Proceedings of the 2004 Winter Simulation Conference
R. G. Ingalls, M. D. Rossetti, J. S. Smith, and B. A. Peters, eds.

INPUT MODELING USING QUANTILE STATISTICAL METHODS

Abhishek Gupta

Department of Industrial Engineering
Texas A & M University
College Station, TX 77843, U.S.A.

Emanuel Parzen

Department of Statistics
Texas A & M University
College Station, TX 77843, U.S.A.

ABSTRACT

This paper applies quantile data analysis to input modeling in simulation. We introduce the use of QIQ plots to identify suitable distributions fitting the data and comparison distribution *P-P* plots to test the fit. Two examples illustrate the utility of these quantile statistical methods for input modeling. Popular distribution fitting software often give confusing results, which is usually a set of distributions differing marginally in the test statistic value. The methods discussed in this paper can be used for further analysis of the software results.

1 INTRODUCTION

Defining input models to represent data is an important step in the simulation modeling process. In this paper we assume that data exists for the factors of interest. We could possibly use the data or its empirical distribution function to carry out the simulation. Fitting a theoretical distribution provides a structure to the simulation model which is useful in giving intuition into the deeper physical processes underlying an observed phenomenon and how the system behavior depends on specific characteristics of randomness associated with its different aspects.

The problem of selecting appropriate distributions to represent the different random quantities of the simulation model is an interesting and difficult problem from the point of view of the practitioner. The main difficulty is that it is a very soft issue. There are no hard and fast rules that can be followed in selecting distributions.

Many software products are now available which fit distribution functions to the data. Popular examples include BestFit, Arena Input Analyzer and ExpertFit. Input modeling that involves fitting standard univariate parametric probability distributions is typically performed using such an input modeling package. These packages fit several distributions to a data set, then determine the distribution with the best fit by comparing goodness-of-fit statistics like the Chi-Squared, Kolmogorov-Smirnov and Anderson-Darling. Cheng (1992) gives a good description of these tests with their advantages and disadvantages. These tests should be considered consultive only and using these tests to rank the fitting of the distributions leads to following problems:

1. The ranking of distributions is done by their p-values. Often the different distributions hardly differ in their p-values and the ranking in that case is indicative of the randomness of the data rather than the distribution fit.
2. The model chosen is often an overfit of the data. The whole goal of modeling is to get a size of the error which is reproducible, else a future sample will not agree with the current model.

To illustrate our point we will discuss the results from distribution-fitting software in the next section.

The purpose of this paper is to introduce the use of quantile statistical methods in simulation input modeling. These methods can be useful in giving additonal evidence in favour of or against the use of the selections suggested by distribution-fitting software. We try to layout steps to systematically choose the distribution functions which best fit the data using quantile statistical modeling methods as have been developed by Parzen (2003).

The format of the remainder of the paper is as follows. Section 2 discusses some issues that arise while using distribution-fitting software. Quantile methods are discussed in section 3. Section 3 also gives the proposed methodology for using these methods in simulation input modeling. Examples are in section 4 and section 5 provides the concluding remarks.

2 NOTE ON DISTRIBUTION-FITTING SOFTWARE

We consider the following data set that originally comes from the life testing literature (Lawless 1982, p.228). Here

we present this data sample of size $n = 23$ as order statistics (in increasing order).

17.88 28.92 33 41.52 42.12 45.6 48.48 51.84 51.96 54.12 55.56 67.8 68.64 68.64 68.88 84.12 93.12 98.64 105.12 105.84 127.92 128.04 173.4

This data set comes originally from a study on the fatigue life of deep groove ball bearings by Lieblein and Zelen (1956) and was analyzed as coming from a Weibull distribution. Lawless (1982) suggested a possible lognormal fit to the data and estimated the parameters and conducted a comparison of fits between Weibull and lognormal. This data set has been studied in detail from the simulation input modeling perspective by Leemis (2001) who regarded the data as service times in seconds. Leemis analyzed the Weibull in detail as a fit for the data.

To get an idea of the possible fitting distributions we used the distribution fitting software BestFit. Table 1 gives a summary of results based on the Anderson-Darling test (A-D test) and from the test statistics we can conclude that BestFit regards Inverse Gaussian, Lognormal, Pearson 5, Log Logistic and Extreme Value distributions as possible candidates.

Table 1: Results from Best-Fit for Lieblein and Zelen Data

Distribution	A-D Test Statistic value	A-D Test p-value
InvGauss	0.1849	N/A
Lognorm2	0.1863	N/A
Pearson5	0.1895	N/A
LogLogistic	0.1966	N/A
ExtValue	0.2302	> 0.25
Logistic	0.5114	$0.1 \leq p \leq 0.25$

BestFit does not fit Weibull to the data at all, which is in contradiction to the analysis available in the literature on this data set. As will be seen from the preliminary diagnostics (to be presented in the next section), at least from initial results we have no basis to reject Weibull or gamma.

These results from BestFit raise important questions about the results that we get from distribution fitting software.

One observation that we can make is that the A-D test statistic values are almost same for Inverse Gaussian, Lognormal, Pearson 5 and Log Logistic. If we analyze the *P-P* plots, they also turn out to be almost the same. The extreme value distribution also has a *P-P* plot which is only slightly different, and this is because the A-D test statistic value is not very different. Thus we have five distributions which characterize the data exactly the same way.

