



OBJECT ORIENTED SOFTWARE FOR AUTOMATIC INFERENCE

Ronald S. Reagan
Department of Computer Science
University of Southern Mississippi

Satya N. Mishra
Department of Mathematics and Statistics
University of South Alabama

ABSTRACT

In this paper we propose a method of distribution recognition using the idea of classical maximum likelihood estimation technique (along with chi-square and Kolmogorov-Smirnov tests). Two examples, one continuous and the other discrete, are given and the corresponding data sets are analyzed using the software RVL.

1 INTRODUCTION

The program product RVL (Random Variable Laboratory) was developed for use in teaching simulation and modeling courses. It is especially useful for course projects that require fitting distributions to independent observations of a random variable.

Object-Oriented Programming (OOP) as presented by Booch(1991) was used in analysis, design, and implementation of RVL. OOP provides convenient facilities for using the software engineering concepts of encapsulation, inheritance, and polymorphism (Jones,1990). Encapsulation allows the packaging of data and functions together in a class and the protection of the data from unauthorized access. Inheritance permits a new class to be derived from a parent class. A child class may use or override functions of the parent and define new data and functions. The classes are not objects, but templates for objects. Use of a class requires instantiation of an object from the class. In polymorphism a single function name is able to operate on each element of a set (container class) of similar but different objects.

The RVL program product is the result of applying OOP to the basic problem of fitting distributions to data. The question "Can a Random Variable be treated as an object?", popped into Reagan's head during a Statistics lecture. The answer to this question has proven to be a resounding "YES!".

Several uses can be envisioned for RVL and its ca-

pability to rapidly fit distributions to data. The previously mentioned use in simulation model building is foremost. In addition to modeling input random variables, RVL may be used to fit a distribution to an output random variable if the replicates are independent. An output distribution could be used as an input distribution to another simulation. RVL is also useful in teaching students (and faculty) about distributions and the kinds of data sets they fit. We learned much about characteristics of distributions during the development of RVL.

The product has been used in Simulation and Modeling and Mathematical Statistics courses at the University of Southern Mississippi and the University of South Alabama. The students enjoyed the search for the "best" distributions and provided valuable feedback about RVL and its user interface.

2 RVL ALGORITHM

RVL enables fitting distributions to data using the maximum likelihood estimation technique. Usually a numerical procedure is required to find the optimum parameter estimates. The random variables then are ranked based on the likelihood values obtained. Although ranking based on likelihood values has been proposed before (Hogg et al., 1972), we are unaware of its use in a distribution fitting code. Since the parameter estimates for each fit maximize the likelihood, the fit which optimizes the maxima is deemed to be an optimum in some sense.

Most numerical procedures in RVL were taken from the Pascal version of Press, et al (1989), including the minimization solver using Powell's method, the Beta function, the Gamma function, and the iterated midpoint integration module. An original algorithm was used to compute integral of the densities near the singularities of the Beta, Gamma, and Weibull densities. A future version of RVL will use the simplex method (Amoeba in Press et al. 1989) as prelimi-

nary testing has shown it to be 3 to 10 times faster than the Powell method. To facilitate extensive testing of RVL, random variable generation is provided. The user chooses a random variable, the number of variates, and the parameter values.

The major steps in the overall algorithm are:

1. If the user wants to generate a random variable then generate the random variable, else read user's datafile. Sort the data and compute the summary statistics.
2. Fit various distributions chosen by the user to data using maximum likelihood. For most distributions the maximization problem is solved numerically.
3. Rank the fits based on values of the likelihood function.
4. Compute other goodness of fit statistics including Kolmogorov-Smirnov and Chi-Square.
5. Plot selected densities against a histogram.
6. Plot selected cumulative distribution functions and the empirical cumulative distribution function.
7. Compute the (population) mean and variance for each distribution fit.
8. Plot P-P and Q-Q probability plots for selected distributions.
9. Print textual and graphical results in a useful form.

After the fitting in step 2 is complete the user is free to view or review the resulting screens in any order.

RVL was analyzed, designed, and implemented using Object-Oriented Programming (OOP). Turbo Pascal 6.0 was used to implement the RVL design.

