"An Empirically-Grounded Comparison of the Johnson System versus the Beta as Crop Yield Distribution Models"

by

Octavio A. Ramirez[1]

and

Tanya U. McDonald[2]

Paper Accepted for Presentation at the Annual Meeting of the
American Agricultural Economics Association
Portland, Oregon, July 29-August 1 2007

[1] Professor and Head, Department of Agricultural Economics and Agricultural Business, New Mexico State University, Box 30003, MSC 3169, Las Cruces, NM 88003-8003, e-mail: oramirez@nmsu.edu, phone: (505) 646-3215.

[2] Research Specialist, Department of Agricultural Economics and Agricultural Business, New Mexico State University.

This research was supported by the National Research Initiative of the Cooperative State Research, Education and Extension Service, USDA, Grant # 2004-35400-14194 and by the Agricultural Experiment Station of New Mexico State University.

 The authors would like to thank Bruce Sherrick and Jonathan Norvell for graciously sharing the high quality yield data from the University of Illinois Endowment Farms.

For many years, agricultural economists have recognized that the choice of an appropriate probability distribution to represent crop yields is critical for an accurate measurement of the risks associated with crop production. Recent research on this issue has been conducted by Gallagher 1987; Nelson and Preckel 1989; Moss and Shonkwiler 1993; Ramirez, Moss, and Boggess 1994; Coble, Knight, Pope, and Williams 1996; and Ramirez 1997; among several others. This research has provided statistical evidence of non-normality and heteroskedasticity in crop-yield distributions, specifically of the existence of positive kurtosis and negative skewness in most cases.

Gallagher (1987) used the well-known Gamma density as a parametric model for the distribution of soybean yields. Nelson and Preckel (1989) proposed a conditional Beta distribution to model corn yields. Taylor (1990) estimated multivariate non-normal densities using a conditional distribution approach based on the hyperbolic tangent transformation. Ramirez (1997) introduced a modified inverse hyperbolic sine (IHS) transformation (also known as the S_U family) as a possible non-normal, heteroskedastic multivariate probability distribution model. Ker and Coble (2003) proposed a semi-parametric model based on the Normal and the Beta densities to represent crop yields.

The three major statistical approaches that have been used for the modeling and simulation of crop yield distributions, namely the parametric, non-parametric and semi-parametric methods, all have distinct advantages and disadvantages. The parametric method is based on assuming that the stochastic behavior of the underlying the variable of interest can be adequately represented by a particular parametric probability distribution function. For this reason, its main weakness is the potential error resulting from assuming a probability distribution that is not flexible enough to properly represent

the yield data. The main advantage of the parametric method is that it performs relatively well in small sample applications. The leading distributions that have been used as a basis for this method are the Beta and the Gamma (Norwood, Roberts, and Lusk 2004).

Most recently, however, Ramirez and McDonald (2006a) introduced an expanded form of the Johnson system, which is composed of the S_U, S_B and S_L distributions. They hypothesize that, because their expanded Johnson system can accommodate all mean-variance-skewness-kurtosis (MVSK) combinations that may be theoretically exhibited by a random variable, it should provide for a reasonably accurate modeling of any probability distribution that might be encountered in practice. This would clearly address the main disadvantage of parametric models cited in the literature, i.e. their lack of flexibility and the resulting specification error risk, and provide for a system that supersedes all other densities that have been considered as a basis for these models.

This hypothesis, however, has not been empirically tested. In fact, because the precise stochastic behavior of a random variable (i.e. the exact shape of its density function) is characterized by an infinite number of central moments, it is possible that accommodating its first four moments does not provide for a sufficiently accurate representation of the variable's probabilistic behavior. The validity of Ramirez and McDonald (2006a) hypothesis is explored in this article.

The Expanded Johnson System

Unlike other frequently assumed distributions such as the Beta and the Gamma, the original Johnson system exhibits the key property of being able to accommodate any theoretically feasible skewness-kurtosis combination (figure 1), although each of those combinations is inherently associated with a fixed set of mean-variance values. Ramirez

and McDonald (2006a) developed an expanded parameterization of the Johnson system that can accommodate the same skewness-kutosis (S-K) combinations allowed by the original system in conjunction with any mean and variance.

