Proceedings of the 2011 Winter Simulation Conference
S. Jain, R.R. Creasey, J. Himmelspach, K.P. White, and M. Fu, eds.

USING PEARSON TYPE IV AND OTHER CINDERELLA DISTRIBUTIONS IN SIMULATION

Russell Cheng

University of Southampton
Highfield
Southampton, SO17 1BJ, UNITED KINGDOM

ABSTRACT

Univariate continuous distributions with unbounded range of variation have not been so widely used in simulation as those that are bounded (usually to the left). However situations do occur when they are needed, particularly in operations research and financial applications. Two distributions that have such unbounded range are the Pearson Type IV and Johnson SU distributions. Though both are well known in statistics, there is still a lack of methods in the literature for fitting these distributions to data which are both efficient and comprehensively reliable. Indeed the Pearson Type IV has the reputation of being difficult to fit. In this paper we identify the pitfalls and propose a fitting method that avoids them. We also show how to test the goodness of fit of estimated distributions. All the procedures described are included as VBA code in an accompanying Excel workbook. Two numerical examples are described in detail.

1 INTRODUCTION

In simulations there is occasionally the requirement to model continuous random variates where the potential range of variation is unbounded both to left and right. The normal distribution is the preeminent distribution with this property, but problems do occur where a more flexible distribution is needed that is more skew say, or more heavy tailed. Two well-known distributions, the Pearson Type IV (PT IV) with probability density function (PDF):

$$f^{IV}(y) = \frac{\left|\Gamma(b+bci\right|^2\Gamma(b)}{(\Gamma(b))^2\Gamma(b-0.5)\Gamma(0.5)}\frac{\exp\{2bc\arctan[(y-\mu)/\tau]\}}{\tau\{1+[(y-\mu)/\tau]^2\}^b} \tag{1}$$

and the Johnson SU (JSU), with PDF:

$$f^{SU}(y) = \frac{\delta}{\sqrt{2\pi}\sqrt{(y-\xi)^2+\lambda^2}}\exp\left\{-\frac{1}{2}\left(c+\delta\log\left[(y-\xi)/\lambda+\sqrt{1+[(y-\xi)/\lambda]^2}\right]\right)\right\} \tag{2}$$

are unbounded in this way. However, despite their long history, the literature still lacks a definitive account of fitting methods that are both efficient and comprehensively reliable, and indeed the PT IV has the reputation of being difficult to fit. We might regard them as 'Cinderella' distributions, possibly attractive but still not widely accepted and not reaching their full potential.

In this paper we consider how both distributions can be reliably fitted to a random sample $\mathbf{y} = (y_1, y_2, ..., y_n)$ of n independently and identically distributed observations, using the method of maximum likelihood (ML).

We begin by reviewing salient properties of each distribution and methods that have been proposed for fitting them to data, pointing out the pitfalls involved.

Firstly, each distribution is one of a larger family, members of which are conveniently characterised by the values of their squared skewness, $\gamma^2 = \beta_1$, and kurtosis, β_2. For both families the (β_1, β_2) plane can be divided into disjoint regions, each representing a particular distribution type belonging to the given family. Figure 1 depicts this, with the plot following the convention, established by Pearson (1916), of β_2 increasing downwards. Our plot depicts a wider range of β_1 and β_2 than conventionally shown, and highlights certain characteristics of each region, not apparent in most published plots.

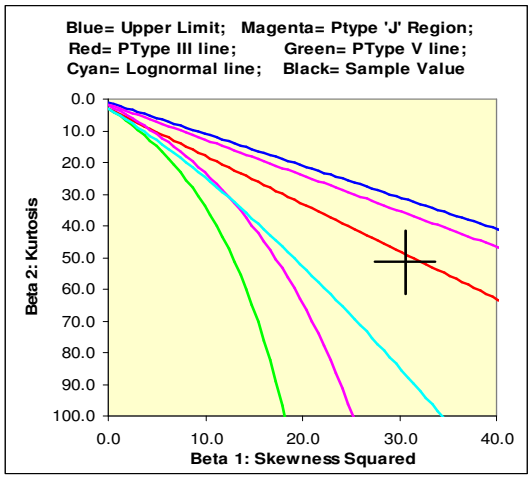

Figure 1: β_1 v. β_2 plot showing the boundaries of regions covered by different members of the Pearson and Johnson families of distributions. The entire region below the light green line is the Pearson type IV (PT IV) region. The entire region below the cyan line is the Johnson SU (JSU) region. The black cross shows the sample point corresponding to the hospital data set discussed in Section 5.2

A point of immediate note is that each point in the (β_1, β_2) plane represents TWO distributions from each family, according to the sign of the skewness γ so that one distribution is the mirror image of the other. There is an immediate issue here in that both distributions may not be representable by the functional form used to define the given member of the family. This problem does not occur in the specific cases of either the PT IV or JSU distributions, but does with certain other members. For example with the gamma distribution, which is a Pearson Type III, the conventional representation is appropriate only for positive skewness; the mirror distribution requires certain changes of sign in the formula for the PDF. When considering such members this slight technicality would have to be properly addressed if both negative and positive skewness is to be allowed.