The class *Sample* is used to hold the data set, its title, and statistics computed from the sample. The functions used to plot the histogram and the empirical cumulative distribution function and all other functions and data associated with the data are encapsulated in the *Sample* class.

Each random variable (distribution) is given a descriptive name e.g., Gamma to which is concatenated the number of unknown parameters. All RV's (Random Variables) are inherited from the base class *Rand* which contains data and functions that are used by all of its children. The classes *Cont* for continuous RV's and *Disc* for discrete RV's are inherited from *Rand*.

All the continuous RV's are inherited from *Cont*, and all the discrete RV's are inherited from the class *Disc*. These relationships are illustrated in the tree shown in Figure 1.

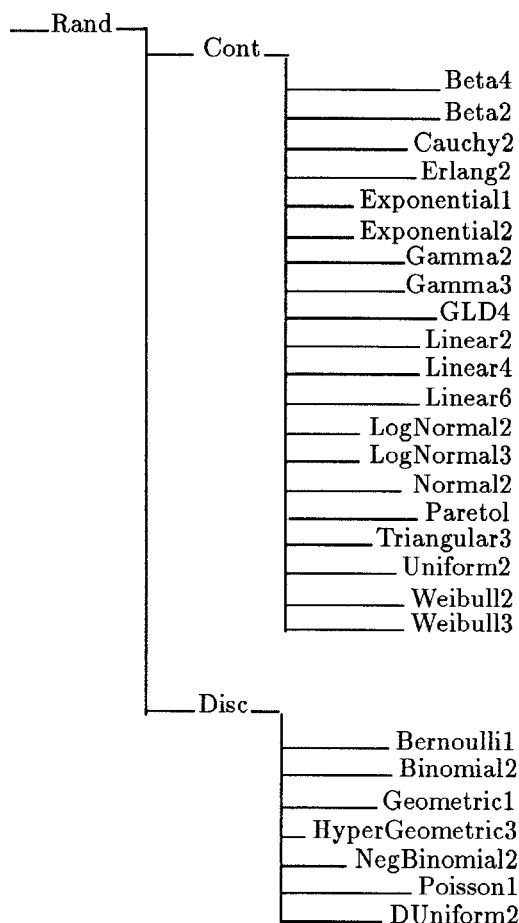


Figure 1: Class Relationship Tree

The definitions and order of the parameters in the RVL output follow those used in several references. The GLD4 distribution is described in Dudewicz and Mishra (1988) and Zaven and Dudewicz (1991). The Cauchy2 is described in Johnson and Kotz (1970). The remaining distributions follow Law and Kelton's (1991) compendium but with the addition of a shifted version of most entries.

The *Linear2*, *Linear4*, and *Linear6* densities are piecewise linear with 2, 4, and 6 independent parameters which are taken as the density values at equally spaced points. For example the *Linear6* has six panels with 7 density values needed to draw the distribution. The rightmost density value is computed from

the other 6 values to force the area under the function to be one. These are included to provide an ordered family of probability models. The Linear6 usually fits better than models with fewer parameters and thus provides a good comparison case for the more common densities. Linear6 is capable of modeling a trimodal density. Linear6 could be used as a practical input random variable. It also would be particularly useful in automated systems with real time updating of the process distribution, such as a traffic control system.

The list of RV's chosen by the user to be fit contains mixed objects since the random variables have different numbers of parameters, different density functions, etc. The concept of polymorphism is used to allow a single function name like FitMLE (MLE = Maximum Likelihood Estimate) to be applied to each RV in the list.

A World Graphics Interface Class (WGI) was used for the plots. This reusable class was originally written in Turbo C++ and later converted to Turbo Pascal. A WGI object is a rectangular region of the screen positioned by device independent coordinates locating the vertices.

A major advantage of the OOP approach is that RVL is extensible. This means that a user can add a new RV to her (or his) version of RVL by writing several short Pascal procedures which specify the basic properties of the random variable. The new RV inherits all the capabilities of RVL for fitting and plotting.

3 EXAMPLES

The results for two examples - one continuous and one discrete - are included. The data sets both came from the exercises from Chapter 6 in Law and Kelton (1991a). The results presented agree with the ones given in the solution manual for the Law and Kelton text (1991b).