Figure 1 illustrates the different skewness-kurtosis regions covered by each of the three families in the Johnson system, as well as by the Beta and the Gamma distributions. Note again that any theoretically feasible S-K combination can be accommodated by one of the three families in this system. In fact, just the S_U and S_B are sufficient for this purpose, as the S_L only spans the curvilinear boundary between the S_U and S_B. The lower bound of the S_B distribution is given by $K = S^2 - 2$, which is also the upper bound for the theoretically impossible S-K region.

In contrast, note that the Gamma distribution only spans a curvilinear segment on the upper right quadrant of the S-K plane. Although, as the S_L, the Gamma distribution can be adapted to cover the mirror image of this segment on the upper left quadrant, the combinations of S-K values allowed by it are still very limited. Also note that the Gamma segment is the upper boundary of the S-K area covered by the Beta distribution. Although the Beta covers a significant area of the S-K plane, the S_B can accommodate all S-K combinations allowed by the Beta while the Beta only covers a subset of the S-K area spanned by the S_B.

In addition to their limited coverage of the S-K plane, the Gamma and the Beta exhibit the same handicap of the original Johnson system. That is, because they are two-parameter distributions, any particular S-K combination is always arbitrarily associated with a specific set of mean and variance values.

Estimation of the Expanded Johnson System

Estimation of the expanded Johnson system can be accomplished by maximum likelihood procedures. The log-likelihood functions to be maximized in order to estimate the parameters of each of the three distributions in the system (S_U, S_B and S_L) are (Ramirez and McDonald 2006a):

(1) $$\sum_{t=1}^{T} Ln\{P(\mathrm{Y}_t^F)\} = 0.5\sum_{t=1}^{T}\ln(G_t) - 0.5\delta^2\sum_{t=1}^{T} H_t^2 \;;$$

where:

(2) $$G_t = \frac{\delta^2 G_{SU}}{2\pi(Z_t\sigma)^2(1+R_{SU\,t}^2)},$$

$$H_t = \ln[R_{SU\,t} + \sqrt{1+R_{SU\,t}^2}] + \frac{\gamma}{\delta} = \sinh^{-1}(R_{SU\,t}) + \frac{\gamma}{\delta},$$

$$R_{SU\,t} = \frac{(\mathrm{Y}_t^F - \mathrm{X}_t\beta)\times G_{SU}^{\,1/2}}{Z_t\sigma} + F_{SU},$$

(3) $$G_t = \frac{\delta^2 G_{SL}}{2\pi(Z_t\sigma)^2 R_{SL\,t}^2},$$

$$H_t = \ln[R_{SL\,t}] + \frac{\gamma}{\delta},$$

$$R_{SL\,t} = \frac{(\mathrm{Y}_t^F - \mathrm{X}_t\beta)\times G_{SL}^{\,1/2}}{Z_t\sigma} + F_{SL},$$

(4) $$G_t = \frac{\delta^2 G_{SB}}{2\pi(Z_t\sigma)^2\ R_{SB\,t}^2(1-R_{SB\,t})^2},$$

$$H_t = \ln[R_{SB\,t}/(1-R_{SB\,t})] + \frac{\gamma}{\delta},$$

$$R_{SB\,t} = \frac{(\mathrm{Y}_t^F - \mathrm{X}_t\beta)\times G_{SB}^{\,1/2}}{Z_t\sigma} + F_{SB};$$

for the S_U, S_L and S_B distributions, respectively; $G_t > 0$; and G_{SU}, F_{SU}, G_{SL}, F_{SLt}, G_{SB} and F_{SBt} are the exponential and trigonometric functions the shape parameters γ and δ. Ramirez and McDonald (2006a) also show that:

(5) $E[Y_t^F] = \mathrm{M}_t = X_t\beta$, and

$V[Y_t^F] = \sigma_t^2 = (Z_t\sigma)^2$;

where X_t and Z_t represent vectors of explanatory variables believed to affect the means and variances of the distributions, and β and σ are conformable parameter vectors. In short, the mean and the variance of the S_U, S_B and S_L random variables ($Y^F{}_t$) can be independently controlled by $\mathrm{M}_t = X_t\beta$ and $(Z_t\sigma)^2$ while the shape parameters γ and δ separately determine the distribution's skewness and kurtosis.