The aim of input modeling is to choose a model with some distinguishing characteristics but here all the 5 models fit the same. We cannot differentiate one model from the other. We really cannot pick a better model if they all have same goodness of fit!

Why do these models fit the same? This is because BestFit is not using maximum likelihood estimators (mle) for parameter estimation. In fact, the mle is just an initial guess for an algorithm that uses the Levenberg-Marquardt method (Jankauskas and McLafferty 1996). The optimization routine aims to minimize the test statistic value by iteratively modifying the parameter estimates and thus get a better goodness-of-fit between the data set and a distribution function. This estimation procedure does not have any good properties of maximum likelihood or the method of moments, and we do need to ask this question: "is this fitting or over-fitting?". This seems more of over fitting and a distribution which is recommended using this procedure does not stand a chance of repeating its performance (i.e. of getting a low test statistic value) to fit another sample.

The method of ranking distributions on the basis of test statistic value is also fallacious. Most distribution-fitting software use a *p-value* to represent the test. Since we consider random data, the test statistics or the p-value can be different for different samples from the same population. We need to remember that we are testing statistical significance, and any distribution which passes the test irrespective of the test statistic value is an equally good choice from the point of view of the goodness of fit test. In fact if the distribution (with parameters being estimated as in this example) is a very good fit for one sample, then it is very unlikely that it will be a good fit for another sample.

An interesting observation is that if we generate a dataset of 250 from the standard Weibull and use BestFit for distribution fitting the Weibull does not fit at all. This is surprising as we would expect Weibull to at least pass the goodness-of-fit test. The result from Arena's Input Analyzer is different. Input Analyzer identifies Weibull as the best fit. One of the reasons why it outperforms BestFit may be because Input Analyzer decides from a pool of 12 distributions as compared to BestFit which compares 28 distributions.

We need to remember that we do not want to answer the question which is the best fitting model. We aim to provide good alternatives which fit the data well. An option should be for the final decision to be taken by the simulation practitioner based on his experience and any physical interpretation that a chosen distribution can give to the observed process.

3 THE QUANTILE METHODS APPROACH

Any input modeling exercise involves the following steps:

1. Testing the assumptions.
2. Selection of possible models.
3. Estimation of parameters.
4. Analysis of the fit.

It is in the selection of possible models and analysis of the fit that quantile methods can play a role in facilitating in modeling. We suggest an algorithmic approach to model selection and testing. First we give a brief explaination of the key concepts and then their incorporation in input modeling.

For the random variable Y, if the distribution function is $F(y) = P[Y \leq y]$ then the quantile function is given by $Q(u) = F^{-1}(u)$. The general definition is given as

$$Q(u) = F^{-1}(u) = \inf(y : F(y) \geq u) \qquad \text{for} \quad 0 \leq u \leq 1.$$

Important concepts are quartiles $Q_j = Q(j/4)$ for $j = 1, 2, 3$, and mid-distribution $F^{\text{mid}}(y) = F(y) - .5P[Y = y]$ which treats tied data by computing their mid-ranks.

Defining mid-quartile $MQ = 0.5(Q_1 + Q_3)$ and measure of deviation $IQR2 = 2(Q_3 - Q_1) = 2 \times$ interquartile range, the quantile/quartile function $QIQ(u)$ as defined by Parzen (2003) is

$$QIQ(u) = \frac{Q(u) - MQ}{IQR2}. \tag{1}$$

One explanation of the powerful insight provided by the quantile-quartile function $QIQ(u)$ is that its five percentile summary $QIQ(u), u = .05, .25, .5, .75, .95$ has the following interpretations: $QIQ(.25) = -.25$, $QIQ(.75) = .25$, universal normalizations; $QIQ(.5)$, identify skewness; $QIQ(.05)$ and $QIQ(.95)$ identify respectively left tail and right tail behavior.

For data analysis we plot the sample quantile/quartile function $Q^{\sim c}IQ^{\sim c}(u)$. The $Q^{\sim c}IQ^{\sim c}(u)$ can be used to draw conclusions about possible distributions fitting the data. The $QIQ(u)$ plot can be interpreted as a Goodness of Fit test without using the location and scale parameters and just using the shape for comparison.

The well known quantile-quantile plots (QQ plots) are commonly used to test the normality assumption which is required by statistical procedures like those using t or F distributions. The quantile-quartile plots (QIQ plots) are NOT QQ plots; they are normalized to satisfy $QIQ(0.25) = -0.25$ and $QIQ(0.75) = 0.25$. They help identify for a sample distribution skewness, tail-behavior, possible bimodality of the data and closeness to a population location-scale distribution.

To test the closeness of a distribution function F_1 and distribution G, the comparison distribution (**?**) is defined as $D(u; F_1, G) = G(F_1^{-1}(u)) = u$ for $0 \leq u \leq 1$. When F_1 is a sample distribution and G is a population distribution; uniformity is equivalent to testing the null hypothesis of equality of F_1 and G. A test can be performed by constructing the graph of D and checking whether it lies on a 45-degree line. The graph of D is known as the percentile-percentile or probability-probability(*P-P*) plot. While QIQ plots can be very powerful in identifying distributions, they are a preliminary diagnostic to be followed by *P-P* plots which require the estimation of the parameters of the fitted distribution, and thus can be used to draw final conclusions regarding the distribution fits. Thus QIQ plots can be used to screen distributions and *P-P* plots can be used to give judgement on fit of the distributions. These methods can help us and should be used, to gain a better understanding of the recommendations being made by the distribution-fitting software. They enable the final decision on the choice of the distribution to be made by the simulation practitioner based on physical interpretation and experience.