3.1 Example 1

The continuous problem apparently arose in quality control and consists of 154 values of errors in the diameters of ball bearings (Law and Kelton 1991a, problem 6.21, p 415). The tabular results from RVL are given in Tables 1-3. Table 1 contains the summary statistics for the data set. The ranked distributions, *fave* (the *n*th root of the likelihood function), and the parameter estimates are listed in Table 2. The expected values for each fit population and the values of the Kolmogorov-Smirnov and ChiSquare statistics are arranged in Table 3. Linear6 provides the best fit

with Linear4 and Normal2 close behind. Other distributions fit well and could be used in a simulation model.

The graphics for the continuous example are plotted in Figures 2-5. The four best densities appear in Figure 2 along with a histogram which has about 5 data points per cell. The numerically integrated cumulative distribution functions are plotted in Figure 3 with the empiric. P-P and Q-Q probability plots are in Figures 4 and 5, resp. The plots graphically confirm the numerical ranking in Table 2. NOTE: The different plots are shown in color on the display when the program is run. However the hardcopy provided herein is black and white. Different line styles are also used to help differentiate the curves.

3.2 Example 2

The discrete example contains 76 values of the number of items demanded per day from an inventory (Law and Kelton 1991a, problem 6.22, page 416). The tabular results are in Tables 4-6. The results agree with Law and Kelton (1991b) with the NegativeBinomial2 winning the contest and Geometric1 a close second.

The graphics for the discrete example are in Figures 6-8. The plots confirm the ranking by likelihood.

4 PITFALLS AND REMARKS

- (i) RVL is a mixed bag of numerical analysis, experience, and software engineering. It contains many iterative procedures which are not guaranteed to converge. Nevertheless RVL's probability of failure is low enough that it is still a very useful tool.
- (ii) All integration within RVL is done numerically. Integration of the density to compute probabilities is the primary operation. Since some of the densities (Beta, Gamma, etc.) may be infinite for some parameter values and others while finite are still badly behaved (LogNormal for example) the integration is a tricky business. However the MLE fitting process does not involve integrals. Thus the integration only provides "window dressing" to be used in visual verification of goodness of fit.
- (iii) Maximum likelihood estimation is an optimal technique for statisticians. A data analyst probes in many directions to get an optimal result for the problem. While exploring the different methods of distribution fitting we decided

Table 1: Sample Statistics for Example 1 — Errors in Diameters of Ball Bearings

	Kind	Size	Mean	Median	Mode		
CONTINUOUS		154	1.230	1.220	1.09		
	Variance	Skewness	Kurtosis	SD	Minimum	Maximum	Range
	0.852	-0.225	3.654	0.923	-1.720	4.010	5.730

Table 2: Distribution Rankings for Example 1

DISTRIBUTION	fave	----- Parameter Estimates -----					
Histogram	0.30575						
Linear6	0.27048	0.046	0.000	0.298	0.406	0.278	0.032
Linear4	0.26514	0.019	0.032	0.501	0.148		
Normal2	0.26296	1.230	0.852				
Weibull3	0.26035	4.442	3.895	-2.337			
Beta4	0.26001	10.466	8.827	-3.528	5.209		
LogNormal3	0.25752	2.296	0.009	-8.722			
GLD4	0.24173	1.145	0.349	4.487	5.807		
Gamma3	0.23945	10.006	0.370	-2.423			
Triangular3	0.23889	-1.806	4.073	1.300			
Linear2	0.23871	0.010	0.340				
Cauchy2	0.22505	0.544	1.265				
Uniform2	0.17452	-1.720	4.010				
Exponential2	0.12469	2.950	-1.720				

Table 3: Expected Values and Statistics for Example 1

RV	E(X)	V(X)	-Ln(L'hood)/n	KS	Chi2
Linear6	1.172	0.928	1.308	0.086	28.734
Linear4	1.386	0.926	1.327	0.082	33.442
Normal2	1.230	0.852	1.336	0.044	62.177
Weibull3	1.214	0.821	1.346	0.047	80.835
Beta4	1.212	0.934	1.347	0.062	58.773
LogNormal3	1.253	0.884	1.357	0.054	154.397
GLD4	1.275	1.387	1.420	0.097	58.512
Gamma3	1.275	1.366	1.429	0.094	180.064
Triangular3	1.189	1.442	1.432	0.129	57.178
Linear2	1.153	1.415	1.433	0.140	57.278
Cauchy2	999.999	999.999	1.491	0.103	57.946
Uniform2	1.145	2.736	1.746	0.245	163.143
Exponential2	1.230	8.705	2.082	0.388	277.667
Sample	1.230	0.852			