A final adjustment that facilitates estimation and interpretation is re-defining these distributional shape parameters as follows: for the S_U γ=-μ, for the $S_{\mathbf{B}}$ γ=μ, and for all three families δ=1/θ. Also in the case of the $S_{\mathbf{L}}$, after re-parameterization, γ becomes a redundant coefficient and, thus has to be set to zero. Then, for both the S_U and the S_B μ<0, μ=0 and μ>0 are associated with negative, zero and positive skewness, respectively, and all three families approach a normal distribution as θ goes to zero. This also allows for testing the null hypothesis of normality as Ho: θ=μ=0.

The Expanded Beta Distribution

An expanded parameterization of the Beta distribution that can accommodate any mean and variance in conjunction with all skewness-kurtosis combinations allowed by the original Beta is needed for the purposes of this research. This expanded Beta distribution is obtained by applying the procedure outlined by Ramirez and McDonald (2006b). With

this procedure, any two-parameter non-normal distribution pdf(y)=f(y,δ,λ) with mean $E[y]=f_1(\delta,\lambda)$, variance $E[(y-E[y])^2]=V[y]=f_2(\delta,\lambda)$, skewness $E[(y-E[y])^3/f_2(\delta,\lambda)^{3/2}]=S[y]=f_3(\delta,\lambda)\neq 0$, and kurtosis $E[(y-E[y])^4/f_2(\delta,\lambda)^2]=K[y]=f_4(\delta,\lambda)\neq 0$, can be expanded as follows:

(6) $y' = \{y- f_1(\delta,\lambda)\}/f_2(\delta,\lambda)^{1/2}$

yields a pdf {pdf'(y')} with a constant mean (E[y']=0) and variance (V[y']=1) without altering its skewness and kurtosis coefficients. Then,

(7) $y'' = \sigma y'+\mu$

yields an expanded, more flexible, pdf {pdf''(y'')=f''(y'',μ,σ,δ,λ)} which mean and variance are solely determined by μ and σ^2, respectively (i.e. E[y'']=μ and $V[y'']=\sigma^2$), while its skewness and kurtosis coefficients depend on the original distributional shape parameters (δ and λ) only. As in the case of the Johnson system, the mean and the variance can be specified as linear functions of relevant explanatory variables:

(8) $\mu_t = X_t\beta$, and $\sigma_t = Z_t\boldsymbol{\sigma}$,

where X_i and Z_i are the explanatory variable matrices, and β and $\boldsymbol{\sigma}$ are parameter vectors.

In the case of the Beta distribution:

(9) $f_1(\delta,\lambda) = F_B = \delta/(\delta+\lambda)$

$$f_2(\delta,\lambda) = G_B = \frac{\delta\lambda}{(\delta+\lambda+1)(\delta+\lambda)^2}$$

Thus, the transformation from the original Beta distributed variable (y) into the random variable exhibiting the expanded Beta distribution (y'') is:

(10) $y'' = \sigma_t(y - F_B)/G^{1/2}{}_B + \mu_t$.

The probability density function for y" is obtained through a straightforward application of the transformation technique (Mood, Graybill, and Boes 1974), which leads to the following log-likelihood function:

$$LL_B = \sum_{t=1}^{T} \ln \left| \frac{\sqrt{G_B}}{\sigma_t^2} \right| + n \ln \Gamma(\delta + \lambda) - n \ln \Gamma(\delta) - n \ln \Gamma(\lambda) + (\delta - 1) \times \left(\sum_{t=1}^{T} \ln P_t \right) + (\lambda - 1) \left(\sum_{t=1}^{T} \ln (1 - P_t) \right) \quad (11)$$

where

$P_t = \frac{\delta}{\delta + \lambda} + \frac{(Y_t - \mu_t) \times \sqrt{G_B}}{\sigma_t^2}$, Γ represents the Gamma function, $\mu_t = X_t\beta$ and $\sigma_t = Z_t\boldsymbol{\sigma}$.