For simplicity of exposition we consider just the case where data has come from a positively skew, or at worst a symmetric distribution. We hope to discuss the problem of finding the best fit from the full Pearson family, or from the full Johnson family, elsewhere. In this paper we focus simply on the problem of finding the best PT IV distribution or the best JSU distribution.

However we will allow the possibility that the best fit occurs at a boundary point. In the PT IV case the key boundary corresponds to the PT V (inverse gamma) distribution with PDF

$$f^V(y) = \frac{1}{\lambda\Gamma(\beta)}\left(\frac{\lambda}{(y-\xi)}\right)^{\beta+1} \exp\left(-\frac{\lambda}{(y-\xi)}\right), \quad y-\xi>0,\ c\neq 0\ . \tag{3}$$

This PT V boundary line is shown in Figure 1. The line is defined, for example, in Stuart and Ord (1987, Section 6.2). The line comprises points (β_1,β_2) that are valid solutions of the equation

$$\beta_1(\beta_2+3)^2 = 4(4\beta_2-3\beta_1)(2\beta_2-3\beta_1-6)\,,$$

i.e. solutions lying below, in the sense of Figure 1, the line $\beta_2-\beta_1-1=0$, this latter condition being satisfied by all distributions. The valid solutions can be written in explicit form as

$$\beta_2 = \beta_2^V(\beta_1) = 3\frac{16+13\beta_1+(2\beta_1+8)\sqrt{\beta_1+4)}}{32-\beta_1},\ 0\le\beta_1<32\,. \tag{4}$$

The endpoint $(\beta_1,\beta_2)=(0,3)$ corresponds to the normal distribution.

In Figure 1, the line $\beta_1=0$ also appears to be a boundary, but this is not really the case, as $\beta_1=0$ merely corresponds to the symmetric case where c = 0 in (1). The negatively skew versions of PT IV correspond to negative c without any change required in the functional form of (1).

In the JSU case the key boundary in the (β_1,β_2) plane, which divides the SU from the SB subfamily, is that corresponding to the lognormal distribution with PDF:

$$f^\Lambda(y) = \frac{1}{\sqrt{2\pi}\sigma(y-\theta)}\exp\left(-\frac{(\log(y-\theta)-\mu)^2}{2\sigma^2}\right). \tag{5}$$

This boundary, called the lognormal line is also shown in Figure 1. Like the PT V line it can also be plotted in explicit form with

$$\beta_2 = \beta_2^\Lambda(\beta_1) = \omega^4(\beta_1)+2\omega^3(\beta_1)+3\omega^2(\beta_1)-3\,, \tag{6a}$$

where

$$\omega(\beta_1) = -\frac{1}{2}[8+4\beta_1+4(4\beta_1+\beta_1^2)^{1/2}]^{1/3}+2[8+4\beta_1+4(4\beta_1+\beta_1^2)^{1/2}]^{-1/3}-1\,. \tag{6b}$$

This follows directly from known expressions for β_1 and β_2, see Johnson, Kotz and Balakrishnan (1994, equations 14.9a and 14.9b). From the form of (6a) and (6b) it is clear that $\beta_2^\Lambda(\beta_1)$ is finite for finite $\beta_1\ge 0$. The lognormal line lies entirely above the PT V line except for the endpoint $(\beta_1,\beta_2)=(0,3)$ which again corresponds to the normal distribution.

In the next Section we consider the problems of fitting the PT IV and JSU distributions, but allow for the possibility that the best fit is obtained on the PT V and lognormal boundary lines.

2 FITTING THE PT IV AND JSU DISTRIBUTIONS

2.1 Fitting the JSU Distribution

We discuss estimating the JSU distribution first, as estimation methods for this are more established. The main methods proposed in the literature are the (i) method of moments, and (ii) quantile methods. The method of moments is discussed in detail by Elderton and Johnson (1969) and a FORTRAN implementation, actually for the full Johnson family, is given by Hill, Hill, and Holder (1976). Ingenious quantile methods are described by Slifker and Shapiro (1980) and Wheeler (1980). The ML estimation method is often mentioned but a full algorithmic description does not seem available.

We shall use the ML method for estimating parameters. The method has well documented general advantages, for example it is well known, in general, to be more efficient and less affect by outliers than the method of moments. The ML method is actually very straightforward, if a numerical optimization procedure, such as the Nelder-Mead simplex algorithm (our preferred method), is used to maximize the loglikelihood. Though there is little theoretical work on the Nelder-Mead, it nevertheless appears to be quite robust in practice and seems widely used. It is uncomplicate to implement, and in particular it is easily adapted to handle restrictions on the range of values allowed for certain parameters, such as positivity conditions.

The Nelder-Mead is a search method that requires initial parameter values to be provided by the user. The choice of initial parameter values is the only major issue in the problems we consider. The method of moments is often suggested, see Hill, Hill, and Holder (1976) for example, as a good starting point for the application of ML estimation. For the JSU distribution, the squared skewness β_1 and the kurtosis β_2 depend only on the c and δ parameters in (2). Inversion allows c and δ to be treated as functions, $c = c(\beta_1, \beta_2)$ and $\delta = \delta(\beta_1, \beta_2)$, of β_1 and β_2. The moment estimators are simply

$$\tilde{c} = c(\tilde{\beta}_1, \tilde{\beta}_2) \text{ and } \tilde{\delta} = \delta(\tilde{\beta}_1, \tilde{\beta}_2) \tag{7}$$

where $\tilde{\beta}_1$ and $\tilde{\beta}_2$ are the *sample* values obtained from the data sample. However there is a clear problem here. If these are to be used as the initial values, then the sample point $(\tilde{\beta}_1, \tilde{\beta}_2)$ *has* to lie in the SU region, depicted in Figure 1. The example sample point shown in Figure 1 lies in the SB region. In this case the formula (7) is not meaningful and fails.