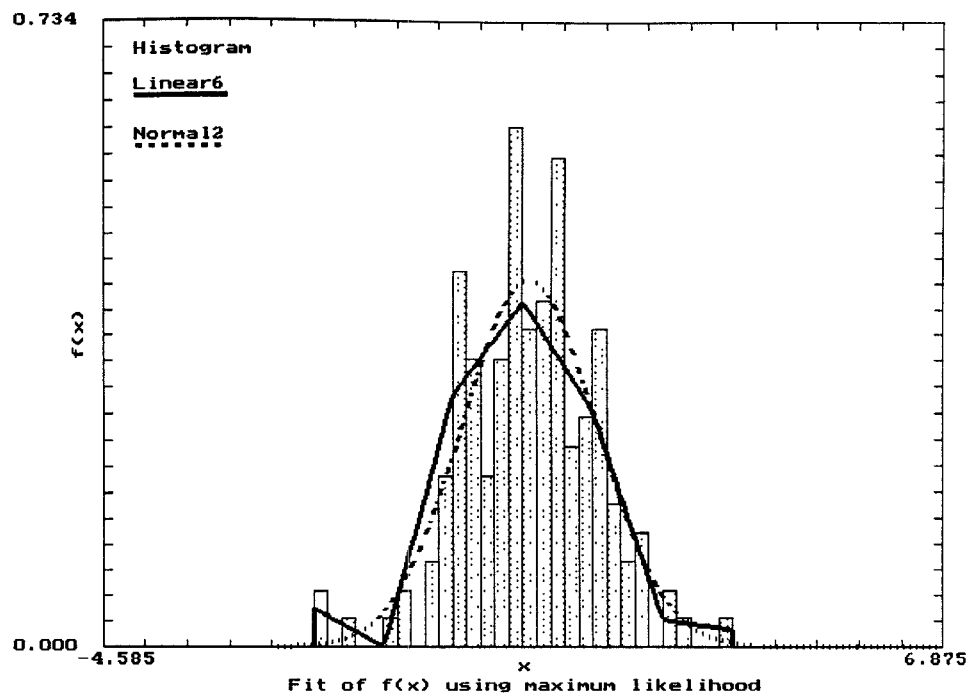


Figure 2: Density Plots for Example 1

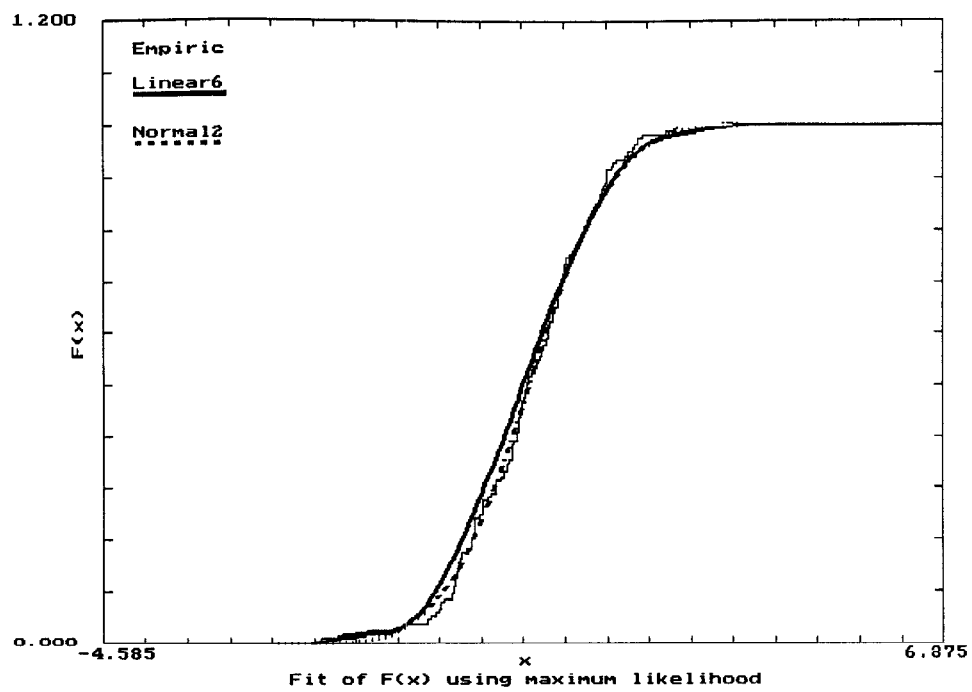


Figure 3: Cumulative Distribution Plots for Example 1

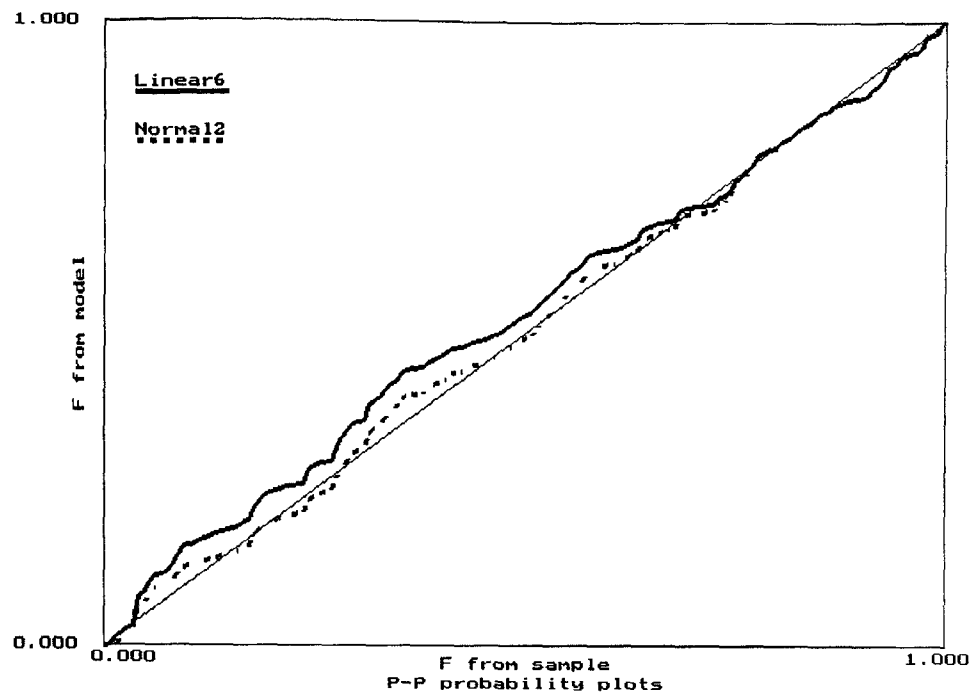


Figure 4: P-P Probability Plots for Example 1

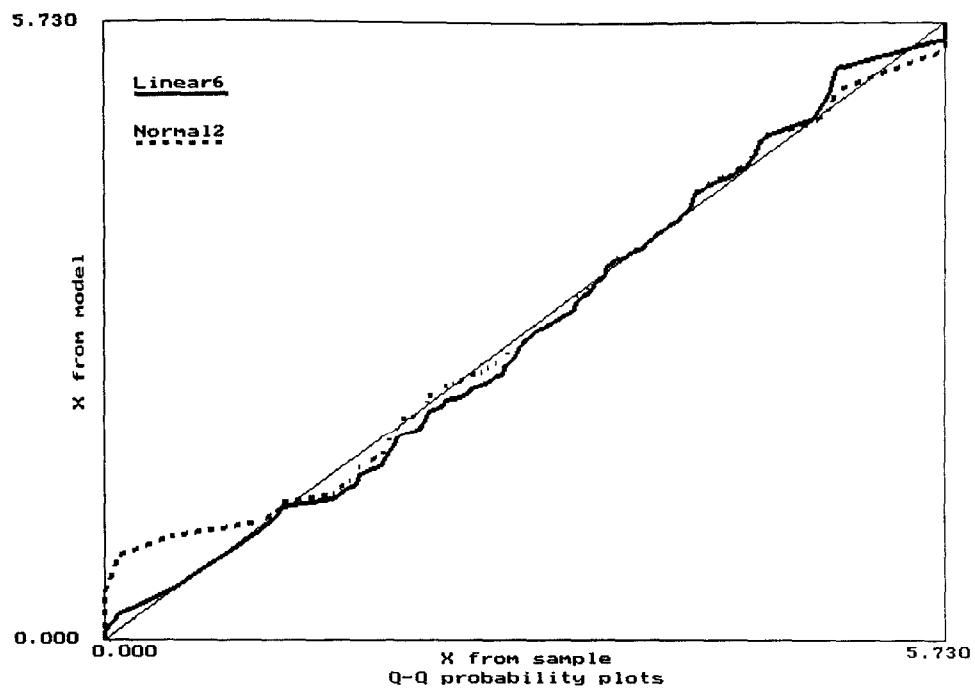


Figure 5: Q-Q Probability Plots for Example 1

Table 4: Sample Statistic Values for Example 2 — Demand Data

Kind	Size	Mean	Median	Mode		
Discrete	76	3.974	3.000	1		
Variance	Skewness	Kurtosis	SD	Minimum	Maximum	Range
12.426	1.559	6.982	3.525	0	20	20

Table 5: Distribution Rankings for Example 2

Distribution	fave	----- Parameter Estimates -----	
Empiric	0.09604		
NegBinom2	0.08637	2.000	0.320
Geometric1	0.08240	0.201	
Poisson1	0.05678	3.974	
DUniform2	0.04762	0.000	20.000
Binomial2	_____	_____	_____
HypGeom3	_____	_____	_____

Table 6: Expected Values and Statistics for Example 2

RV	E(X)	V(X)	-Ln(L'hood)/n	KS	Chi2
NegBinom2	4.251	13.224	2.449	0.064	21.800
Geometric1	3.965	19.478	2.496	0.096	21.459
Poisson1	3.974	3.974	2.869	0.205	1.77E+0006
DUniform2	10.000	36.667	3.045	0.492	103.053
Sample	3.974	12.426			