The *P-P* plot that we use (and recommend because of theoretical properties) is slightly different from those that we get from standard software like SAS or BestFit. We use the mid-distribution function to plot the *P-P* plot, thus our treatment of ties is different as explained earlier. But like SAS we recommend plotting the *P-P* plot with the sample mid-distribution function on the Y-axis and the fitted cumulative distribution function on the X-axis. This is advantageous as it gives an intuition on non-parametric density estimation of the unknown theoretical distribution of the observed data.

The rest of the paper primarily discusses the applications of QIQ plots and *P-P* plots to input modeling. We propose the following steps to be incorporated when using quantile methods :

Stage 1	Compute the sample quantile function $Q^{\sim c}(u)$.
Stage 2	From $Q^{\sim c}(u)$ we define the median, quartiles, IQR2 and the other diagnostics.
Stage 3	Plot the QIQ function and compare with the QIQ plots of exponential and normal to get a feeling of the distribution. The normal and exponential are chosen as reference distributions because they are amongst the most popular univariate distributions that are encountered in theoretical and applied statistics.
Stage 4	Compare the QIQ function plot with possible fits.
Stage 5	Do the *P-P* plot amongst possible candidate distributions to judge the best fit.

4 EXAMPLES

We now illustrate the use of quantile methods in simulation using two examples from literature.

4.1 Service Time Model

We consider the data from Lieblein and Zelen (1956) which was introduced in section 2. As mentioned earlier, Leemis (2001) considered this data set as service times of a queuing system.

4.1.1 Quantile Methods Analysis

First we compute the sample quantile function $Q^{\sim c}(u)$ from the data. We now compute the median, the quartiles, the inter quartile range and other diagnostics. Table 2 shows the numerical summary of the data. Since absolute value of the skew index is less than 0.0625, the distribution is symmetric. The tail indices suggest that the distribution has short left tail ($-0.5 \leq QIQ(0.05) \leq -0.25$), medium right tail ($.5 \leq QIQ(0.95) \leq 1$) and no outliers. Details of the quantile/quartile diagnostics of tail are given in Parzen (2003).

Table 2: Numerical Summary for Lieblein and Zelen Data

Sample MIN	17.88
Sample MAX	173.4
$Q1$	46.32
$Q2$	67.80
$Q3$	97.26
MQ	71.79
$IQR2$	101.88
Skew index $(Q2 - MQ)/IQR2$	-0.0392
Left tail index, $QIQ(0.05)$	-0.4587
Right tail index, $QIQ(0.95)$	0.708
$(MIN - MQ)/IQR2$	-0.53
$(MAX - MQ)/IQR2$	0.99
Conclusion	Symmetric

We compare the sample quantile/quartile function plot with those obtained from normal and exponential (Figure 1). From the plot we note that the distribution is between normal and exponential. We could claim that the exponential may fit in the middle, but as can be seen from the QIQ plot the sample definitely does not exhibit the tail behavior of an exponential distribution. The normal distribution does not fit the sample in the middle (which is unusual); horizontal shape indicates many ties (clustering) in the middle of the sample.

From the plot we conclude that the sample is NOT from normal or exponential but from a distribution like gamma, Weibull or log-normal. Since the data from Lieblein and Zelen was from a survival analysis, we analyze gamma and Weibull in more detail.

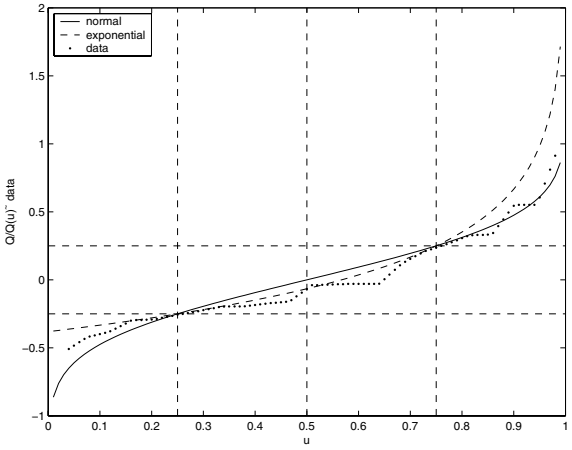


Figure 1: $Q^{\sim c}IQ^{\sim c}(u)$ Plot with QIQ Normal and QIQ Exponential

The Weibull distribution has probability density function given by

$$f(x) = \lambda^{\kappa}\kappa x^{\kappa-1}e^{-(\lambda x)^{\kappa}} \qquad x \geq 0, \tag{2}$$

where λ is a positive scale parameter and κ is a positive shape parameter. For the purposes of QIQ plot we are interested in the standard Weibull (κ) distribution with $\lambda = 1$. Figure 2(a) shows the QIQ plot for κ =2. Further details on choosing the shape parameter for plotting the Weibull and gamma QIQ plots are given in appendix A.