This log-likelihood function is maximized with respect to $\boldsymbol{\delta}$, $\boldsymbol{\lambda}$, β and $\boldsymbol{\sigma}$ in order to obtain estimates for these parameters and parameter vectors.

Evaluating the Flexibility of the Johnson System

Ramirez and McDonald (2006a) hypothesis is that, because their expanded Johnson system can accommodate all MVSK combinations that may be exhibited by a random variable, it should provide for a reasonably accurate modeling of any probability distribution that might be encountered in practice and thus supersede all other densities which have been considered as the basis for parametric models. Of the two distributions that have been widely used for the probabilistic modeling of crop yields, namely the Gamma and the Beta, the latter is clearly the most flexible as it can accommodate a much wider range of skewness-kurtosis combinations (figure 1). Also, theoretically, the Beta distribution is not related to the Johnson system. Therefore, the Beta distribution is selected for use in this comparative evaluation of the Johnson system.

The yield data used by Ramirez and McDonald (2006a) to introduce and illustrate applications of the expanded Johnson system is chosen as a basis for the evaluation. The data, obtained from the University of Illinois Endowment Farms database, includes 26 corn farms located in twelve counties across that State. Data are available from 1959 to 2003, with the sample size varying from 20 to 45.

Gauss 6.0 Constrained Maximum Likelihood (CML) programs are used to estimate the parameters of yield models based on the expanded Beta distribution for each of these 26 farms. The means, variances, skewness and kurtosis coefficients implied by the models are computed through those programs as well (all programs are available from the authors upon request). As in Ramirez and McDonald (2006a), the means and standard deviations are specified as second and first degree polynomial functions of time:

(12) $M_t = X_t\beta = \beta_0 + \beta_1 t + \beta_2 t^2$, and

$\sigma_t = (Z_t\sigma) = \sigma_0 + \sigma_1 t$; t=1,…,T.

As in the case of the S_U and S_B models estimated by Ramirez and McDonald, the Beta models initially include seven parameters (β_0, β_1, β_2, σ_0, σ_1, θ and μ). Select statistics about the estimated Beta models, as well as the S_U, S_B and normal models (from Ramirez and McDonald 2006a), are presented in table 1. The S_L is excluded from the comparison on the basis of Ramirez and McDonald's finding that the S_L models are always outperformed by the S_U and the S_B in this particular application. This is expected since Corn Belt corn yields have been previously found to be left-skewed (Nelson and Preckel 1989; Taylor 1990; Ramirez 1997; Ker and Coble 2003; Harri, Coble, Erdem, and Knight 2006) and the S_L family only allows for positive skewness (figure 1).

Likelihood ratio (LR) tests reject the null hypothesis of normality (Ho: $\theta=\mu=0$; $\chi^2_{2,0.1}=4.61$) in 17 of the 26 S_U models and in 18 of the 26 S_B and Beta models ($\alpha=0.10$). Normality is rejected by both the S_B and Beta models in 17 cases and not rejected by either in seven cases. There is only one instance in which normality is marginally rejected by the Beta model (2*x*MLLFV=5.08) but not by the S_B (2*x*MLLFV=3.8), and one in which it is rejected by the S_B (2*x*MLLFV=4.64) but not by the Beta (2*x*MLLFV=2.20).

Interestingly, in the two of seven cases in which normality is not rejected by the S_B and the Beta models, it is rejected by the S_U model; and in four of the nine cases in which normality is not rejected by the S_U model, it is rejected by the S_B or the Beta models. These results are consistent with the previously discussed theoretical properties of these distributions, i.e. the fact that the S-K areas covered by the S_B and the Beta overlap substantially while the S_U spans an entirely different S-K region.