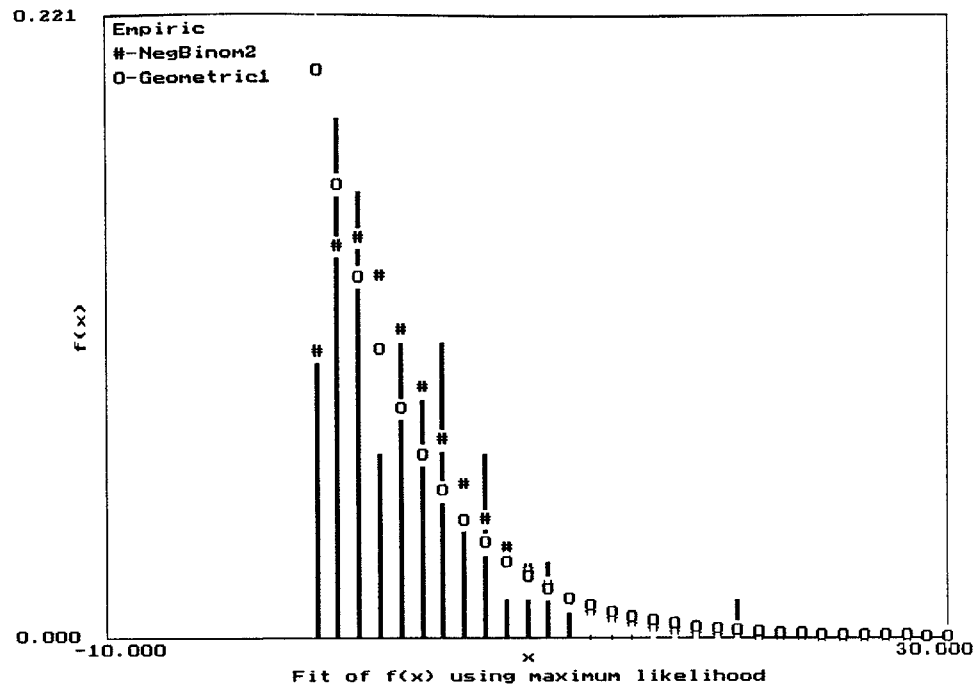


Figure 6: Density Plots for Example 2

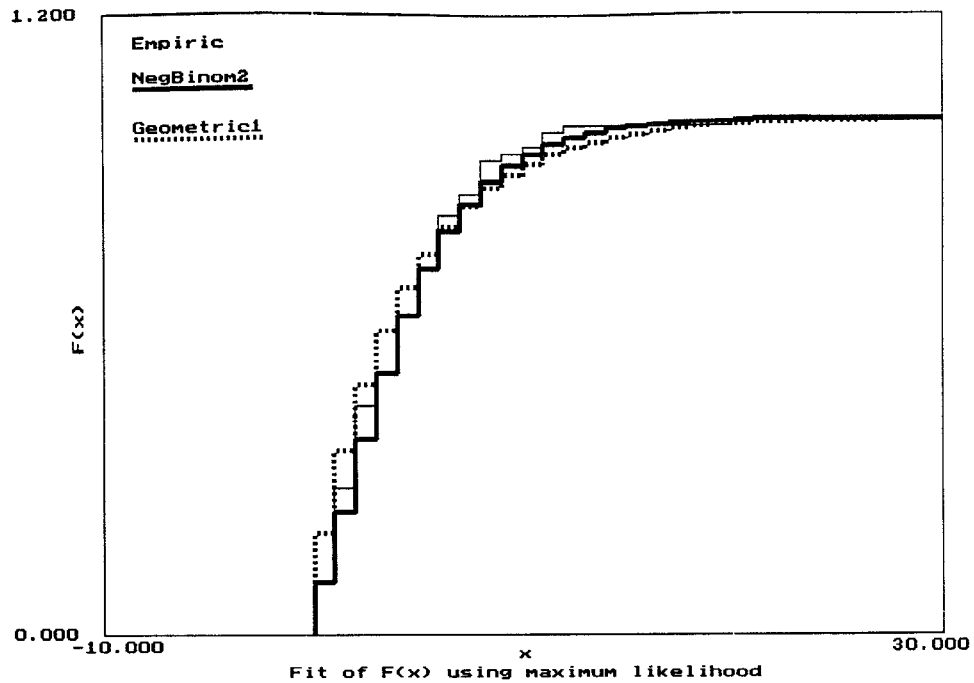


Figure 7: Cumulative Distribution Plots for Example 2

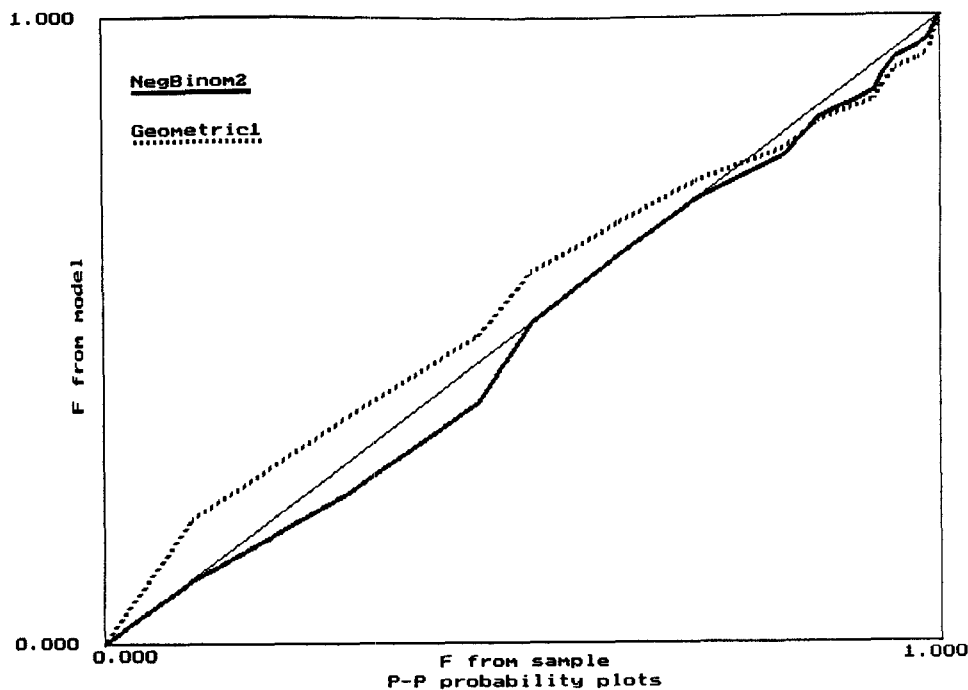


Figure 8: P-P Probability Plots for Example 2

upon maximum likelihood because the “peakedness” of the likelihood function at estimated values of the parameters (if in fact data was taken from that probability model).

- (iv) The piecewise linear family of random variables has several applications. First, comparison of the piecewise linear random variable with a competing more standard distribution, provides reference for using or not using the standard random variable. Secondly, if no one of commonly used distribution fits “adequately”, the linear family may be a quite feasible choice. In fact, use of the piecewise linear density and its smooth CDF is definitely preferable to direct use of the empirical distribution function.

5 AVAILABILITY OF RVL

An executable copy of RVL with on-disk documentation is available to those who request it. Teachers who want to use it in their courses are especially encouraged to acquire it. RVL is a quite usable prototype now (July 1992).