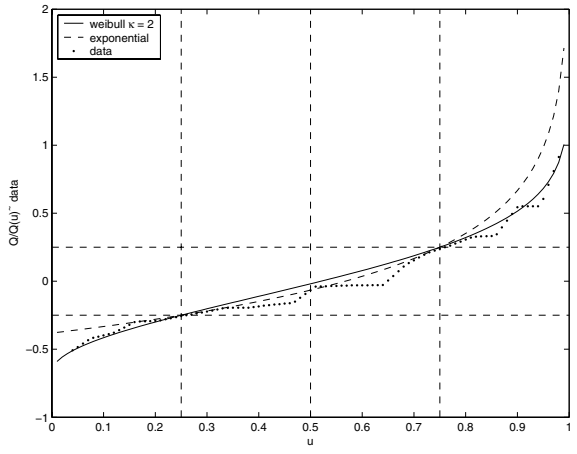
From the QIQ plot in Figure 2(a) we can conclude that the Weibull(κ) fits the data fairly well especially in the tails, but not as well in the middle. The sample QIQ plot shows a flatness in the middle, from $u = 0.4$ to $u = 0.6$. This is because approximately 1/5th of the data is around the median. But this flatness cannot be suggested as a deviation from Weibull, especially since the Weibull does a good job of fitting in the tails. Using the MATLAB function *weibfit* we get the maximum likelihood estimates of the Weibull parameters as $\kappa = 2.102$ and $\lambda = 0.0125$.

The gamma distribution has probability density function given by

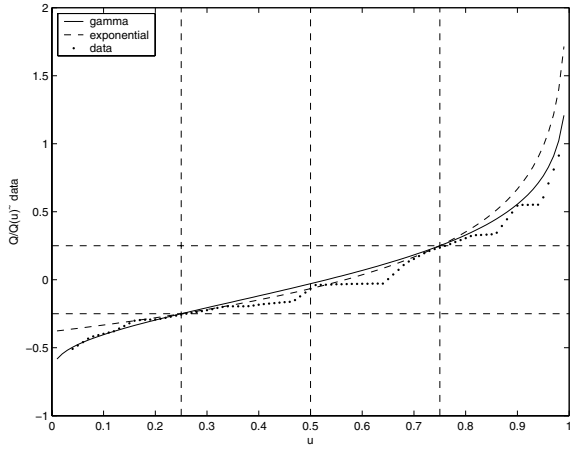
$$f(x) = \frac{1}{\beta^{\gamma}\Gamma(\gamma)}x^{\gamma-1}e^{-\frac{x}{\beta}} \qquad x \geq 0; \gamma, \beta > 0 \tag{3}$$

where γ is a positive shape parameter and β is a positive scale parameter. The standard gamma (γ) distribution has $\beta = 1$. Figure 2(b) shows the QIQ plot for γ =4.

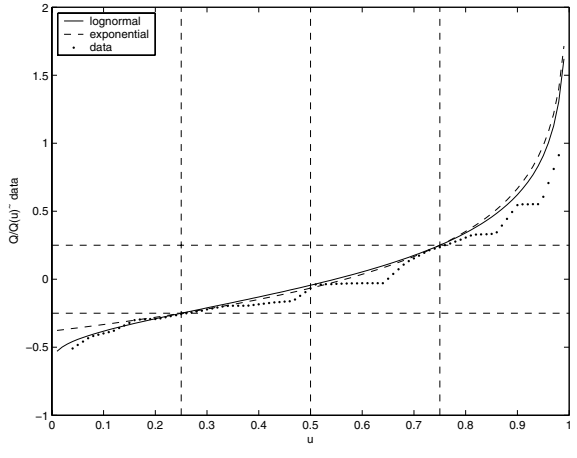
From the QIQ plot we can conclude that the gamma distribution fits the data fairly well too. Using the MATLAB function *gamfit* we get the maximum likelihood estimates of the gamma parameters as $\gamma = 4.0255$ and $\beta = 17.9419$.



(a) $Q^{\sim c}IQ^{\sim c}(u)$ Plot with QIQ Weibull ($\kappa = 2$) and QIQ Exponential



(b) $Q^{\sim c}IQ^{\sim c}(u)$ Plot with QIQ Gamma ($\gamma = 4$) and QIQ Exponential



(c) $Q^{\sim c}IQ^{\sim c}(u)$ Plot with QIQ Lognormal ($\sigma = 0.53$) and QIQ Exponential

Figure 2: QIQ Plots for the Lieblein and Zelen Data

The probability density function of the lognormal distribution is

$$f(x) = \frac{e^{-((\ln(x-\theta)-\zeta)^2/(2\sigma^2))}}{(x-\theta)\sigma\sqrt{(2\pi)}} \qquad x \geq 0; \zeta, \sigma > 0 \quad (4)$$

where σ is the shape parameter, θ is the location parameter and ζ is the scale parameter. The standard lognormal (σ) distribution has $\theta = 0, \zeta = 0$.

The QIQ plot for lognormal is shown in Figure 2(c). We have chosen the shape parameter σ to be 0.53 which is the maximum likelihood estimate. Since the location and scale parameter do not effect the QIQ plot we use the standard lognormal with $\sigma = 0.53$. The lognormal fits well in the left tail; right tail fit is not good because lognormal has longer right tail than sample.