The S_B model shows the highest maximum log-likelihood function value (MLLFV) in 11 of the 26 cases, versus nine for the Beta and six for the S_U. When the S_U, S_B or Beta model with the highest MLLFV is selected as the most suitable non-normal model, the normality hypothesis (Ho: $\theta=\mu=0$) is rejected 21, 17 and 10 out of 26 times at the ten, five and one percent significance levels, respectively. Note than four of the five non-rejections of normality ($\alpha=0.10$) correspond to the smaller ($T\leq30$) sample sizes and the fifth corresponds to a relatively small sample size of 34. This suggests that normality could also be rejected in at least some of those cases if larger sample sizes were available.

Of the 21 models in which normality is rejected ($\alpha=0.10$), the S_B model shows the highest MLLFV in seven cases, versus eight for the Beta and six for the S_U. The S-K combinations corresponding to the highest MLLFV models are presented in figure 2.

Note that three of the estimated S_U distributions exhibit quite large (>50) kurtosis values and are thus not shown in figure 2. The S-K combinations of the remaining 18 non-normal distributions stretch from fairly low to relatively high S-K value combinations.

Because these three models are not nested to each other, a LR test to ascertain if one is statistically superior to the other in a particular application is theoretically inappropriate. Note, however, that no MLLFV differences of more than 1.61 units are found between the eight Beta models with the highest MLLFV and the corresponding S_B models. That is, when the Beta model exhibits the highest MLLFV the corresponding S_B model's MLLFV is no more than 1.61 units lower. The average (Beta-S_B) MLLFV difference in these eight cases is 0.62. This supports the hypothesis that, in any particular application, the S_B distribution model is a fairly close statistical substitute for the Beta in terms of the likelihood of having generated the yield sample corresponding to that application. However, the question remains of how these seemingly small MLLFV differences translate into cumulative probability discrepancies. This question will be explored through simulation analyses in the next section.

Two noticeably larger MLLFV differences (2.86 and 2.59) are found between the seven S_B models with the highest MLLFV and the corresponding Beta models. The average (S_B-Beta) MLLFV difference in these seven cases is 1.12. These results are not surprising since, according to theory, all S-K combinations allowed by the Beta can also be modeled by the S_B but a significant portion of the S-K space spanned by the S_B is unattainable with the Beta. In other words, it is possible that the Beta model is not a close statistical substitute for the S_B in terms of the likelihood of having generated the yield sample corresponding to some applications.

Out of the six non-normal models for which the S_U exhibits the highest MLLFV, there are two cases where the MLLFV corresponding to the S_U model is substantially higher (3.05 and 3.58 units) than the highest of the S_B and Beta models. The average of these six MLLFV differences is 1.33. Alternatively, out of the 15 non-normal models for which the S_B or the Beta model exhibits the highest MLLFV, there are four cases where this highest MLLFV is substantially higher (4.23, 2.09, 2.31, and 2.10 units) than the MLLFV corresponding to the S_U model. The average of these 15 MLLFV differences is 1.19. This is consistent with the theoretical knowledge that the S-K region covered by the S_U model is does not overlap with the areas covered by the S_B or the Beta models.

In short the claim that, because it can accommodate all theoretically possible MVSK combinations, the expanded Johnson system is flexible enough to properly represent the diversity of continuous distributions that might be encountered in practice, is supported by the previously discussed empirical results. Specifically, these results suggest that the while the S_B distribution may be an adequate substitute for the Beta model, the Beta might not be able to effectively replace the S_U and the S_B models in some applications. The following simulation analyses provide further evidence in this regard.