One might think that fitting an SU distribution to a sample whose $(\tilde{\beta}_1, \tilde{\beta}_2)$ falls in the SB region is therefore inappropriate. However this is not necessarily the case. It is known, especially for small samples, that the higher sample moments are very variable. Thus it is perfectly possible for the sample $(\tilde{\beta}_1, \tilde{\beta}_2)$ point not to lie in the region from which the sample was truly drawn.

Our solution takes advantage of the tolerance that the Nelder-Mead algorithm has over its starting point. Our proposed starting point is as follows.

JSU Starting point:

If $(\tilde{\beta}_1, \tilde{\beta}_2)$ lies in the JSU region then use $\tilde{c}$ and $\tilde{\delta}$ as in (7). If $(\tilde{\beta}_1, \tilde{\beta}_2)$ lies in the JSB region then *still* use (7) but with $\tilde{\beta}_2$ replaced by

$$\tilde{\beta}_2^* = 1.1\beta_2^{\Lambda}(\tilde{\beta}_1), \tag{8}$$

where $\beta_2^L(\tilde{\beta}_1)$ is the point on the lognormal line (6) corresponding to $\tilde{\beta}_1$, the observed sample β_1 value.

The starting values, $\tilde{\xi}$ and $\tilde{\lambda}$, of the other two parameters in (2) are found by matching the first two moments of the distribution with the sample mean and sample variance. The formulas are given in Hill, Hill, and Holder (1976).

The overall starting point, by construction, matches the first three moments. The condition (8), contains an inflation factor with a somewhat arbitrary value of 1.1, but which ensures that the modified starting point $(\tilde{\beta}_1, \tilde{\beta}_2^*)$ lies firmly in the SU region.

With the initial point determined, the only other thing to ensure is that, in its search to maximize the loglikelihood, each point selected by the Nelder-Mead algorithm has meaningful parameter values, or values which do not cause numerical instability. The parameter *c* in (2) is a shape parameter to which upper and lower bounds, $\pm S$, can be applied. The parameters δ and ξ are both scale parameters which can be restricted to being greater than some small $\varepsilon > 0$. At each iteration each parameters can be tested against each bound(s) it is to satisfy and reset to value of the bound if the Nelder-Mead has selected a value that breaks the bound.

2.2 Fitting the PT IV distribution

Consider now the PT IV case. Heinrich (2004) gives a clear summary of salient properties of this distribution especially with regard to numerical work. In particular Heinrich (2004) provides a simple product formula for calculating the normalizing constant in the PDF of the PT IV. This formula has been known since Pearson (1895). Nagahara (1999) obtains an alternative form which seems less convenient. A clear discussion of many of the problems that occur when fitting this distribution is provided by Heinrich (2004), who recommends use of the method of moments for estimating the distribution. Parrish (1983) discusses the more general problem of fitting a Pearson distribution from the full family by maximizing the likelihood, and other criteria, treating this as a function of the moments. This latter approach is presumably motivated by the fact that the (β_1, β_2) plot can be separated into disjoint regions corresponding to different Pearson types. Thus when numerically maximizing the likelihood the Pearson type can be selected to match each (β_1, β_2) point at which the likelihood is calculated.

We do not regard these suggestions of Heinrich and Parrish as satisfactory.

The Heinrich suggestion of using the method of moments for estimating the PT IV distribution is not satisfactory because there is no guarantee that the sample value $(\tilde{\beta}_1, \tilde{\beta}_2)$ will fall in the PT IV region. The approach suggested by Parrish (1983) is not satisfactory as the Pearson types IV, V and VI all include cases where some or all moments do not exist. These cases are not represented in the (β_1, β_2). The approach suggested by Parrish will therefore fail to correctly estimate a PT IV, for example, if it is one where β_1 and β_2 are not both finite.

Our ML approach is standard in that the likelihood is defined in terms of the parameters of (1). We shall also use the moment method to obtain starting parameter values, but only when the sample value $(\tilde{\beta}_1, \tilde{\beta}_2)$ falls in the type IV region. When it does not we shall match the parameters to the first *three* sample moments subject to the condition

$$\frac{bc}{b-1} = \tilde{\beta}_1 . \tag{9}$$

Using the formulas for the first three moments, given by Nagahara (1999) or Heinrich (2004) for example, this yields the starting estimates as:

$$\tilde{b}(\tilde{\gamma}) = 2 + \frac{2}{1+\tilde{\gamma}^2}(2 + \sqrt{5+\tilde{\gamma}^2}) \ , \ \ \tilde{c} = g(\tilde{\gamma})\tilde{\gamma} \ , \ \ \tilde{\tau} = t(\tilde{\gamma})s \ , \ \ \tilde{\mu} = \bar{y} - m(\tilde{\gamma})s \tag{10a}$$

where

$$\tilde{\gamma} = \sqrt{\tilde{\beta}_1}\ ,\ g(\gamma) = \frac{(b(\gamma)-1)}{b(\gamma)},\ \ t(\gamma) = \frac{(b(\gamma)-2)}{b(\gamma)},\ m(\gamma) = \frac{1}{2}\gamma(b(\gamma)-2)\,. \tag{10b}$$