REFERENCES

- Booch, G. 1991. *Object Oriented Design With Applications*. Redwood City, CA: Benjamin Cummings.
- Dudewicz, E.J. and Mishra, S.N. 1988. *Modern Mathematical Statistics*. New York: Wiley.
- Hogg, R.V., Uthoff, V.A., Randles, R.H., and Davenport, A.S. (1972), “Selecting the Underlying Distribution and Adaptive Estimation,” *Journal of the American Statistical Association*, 67, 597-600.
- Johnson, N.L. and S. Kotz, 1970. *Continuous Univariate Distributions - 1*. New York: Wiley.
- Jones, G.W. 1990. *Software Engineering*. New York: Wiley.
- Law, A.M. and W.D. Kelton. 1991a. *Simulation Modeling and Analysis* 2d ed. New York: McGraw-Hill.
- Law, A.M. and W.D. Kelton. 1991b. *Solutions Manual* to accompany Simulation Modeling and Analysis, 2d ed. New York: McGraw-Hill.
- Law, A.M. and S. G. Vincent. 1991. UniFit II: Total Support for Simulation Input Modeling. In *Proceedings of the 1991 Winter Simulation Conference*, ed. B. Nelson, W. D. Kelton, and G. M. Clark, 136-142. Institute of Electrical and Electronics Engineers, Phoenix, Arizona.

- Press, W.H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. 1989. *Numerical Recipes* (in Pascal). New York: Cambridge University Press.
- Zaven, A.K, and Dudewicz, E.J. 1991. *Modern Statistical, Systems, and GPSS Simulation: The First Course*. New York. W.H. Freeman.

AUTHOR BIOGRAPHIES

SATYA N. MISHRA is an Associate Professor of Statistics in the Department of Mathematics and Statistics at the University of South Alabama. He received a B.Sc. degree in Mathematics, Physics, and Chemistry from the University of Gorakhpur (India) in 1966, M.Sc. in Applied Mathematics from Benares Hindu University in 1969, M.A. in Pure Mathematics from the University of Massachusetts in 1974, and M.S. and Ph.D. in Statistics from the Ohio State University in 1981 and 1982, respectively. His principal areas of research are selection theory, data analysis, and multiple comparison procedures.

RONALD S. REAGAN is an Associate Professor of Computer Sciences and Statistics at the University of Southern Mississippi Gulf Coast Regional Campus in Long Beach, Mississippi. He teaches Simulation and Modeling, Mathematical Statistics, Software Engineering, and Pascal programming. His research interests are Object Oriented Approaches to Statistical Modeling, Driver Risk Behavior at Grade Crossings, and Models for Occurrence of Hurricanes on the Mississippi Gulf Coast.