are introduced as a means of generating an empirical distribution of the activity. The pros and cons of empirical modelling are discussed and the advantages of modelling activities with known theoretical distributions is introduced. This leads conveniently into the concept of parameter estimation of theoretical distributions and goodness of fit testing.

Students are introduced to three methods of parameter estimation: the maximum likelihood method, the method of moments, and common sense. All of these methods are illustrated using a variety of theoretical distributions. Maximum likelihood estimates are preferred when they are readily available. Students are also introduced to both the chi-square goodness of fit test and Kolmogorov-Smirnov test. Example data sets are analyzed, by hand, in class and students are assigned data sets to analyze, by hand, for homework. Finally, Gof is introduced as a means of quickly trying several alternate distributions on a data set.

Gof is a particularly useful program in a simulation class. Simulation models are quite flexible and allow the modeller to use almost any distribution to represent stochastic activities. Typically, I assign the students to develop a simulation model of a particular system, but instead of giving them the distributions for each activity, I give them a set of data observations for each activity. It is their responsibility to analyze the data and develop distribution models for each activity using Gof. Then they can employ these distribution models with the estimated parameters in their simulation model.

This type of assignment is only possible with the aid of a program such as Gof. It provides a valuable learning experience. Students learn a lot about modelling and a lot about data analysis from this type of activity. They learn to usefully apply the theoretical distributions they studied in their probability classes. They also learn about the indefiniteness of the chi-square test as they perform the test using a variety of cell intervals. With one set of cell intervals, a particular distribution is rejected and with another, it is not. In many cases, they prefer the definiteness of the Kolmogorov-Smirnov test.

6 PROGRAM AVAILABILITY

Copies of this software can be obtained on Internet using anonymous ftp from **apsrisc.aps.muohio.edu** (134.53.3.230) in the directory /pub/gof. Copy the files **gof.exe** and **gofhelp.hlp**. If you are not able to use ftp, you can write me at the address given on the title page.

REFERENCES

Byrkett, D.L. 1989. Classroom goodness of fit testing. In *Proceedings of the Midwest Decisions Sciences Institute*, eds. R. Luebbe and B. Finch, 16-18. Decision Science Institute, Cincinnati, Ohio.

Law, A.M. & Kelton, W.D. 1991. *Simulation modeling and analysis*, Second Edition. New York: McGraw-Hill.

Microsoft Windows Software Development Kit Application Style Guide. 1987. Microsoft Corporation, One Microsoft Way, Redmond, WA. 98052.

Phillips, D.T. 1972. *Applied goodness of fit testing*. Norcross, GA: Institute of Industrial Engineering Monograph Series, Operations Research Division, Pub. No. 1.

AUTHOR BIOGRAPHY

DONALD L. BYRKETT is a professor in the Systems Analysis Department at Miami University. He teaches in the areas of stochastic modeling, simulation, manufacturing systems, and microcomputer systems. He is currently writing an introductory book on the Windows environment.