The condition (9) is to a certain extent arbitrary, but has been chosen to ensure that the initial estimates (10a) do not behave unusually. In fact both $\tilde{b}$ and the multiplying factors *g*, *t* and *m* remain bounded and very stable as γ varies from 0 to infinity. The behaviour of the Nelder-Mead optimization seems very robust with these starting values, and we have been able to use them even when $(\tilde{\beta}_1, \tilde{\beta}_2)$ falls in the type IV region.

As with the JSU distribution, numerical stability in the Nelder-Mead can be maintained by imposing appropriate bounds on parameters, to be satisfied at each iteration of the Nelder-Mead search. The parameter *b* in (1) is a shape parameter. Though in principle it needs only to satisfy $b > 0$, the distribution does not possess *any* finite moments for $b \le 1$. The normalizing constant becomes very unstable to calculate for *b* much less than 0.5. We have therefore chosen to restrict the allowed range to $b \ge 0.5 + \varepsilon$ where ε is a small fixed positive quantity. We also impose an upper bound $b \le S$, with *S* large. The parameter τ is a scale parameter to which we impose the condition $\tau \ge \varepsilon > 0$, with ε a small fixed quantity.

3 BOUNDARY MODELS

When fitting the PT IV (respectively JSU) full model we allow the possibility that the best model corresponds to the PT V (respectively lognormal) boundary model. We can do this in two ways.

One way is to try to fit the full model directly. We then need to know if the Nelder-Mead search is tending towards the boundary model. We therefore need to know how the parameters of the PDF of the full model would behave in this situation.

The other way is to fit the boundary model first, and then have some criterion which determines whether it would be worthwhile extending the fitting process to consideration of the full model.

We consider both approaches for each of the distributions. All we need to do is examine the loglikeliood.

3.1 PT IV Model: Type V Boundary

For the PT IV distribution (1), the loglikelihood can be expanded as a power series in $\alpha = c^{-1}$. We have

$$L^{IV}(b,c,\mu,\tau \mid y) = L^{V}(\beta,\xi,\lambda \mid y) + M^{V}(\beta,\xi,\lambda \mid y)\alpha^2 + O(\alpha^3) \tag{11a}$$

where $L^{V}(\beta,\xi,\lambda \mid y)$ is the loglikelihood of the PT V distribution with PDF (3) and

$$M^{V}(\beta,\xi,\lambda \mid y) = \frac{1}{3(\beta+1)} + \frac{1}{6}(\beta+1) - \frac{1}{2} + \sum_{i=1}^{n}\left(\frac{\lambda^3}{3(\beta+1)^2(y_i-\xi)^3} - \frac{\lambda^2}{2(\beta+1)(y_i-\xi)^2}\right), \tag{11b}$$

where

$$\alpha = c^{-1},\ \beta = 2b-1,\ \xi = \mu \text{ and } \lambda = (\beta+1)c\tau\,. \tag{11c}$$

This shows that if

$$c \to \infty \text{ and } \tau \to 0\,, \text{ with } c\tau \to \lambda/(\beta+1),\ b \to (\beta+1)/2 \text{ and } \mu \to \xi\,, \tag{11d}$$

with β, ξ and λ remaining fixed, then the PT IV distribution (1) tends to that of the PT V distribution (3).

Moreover if $M^V(\cdot) > 0$ then, at least locally, $L^{IV}(\cdot)$ will increase as α increases away from zero. This shows that if we fit the PT V boundary model *first*, and find that $M^V(\cdot) > 0$, then there are PT IV models better than the best PT V. In this case it is worth fitting the full PT IV model.

Alternatively we can fit the full PT IV first, but if we find in the Nelder-Mead search that the parameters are behaving as in (11d), then this would be an indication that the PT V should be tried. But this seems less easy to do than trying the boundary model first.

3.2 JSU Model: Lognormal Boundary

An analogous analysis applies to the JSU distribution. Its loglikelihood can be expanded as a power series in λ. We have

$$L^{SU}(c,\delta,\xi,\lambda \mid y) = L^{\Lambda}(\theta,\mu,\sigma \mid y) + M^{\Lambda}(\theta,\mu,\sigma \mid y)\lambda + O(\lambda^2) \tag{12a}$$

where $L^{\Lambda}(\beta,\xi,\lambda \mid y)$ is the loglikelihood of the lognormal distribution with PDF (5) and

$$M^{\Lambda}(\theta,\mu,\sigma \mid y) = -\sum_{i=1}^{n}\left(\frac{\log(y_i-\theta)-\mu}{4\sigma^2(y_i-\xi)^2}\right), \tag{12b}$$

where

$$\theta = \xi,\ \sigma = \delta^{-1} \text{ and } \mu = \log\lambda - c\delta^{-1} - \log 2. \tag{12c}$$

This shows that if

$$c \to -\infty \text{ and } \lambda \to 0, \text{ with } \log\lambda - c\delta^{-1} - \log 2 \to \mu,\ \delta^{-1} \to \sigma, \text{ and } \xi \to \theta, \tag{12d}$$

with μ, σ and θ remaining fixed, then the JSU distribution (2) tends to that of the lognormal distribution (5).