The preliminary diagnosis using QIQ plots provides evidence that gamma, Weibull and possibly lognormal could be the fitting distribution. We need to do further analysis using *P-P* plots to be able to come to a conclusion. A final conclusion also needs to be based on a possible physical interpretation. Using the mle estimators we plot the *P-P* plots (Figure 3) for Lieblin and Zelen data.

The data has a cluster near the median, and the *P-P* plot gives a clear evidence of a sharp blip in the region from $u = 0.4$ to $u = 0.6$. This is a crucial property of the data and is not being picked out. We can choose to ignore this clustering and go ahead with a parametric distribution, or alternatively we can fit a non-parameter density function depending on our objectives.

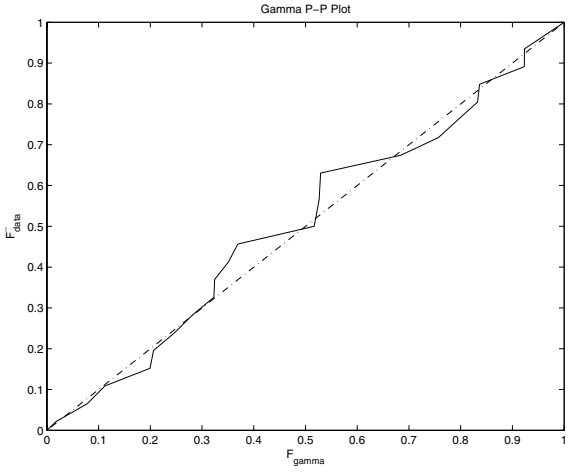
The gamma seems to fit the data better as compared to Weibull (poorer fit near 0) and lognormal (poorer fit near 1).

4.2 LV Model

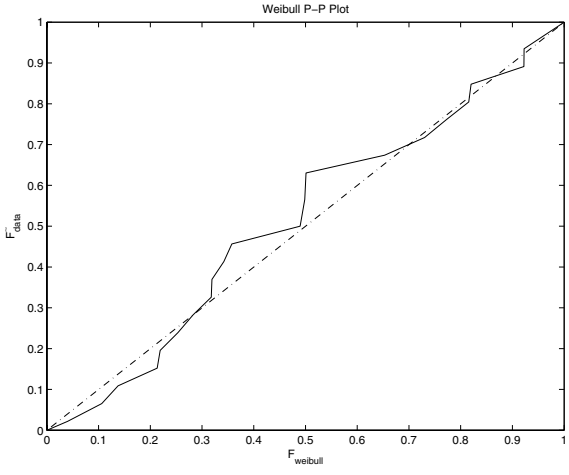
Cheng, Holland, and Hughes (1996) gives the data of the times in seconds to serve Light Vans (LV) at the toll booths of the Severn Bridge River crossing in Britain. The data in ascending order is :

3.1 3.1 3.3 3.5 3.6 3.8 3.9 4 4 4.1 4.2 4.3 4.3 4.4 4.5 4.6 4.7 4.7 4.7 4.8 4.9 5 5.2 5.2 5.7 5.8 5.8 6.1 6.2 6.3 6.3 6.4 6.4 6.6 6.7 6.7 7.2 7.2 7.7 7.8 7.9 8 8 8.2 10.5 10.9 12.5

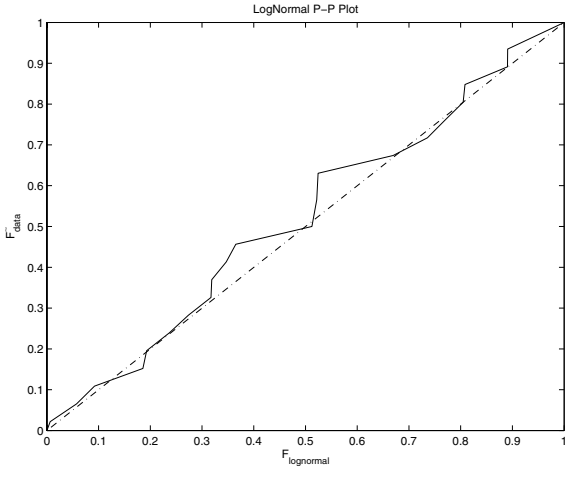
The interesting aspect of this data set (which we refer to as LV data) is the presence of ties. This is an illustration that the assumption of no ties in data (which is used while deriving goodness-of-fit test statistics) may not be valid in real scenarios.



(a) Gamma ($\gamma = 4.03,\ \beta = 17.9$)



(b) Weibull ($\kappa = 2.102,\ \lambda = 0.0122$)



(c) Lognormal($\theta = 0,\ \sigma = .53,\ \zeta = 4.2$)

Figure 3: *P-P* Plots for the Lieblein and Zelen Data

4.2.1 Quantile Methods Analysis

From the sample, we compute the sample quantile function and the associated numerical summary diagnostics like median, quartiles and tail indices. The skew index value ($= -0.0715$) indicates that the distribution is not symmetric (skew index diagnostic absolute value must be less than 0.0625 for symmetry). The tail indices suggest that the distribution has a short left (left tail index absolute value $= -0.4524$, below .5) and long right tail (right tail index $= 1.0159$ above 1).