Moreover if $M^{\Lambda}(\cdot) > 0$ then, at least locally, $L^{SU}(\cdot)$ will increase as λ increases away from zero. This shows that if we fit the lognormal boundary model *first*, and find that $M^{\Lambda}(\cdot) > 0$, then there are JSU models better than the best lognormal. In this case it is worth fitting the full JSU model.

Alternatively we can fit the full JSU model first, but if we find in the Nelder-Mead search that the parameters are behaving as in (12d), then this would be an indication that the JSU should be tried. As in the PT IV case, this seems less easy to do than trying the boundary model first.

4 GOODNESS OF FIT

Once a model has been fitted, an immediate question is whether the model is a good fit. Of the many goodness-of-fit statistics available the Anderson-Darling statistic

$$A^2 = -\{\sum_{i=1}^{n}(2i-1)[\log(F(y_i;\hat{\boldsymbol{\theta}}) + \log(1-F(y_i;\hat{\boldsymbol{\theta}})]\}/n - n \tag{13}$$

is widely regarded as one of the best. In (13) we have taken the formula as given in Stephens (1974) explicitly in the form required to test the fitted model so that $F(y_i;\hat{\boldsymbol{\theta}})$ is the CDF of the fitted model evaluated at the observation y_i, with the parameters set to their estimated ML values.

There is one important issue. Stephens (1974) points out that the null distribution of A^2, when parameters have been estimated, depends on the distribution being fitted and *also* on which parameters have been estimated. For example, for the normal model the 10% percentage point changes from 1.933, when both mean and SD are known, to 0.656 when both are estimated by maximum likelihood.

D'Agostino and Stephens (1986) use Monte-Carlo simulation to produce tables of critical points for various distributions, but not for the PT IV or JSU distributions. However Monte-Carlo simulation is readily incorporated in our estimation problem. Once a model is fitted, and this applies to all the models that we consider, we can estimate the null distribution of A^2 by the following procedure.

(i) Generate *B* samples of size *n from the fitted distribution* $F(\cdot;\hat{\boldsymbol{\theta}})$:

$$\mathbf{y}^{*(j)} = \{y_1^{*(j)}, y_2^{*(j)}, ...y_n^{*(j)}\},\ \ j = 1,2,...,B\,.$$

(ii) Fit the *same* model $F(\cdot;\boldsymbol{\theta})$ to each sample. Write these as $F(\cdot;\hat{\boldsymbol{\theta}}^{*(j)})$, j = 1,2,..., *B*.

(iii) Calculate $A^{*(j)2}$ from (13), with $F(\cdot;\hat{\boldsymbol{\theta}}^{*(j)})$ in place of $F(\cdot;\hat{\boldsymbol{\theta}})$ and $\mathbf{y}^{*(j)}$ in place of the original sample $\mathbf{y}$.

(iv) Use the EDF of the $A^{*(j)2}$, j = 1,2,..., *B* to estimate the required null distribution of A^2.

All that is required is a method for generating variates from the fitted distribution.

Generators are well known for the PT V, JSU and lognormal cases. For the PT IV distribution of (1), a neat method is given by Devroye (1986), with an explicit implementation given by Heinrich (2004), provided the parameter $b > 1$. This condition is the same as that ensuring the mean of the distribution is finite. With this proviso, the goodness of fit method just described is implemented in the Excel workbook accompanying this paper, for all the models considered in this paper.

Stephens (1974) used a value of *B* = 10,000. In our examples, which are for illustration only, we used *B* = 1,000. In exploratory work, *B* = 100 is usually quite sufficient to provide a clear indication of whether a fitted model is adequate or not.

5 NUMERICAL EXAMPLES

We give two examples. We have selected two extreme cases to illustrate the robustness of our proposed fitting methods. Both data sets are included in the accompanying Excel workbook.

5.1 UK Stock Exchange FTSE Data Set

The first comes from a financial application. In a study of the movement of the stock market (with a view to generating similar data for use in a simulation) we consider fitting the PT IV model to a set of data of the form:

$$y_i = \log(p_i / p_{i-1})\,,\ i = 1,2,\ ....,\ n$$

where p_i is the closing FTSE100 index on day *i*. The data set comprises *n* = 250 observations, with the last day observed being 17 March 2011. There is the possibility of correlation between succeeding observations, but the lag-one autocorrelation is fairly small at 0.016. As the example is for illustration only, we have therefore treated it as a random sample.

For this data set $(\tilde{\beta}_1, \tilde{\beta}_2)$ = (0.023, 4.84) placing the point well in the PT IV region of Figure 1. We fitted the PT IV distribution to this data set and obtained the ML estimates of $\hat{b} = 3.03$ $\hat{c} = -0.0261$ $\hat{\mu} = 0.000766$ and $\hat{\tau} = 0.0196$. Figure 2 shows the fitted CDF and PDF.

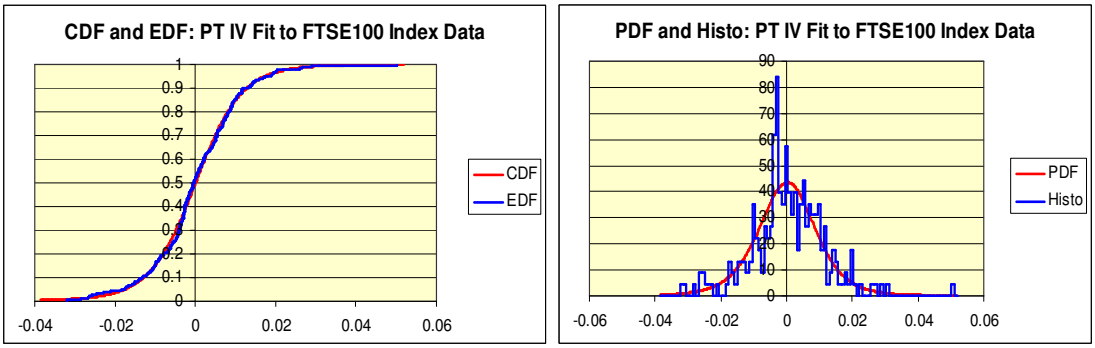

Figure 2: CDF and PDF of the PT IV distribution fitted to the FTSE financial data

We used the resampling method of Section 4 to calculate, under the null hypothesis, the EDF of A^2 of, the Anderson-Darling goodness of fit statistic. This gave A^2 = 0.333. This corresponds to a p-value, obtained from the null EDF, of 0.16, so that we would not reject the fit at the 10% level. The 10% critical point obtained from the EDF was 0.372. The EDF is shown in Figure 3 to illustrate the typical form that the EDF takes.

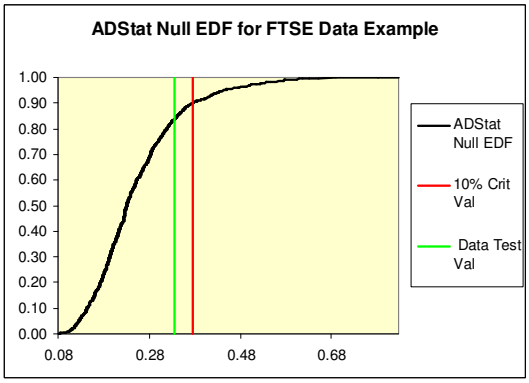

Figure 3: Null Distribution of the A^2 statistic for the PT IV model fitted to the FTSE data set

For the boundary PT V model given in (3), the parameter estimates were $\hat{\xi} = -0.103$, $\hat{\beta} = 86.4$ and $\hat{\lambda} = 8.89$. The large $\hat{\beta}$ indicates that the model is near normal. The fitted CDF and PDFs are shown in Figure 4. The goodness of fit test gave A^2 = 1.342 with a 10% critical value of 1.341 that was almost identical, indicating that the fit is very marginal at the 10% critical level. This agrees with the value of the gradient factor of (11b), which was M^V = 0.098 > 0 indicating that the PT IV distribution would provide a better fit.

As a further comparison we also fitted the normal distribution to the data giving ML estimates of $\hat{\mu} = 0.00$ and $\hat{\sigma} = 0.011$. We have not shown the fitted CDF and PDF graphically as the plots are visually identical to those for the PT V fit given in Figure 4. The goodness of fit test statistic value of A^2 = 1.342 was effectively the same as in the PT V case. However the 10% critical point (obtained under the null assumption of a normal model, was 0.629 using B = 1000. A more refined experiment with B = 10,000 gave the 10% point as 0.635.

The normal model with both mean and SD fitted is one of the cases for which there are published tabulated critical values. Our values are close to the tabulated value of 0.632 for the 10% critical value in Table 6.18 of Law (2007).

Our results show that the normal model is not a very good fit. Inspection of the plots show that the more sharply pointed (i.e. leptokurtic) fit of the PT IV distribution better captures the form of the data. As pointed out in Stuart and Ord (1987, Section 3.32) this could be an indication that the data is more fat tailed than the normal.