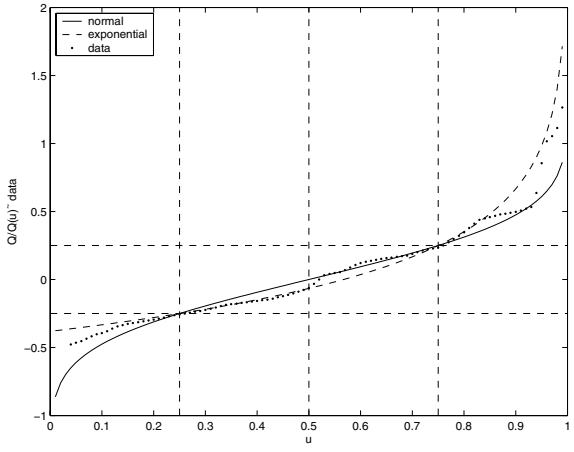


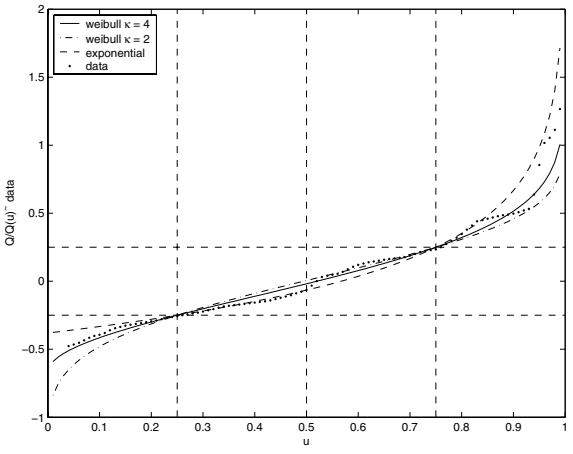
Figure 4: $Q^{\sim c}IQ^{\sim c}(u)$ Plot with QIQ Normal and QIQ Exponential for the LV Data

As in the previous example, we first compare the sample quantile/quartile function plot with those obtained from normal and exponential (Figure 4). From the plot we can conclude that the sample is NOT from normal or exponential but from a distribution like gamma, Weibull or log-normal.

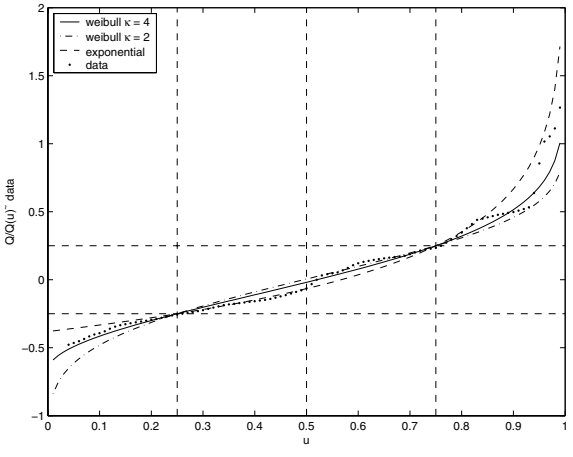
We now plot the QIQ plot using the standard Weibull pdf with values of $\kappa = 2$ and $\kappa = 4$. Figure 5(a) shows the QIQ plots. From the QIQ plot we can conclude that the Weibull does not fit the data well in the right tail. Using the MATLAB function *weibfit* we get the maximum likelihood estimates of the Weibull parameters as $\kappa = 2.9347$ and $\lambda = 0.1536$.

Figure 5(b) shows the QIQ plot for gamma with $\gamma = 4$. From the QIQ plot we can conclude that the gamma distribution fits the data fairly well. Using the MATLAB function *gamfit* we get the maximum likelihood estimates of the gamma parameters as $\gamma = 9.2047$ and $\beta = 0.6306$.

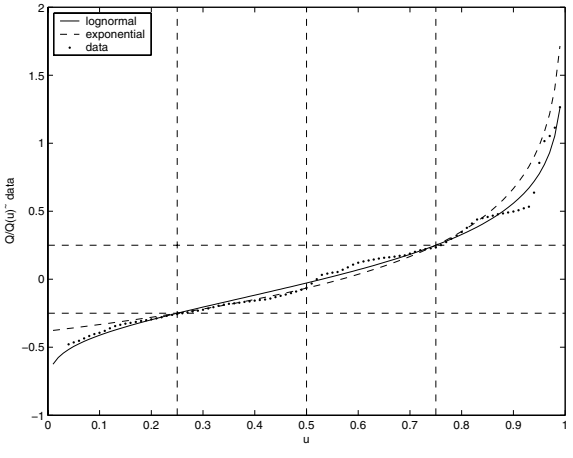
From the QIQ plot for lognormal in Figure 5(c), we conclude that the lognormal distribution does a good job in fitting the data well. It fits well especially in the right tail, which is long for the sample and long for the distribution. The maximum likelihood estimates of lognormal are calculated to be $\sigma = 0.33$ and $\zeta = 1.7$. We use the shape parameter to be the maximum likelihood value while plotting the QIQ plot.