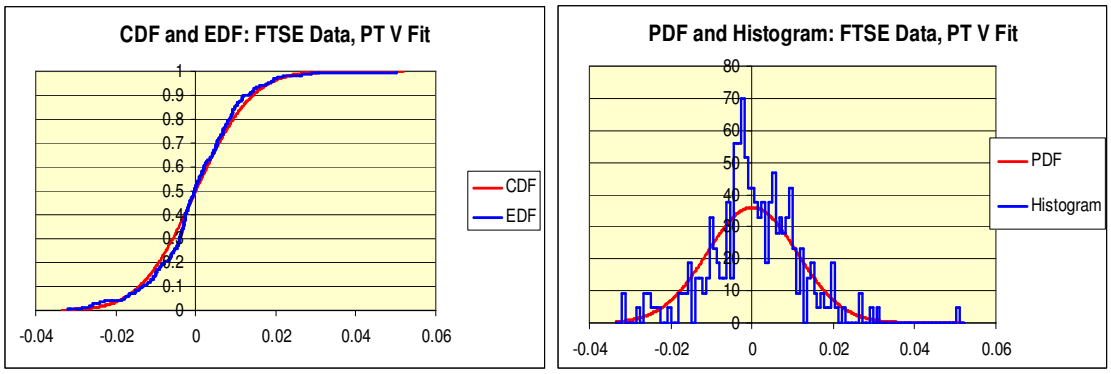

Figure 4: CDF and PDF of PT V distribution fitted to the FTSE financial data

5.2 Hospital Length of Stay Data

The second data set is a sample of 1073 actual observations of hospital lengths of stay (in days). These were used in a simulation study investigating hospital bed allocation policies. This is a very positively skewed sample. The point $(\tilde{\beta}_1, \tilde{\beta}_2)$ = (30.6, 51.4) is the sample point shown in Figure 1. In the actual study a gamma distribution was tried initially but not found satisfactory. We first verify this finding, by fitting the gamma distribution. (The attached Excel workbook has the option of fitting the gamma model.)

One immediate issue concerns the left tail of the distribution. As the lengths of stay are recorded to the nearest day starting with one day's stay it seems reasonable make the left tail behavior explicit and to use the shifted gamma model with PDF

$$f^G(y) = \Gamma^{-1}(\alpha)\beta^{-\alpha} y^{\alpha-1} \exp[-(y-\theta)/\beta]$$

where the threshold value θ is set to $\theta = 0.5$. This allows us to illustrate one advantage of our proposed goodness of fit procedure, which is that calculation of the critical values of the A^2 will automatically handle situations where certain parameter values are fixed.

The ML estimates for the gamma model were $\hat{\theta} = 0.5$ $\hat{\alpha} = 0.718$, $\hat{\beta} = 6.09$. The fitted CDF and EDF are depicted in Figure 5. The goodness of fit test gave a value of A^2 = 2629, with a p-value not measurably different from zero. The 10% critical value was 1776.

These results confirm the finding in the initial study of the unsatisfactory nature of the gamma model for this particular data set.

We now consider fitting the Johnson SU distribution to the data. When the JSU distribution of (2) is fitted to very positively skew data the parameter ξ effectively becomes the left threshold. We therefore held this parameter fixed at $\xi = 0.5$.

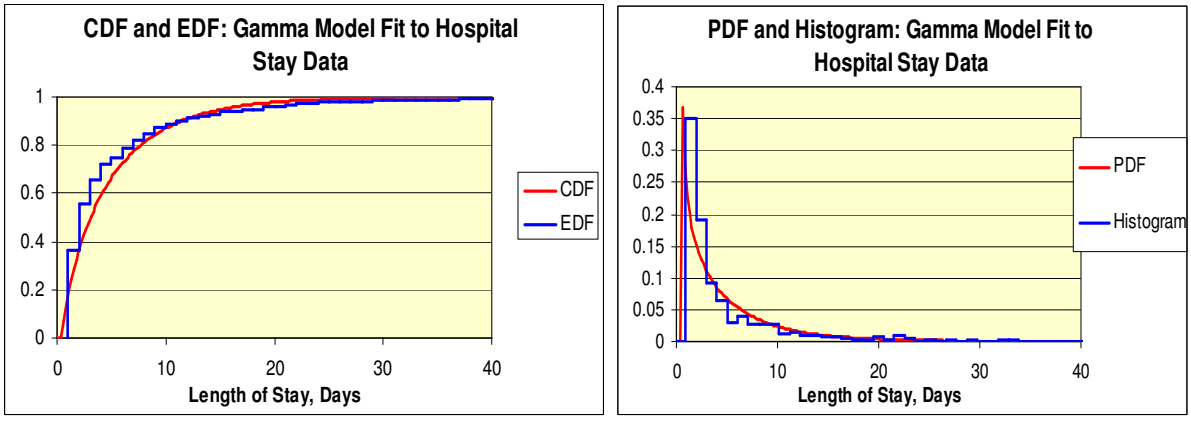

Figure 5: CDF and PDF of the gamma distribution fitted to the hospital length of data

The ML estimates of the JSU distribution were $\hat{\xi} = 0.5$, $\hat{c} = -4.26$, $\hat{\delta} = 0.806$, and $\hat{\lambda} = 0.0191$. The fitted CDF and PDF are depicted in Figure 6.

The goodness of fit statistic was A^2 =1054 with a p-value of 0.5, whilst the 10% critical point was much higher at 1212. We conclude that the JSU model was a reasonably good fit to the data set.

We also fitted the lognormal model (5) to this sample. The ML estimates were $\hat{\theta} = 0.5$, $\hat{\mu} = 0.636$ and $\hat{\sigma} = 1.24$. The goodness of fit test gave a value of A^2 = 1053, with a p-value 0f 0.489. The 10% critical point was 1202. The values are very similar to those obtained in the JSU fit, even though the gradient value of (12b) was M^{Λ}=0.314 > 0 indicating that the JSU model should be a better fit. The CDF and PDF of the fitted lognormal model was visually identical to those of the JSU model depicted in Figure 6, and are not shown here. We conclude that in this case the boundary model is quite adequate.

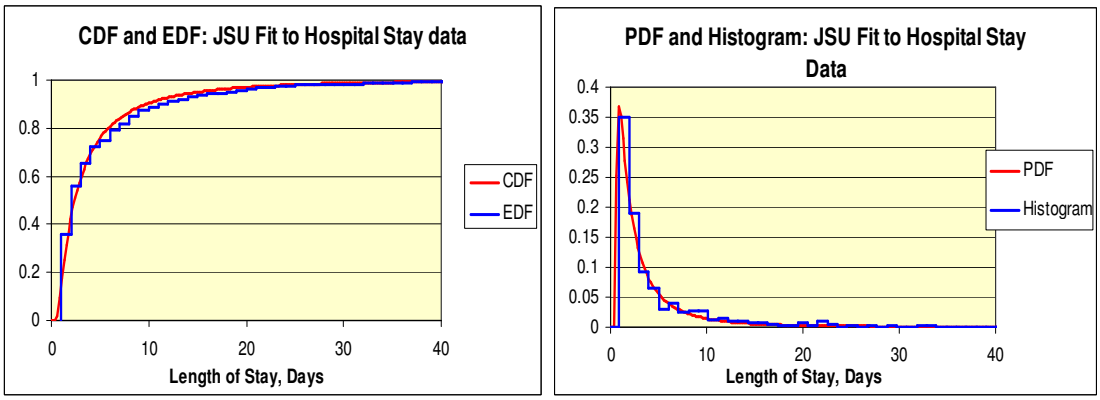

Figure 6: CDF and PDF of JSU distribution fitted to the hospital length of stay data

6 CONCLUSIONS

This paper has investigated fitting PT IV and JSU distributions by maximum likelihood where optimization of the likelihood is carried out by numerical search using the Nelder Mead algorithm. Explicit start-

ing parameter values are provided which match the first three moments of the starting distribution to the sample moments of the sample. For the particularly difficult case of the PT IV the starting values allow a model to be fitted for which moments do not exist, clearly not possible with a method like the method of moments.

The paper also points out that a goodness of fit test using the Anderson Darling statistic is straightforward if critical test values are estimated using Monte Carlo simulation.

All the procedures discussed in the paper are available as VBA code, implemented in an Excel workbook accompanying this paper. [The easiest way is to go to my personal web page and follow the link at the bottom of the first page.]

REFERENCES

D'Agostino, R. B. and M. A. Stephens. 1986. *Goodness-of-Fit Techniques*. New York: Marcel Dekker.

Devroye, L. 1986. *Non-uniform Random Variate Generation.* New York: Springer-Verlag.

Elderton, W. P. and N. L. Johnson. 1969. *Systems of Frequency Curves*. Cambridge: The University Press.

Johnson, N. L., S. Kotz, and N. Balakrishnan. 1994. *Continuous Univariate Distributions. Volume 1.* 2nd ed. New York: Wiley, Inc.

Heinrich, J. 2004. "A Guide to the Pearson type IV Distribution." *CDF/MEMO/STATISTICS/PUBLIC /6820*. http://www-cdf.fnal.gov/physics/statistics/notes/cdf6820_pearson4.pdf

Hill, I. D., R. Hill, and R. L. Holder. 1976. "Algorithm AS 99: Fitting Johnson Curves by Moments." *J. Roy. Statist. Soc. Series C.* 25:180-189.

Law, A. M. 2007 *Simulation Modeling and Analysis*. 4th ed. Boston:McGraw-Hill.

Nagahara, Y. 1999. "The PDF and CF of Pearson type IV Distributions and the ML Estimation of the Parameters." *Statistics & Probability Letters* 43:251-264.

Parrish, R. S. 1983. "On an Integrated Approach to Member Selection and Parameter Estimation for Pearson Distributions." *Comput. Stats. & Data Anal.* 1:239-255.

Pearson, K. 1895. "Contributions to the Mathematical Theory of Evolution.—II. Skew Variation in Homogeneous Material", *Phil. Trans. Roy. Soc., A*, 186: 343-414.

Pearson, K. 1916. "Mathematical Contributions to the Theory of Evolution. XIX. Second Supplement to a Memoir on Skew Variation", *Phil. Trans. Roy. Soc., A*, *Containing Papers of a Mathematical or Physical Character*. 186: 429-457.

Slifker, J. F. and S. S. Shapiro. "The Johnson System: Selection and Parameter Estimation." *Technometrics*, 22:239-246.

Stephens, M. A. 1974. "EDF Statistics for Goodness of Fit and Some Comparisons." *J. Amer. Statist. Soc.* 69:730-737.

Stuart, A., and J. K. Ord. 1987. *Kendall's Advanced Theory of Statistics. Volume 1.* 5th ed. London: Griffin & Co. Ltd.

Wheeler, R. E. 1980. "Quantile Estimators of Johnson Curve Parameters." *Biometrika* 67:725-728.

AUTHOR BIOGRAPHY

RUSSELL C. H. CHENG is Emeritus Professor of Operational Research at the University of Southampton. He has an M.A. and the Diploma in Mathematical Statistics from Cambridge University, England. He obtained his Ph.D. from Bath University. He is a former Chairman of the U.K. Simulation Society, a Fellow of the Royal Statistical Society and Fellow of the Institute of Mathematics and Its Applications. His research interests include: design and analysis of simulation experiments and parametric estimation methods. He was a Joint Editor of the *IMA Journal of Management Mathematics*. His email and web addresses are <R.C.H.Cheng@soton.ac.uk> and <www.personal.soton.ac.uk/rchc> .