(a) $Q^{\sim c} IQ^{\sim c}(u)$ Plot with QIQ Weibull ($\kappa = 2$), QIQ Weibull ($\kappa = 4$) and QIQ Exponential

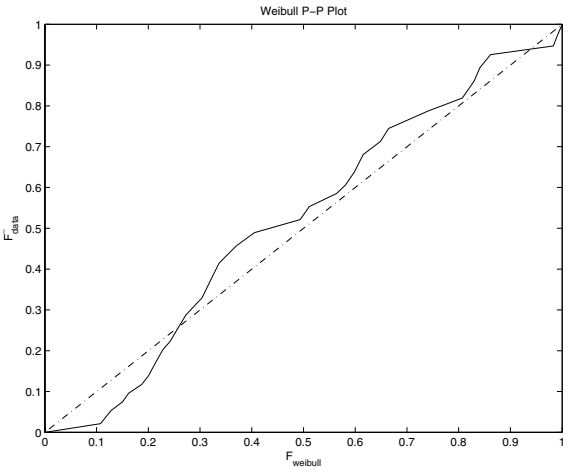


(b) $Q^{\sim c} IQ^{\sim c}(u)$ Plot with QIQ Gamma ($\gamma = 4$) and QIQ Exponential

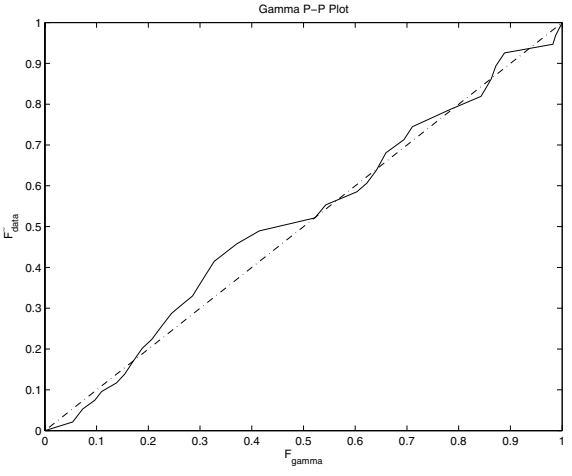


(c) $Q^{\sim c} IQ^{\sim c}(u)$ Plot with QIQ Lognormal ($\sigma = 0.33$) and QIQ Exponential

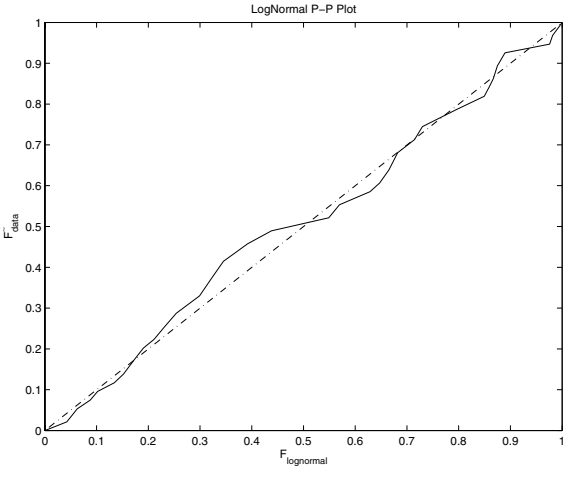
Figure 5: QIQ Plots for the LV Data



(a) Weibull ($\kappa = 2.9347,\ \lambda = 0.1539$)



(b) Gamma ($\gamma = 9.2047,\ \beta = 0.6306$)



(c) Lognormal($\sigma = 0.33, \zeta = 1.7$)

Figure 6: *P-P* Plots for the LV Data

4.2.2 Results from the Software

The BestFit rankings of the fitting distributions for the LV data are given in Table 3. Inverse Gaussian, Weibull, Pearson 5, Log Logistic and Extreme Value appear to be the fitting distributions.

The *P-P* plot of Weibull (Figure 6(b)) shows that Weibull with maximum likelihood estimates does not fit the sample data well in left tail and thus the choice made by BestFit is not fully supported by the quantile analysis. On the other hand, the gamma *P-P* plot in Figure 6(a) shows that gamma with maximum likelihood estimates fits the sample data well.

Another observation that we make is the high A-D test statistic value for the lognormal distribution calculated by BestFit. Cheng, Holland, and Hughes (1996) calculates the value of the A-D statistic as 0.227 as compared to the BestFit value of 18.158. These sharply different values between the computer and the literature raises some concern. The *P-P* plot of lognormal distribution in Figure 6(c) gives evidence that the lognormal with maximum likelihood estimates is a good fit for the sample data. We would like to point out that the formula for the A-D test statistic computation has been derived under the assumption of distinct values.

The A-D test statistic is a numerical diagnostic for the P-P plot; it is a measure of the "distance" between the empirical and theoretical distribution function. Since, the number from the software does not agree with the graph for which it stands we need to ask "Who are you going to believe ? the number or the evidence from the graph for which it stands".

Table 3: Results from Best-Fit

Distribution	A-D Test Statistic value	A-D Test p-value
InvGauss	0.2336	N/A
Weibull	0.2609	N/A
Pearson5	0.2615	N/A
LogLogistic	0.3194	N/A
ExtValue	0.3557	> 0.25
Logistic	0.7847	$0.025 \leq p \leq 0.05$
Normal	1.1138	$0.005 \leq p \leq 0.01$
Expon	1.5131	< 0.01
Uniform	12.3088	N/A
Lognorm2	18.1558	N/A

5 CONCLUDING REMARKS

Input modeling is an important issue in simulation modeling. Recent times have seen advances in software for distribution fitting and these software are now being widely used by simulation practitioners for input modeling. Instead of treating the answers from the software as from a black box, it is crucial to perform one's personal reasoning to interpret numbers and diagnostics. Our practice should regard that analysis is cheap since data is costly. In this spirit we present the quantile methods as a facilitator in improving the results (and the interpretations of the results) that we get from popular software. We develop a systematic input modeling strategy using quantile methods, especially QIQ plots and *P-P* plots. The examples shown illustrate the importance of these methods in improving the results that we get from popular software.

The real issue may be the overdependence of present analysis on goodness of fit tests like Chi-Squared, Kolmogorov-Smirnov and Anderson-Darling. We need to realize the numbers associated with these statistics are just diagnostics for the discrepancy between the empirical sample distribution and the theoretical distribution. The software may not give enough weight to all aspects of the fit, and are especially susceptible to miss out on giving importance to the tails. We cannot just argue on the basis of a number, it is always helpful to support the analysis the picture provided by the *P-P* plots.

Central to the quantile methods is the definition of the mid-distribution function which is required to handle the case of data with ties.

A Deciding the Shape Parameter for Plotting the Gamma and Weibull QIQ Plots

Write the standard gamma distribution density function with $b = 1/\gamma$,

$$f(x) = \frac{x^{\frac{1}{b}-1}}{\Gamma(\frac{1}{b})} e^{-x} \qquad x \geq 0; \quad b > 0.$$

For $b = 1$ we get the exponential distribution and $b \to 0$ gives the extreme value distribution. We check the exponential and normal QIQ plots in the beginning of our analysis, and so for the gamma QIQ plots, we propose plotting $b = 0.25, 0.5, 0.75$. As can be seen from a pdf plot the important features with regards to the exploratory data analysis are captured by this range of b values for the gamma. As the b value comes closer to 0 the graph tends to become more and more flatter to the x-axis. It is clear from the QIQ plots for these b values that the choice of the shape parameter largely depends on the tail behavior exhibited and this can be easily done using these plots.

A similar line of reasoning can be developed for the Weibull distribution and for a preliminary diagnostic we just choose $b = 0.25, 0.5, 0.75, 1$ where $b = 1/\kappa$. Plotting the Weibull pdf for different b values it can easily be seen that as $b \to 0$ the distribution becomes sharper. As was the case for gamma, the choice of shape parameter for Weibull depends on the tail behavior.

REFERENCES

Cheng, R. C. H. 1992. Distribution fitting and random number and variate generation. In *Proceedings of the 1992 Winter Simulation Conference*, ed. J. J. Swain, D. Goldsman, R. C. Crain, and J. R. Wilson, 74–81. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Cheng, R. C. H., W. Holland, and N. A. Hughes. 1996. Selection of input models using bootstrap goodness-of-fit. In *Proceedings of the 1996 Winter Simulation Conference*, ed. J. M. Charnes, D. J. Morrice, D. T. Brunner, and J. J. Swain, 199–206. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Jankauskas, L., and S. McLafferty. 1996. Bestfit, distribution-fitting software by palisade corporation. In *Proceedings of the 1996 Winter Simulation Conference*, ed. J. M. Charnes, D. J. Morrice, D. T. Brunner, and J. J. Swain, 551 – 555. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Lawless, J. F. 1982. *Statistical models and methods for lifetime data*. New York: John Wiley & Sons.

Leemis, L. 2001. Input modeling techniques for discrete-event simulations. In *Proceedings of the 2001 Winter Simulation Conference*, ed. B. Peters, J. Smith, D. Medeiros, and M. Rohrer, 62–73. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Lieblein, J., and M. Zelen. 1956. Statistical investigation of the fatigue life of deep-groove ball bearings. *Journal of Research of the National Bureau of Standards* 57:273–316.

Parzen, E. 2003. Quantile probability and statistical data modeling. Technical report, Department of Statistics, Texas A & M University. Accepted for publication in *Statistical Science*.

AUTHOR BIOGRAPHIES

ABHISHEK GUPTA is a graduate student in the Department of Industrial Engineering at Texas A & M University. He received the Bachelor of Technology degree in Mechanical Engineering from the Indian Institute of Technology, Delhi. In 2004 he received the Mary G. Natrella scholarship from the Quality and Productivity Research section of the American Statistical Association. He is a student member of IIE, ASA, IMS and SIAM. His research interests include engineering statistics and data analysis. His e-mail address is <abhishek.gupta@tamu.edu>.

EMANUEL PARZEN, Distinguished Professor of Statistics at Texas A & M University, was born in New York City on April 21, 1929, and educated at Harvard (B.A. 1949) and University of California Berkeley (Ph.D. 1953). He has served as a Statistics faculty member at Columbia (1953-56), Stanford (1956-70), SUNY Buffalo (1970-1978), Texas A & M (1978-Present), and a visiting faculty at Imperial College London, M.I.T., IBM, Harvard, and The Center for Advanced Study in the Behavioral Sciences. In 1994 he was awarded the Samuel S. Wilks Memorial Medal of the American Statistical Association with the following citation:

For outstanding research in Time Series Analysis, especially for his innovative introduction of reproducing kernel spaces, spectral analysis and spectrum smoothing; for pioneering contributions in quantile and density quantile functions and estimation; for unusually successful and influential textbooks in Probability and Stochastic Processes; for excellent and enthusiastic teaching and dissemination of statistical knowledge; and for a commitment to service on Society Councils, Government Advisory Committees and Editorial boards. His e-mail address is <eparzen@stat.tamu.edu>.