Simulation Evidence of the Flexibility of the Johnson System

While the previous section provides interesting insights about the flexibility of the Johnson system, an assessment of how well this system can approximate a variety of distributional shapes generated from the Beta density is a more definitive means to test Ramirez and McDonald (2006a) hypothesis. Such evaluation is more credible if the Beta-generated distributional shapes are empirically motivated, i.e. derived from parametric Beta models that have been estimated on the basis of actual yield data. The previously

discussed S_U, S_B and Beta models are used for this purpose. Specifically, 21 datasets of 100,000 observations each are simulated on the basis of the six S_U, seven S_B and eight estimated Beta models. The following simulation formulas are based on equations (10) to (13) in Ramirez and McDonald (2006a) and equation (10) above:

(13) $$SS_U = M_t + \{\sigma_t[\sinh(\theta\{Z+\mu\}) - F_{SU}] \div \theta\sqrt{G_{SU}}\}$$

(14) $$SS_{SB} = M_t + \{\sigma_t \exp(\theta[Z-\mu]) \div \sqrt{G_{SB}}[1+\exp(\theta[Z-\mu]) - F_{SB}]\}$$

(15) $$S_B = M_t + \sigma_t\{(B - F_B) \div \sqrt{G_B}\}$$

where Z is a draw from a standard normal and B is a draw from a Beta distribution with parameters μ and 1/θ. Also in order to replicate the original data-generating process M_t and σ_t are computed 2500 times for values of t ranging from 1 to 40 to obtain the desired total of 100,000 observations. Although most yield distribution estimation applications involve small samples, very large simulated samples are required to precisely evaluate how closely a probability distribution function can approximate another.

Next, a second round of S_U, S_B, Beta and Normal models are estimated on the basis of each of those 21 datasets. Key statistics about these models are presented in table 2 (data-generating process=S_B), 3 (data-generating process=Beta) and 4 (data-generating process=S_U). The MLLFVs reported in these tables are divided by 2500 in order to make them comparable in magnitude to those that would be expected from a sample of size 40.

As expected, in all 21 cases, the models that exhibit the highest MLLFVs are those that are based on the probability distribution (S_U, S_B, or Beta) that was used to simulate the data, and the parameter estimates corresponding to those models (available form the authors) are very close to the parameter values used for the simulations.

In the case of the seven sets of models corresponding to the S_B-generated datasets (farms B, C, J, K, N, O and P in table 2) the MLLFVs of the Beta models seem relatively close to those of the S_B models, with differences ranging from 0.03 (farm P) to 2.05 (farm N) and averaging 1.02 units. At 2.53 units, the average MLLFV difference between the S_B and the S_U models is substantially larger. The normal models show substantially lower MLLFVs than any of the three non-normal models in all cases.

In the case of the eight sets of models corresponding to the Beta-generated datasets (farms E, G, M, Q, T, U, V, and Y in table 3), with differences ranging from 0.06 (farm M) to 0.68 (farm Q) and averaging 0.39 units, the MLLFVs of the S_B models are noticeably closer to those of the Beta models. At 1.26 units, the average MLLFV difference between the Beta and the S_U models is again markedly larger. As before, the normal models show much lower MLLFVs than any of the three non-normal models.

In the case of the six sets of models corresponding to the S_U-generated datasets (farms A, D, I, R, S and X in table 4), both the S_B and the Beta models yield MLLFVs that are substantially lower than those of the S_U models. On average, the MLLFVs are 11.71 units lower in the S_B models, 12.60 units lower in the Beta models, and 13.94 units lower in the normal models. In two of the six cases (farms I and X), the S_B and the Beta models can not do any better than the normal, while in the other four cases the low skewness and kurtosis and the MLLFVs suggest a relative closeness to normality.

In short, the MLLFV analysis suggests that the S_U model is not a good substitute for either the S_B or the Beta, and the S_B and the Beta models are poor surrogates for the S_U. On the positive side, it appears that the S_B and the Beta models could be acceptable substitutes for each other, with the S_B being a better surrogate for the Beta than the Beta

is for the S_B. However, the question remains of exactly how well these non-normal models can substitute for each other. To answer this question, the cumulative distribution functions (CDFs) implied by the second-round S_U, S_B, Beta and Normal models, are derived for each of the 21 cases. These are based on a second round of yield simulations (n=20 million) using these models' parameter estimates and equations (12)-(15) setting t=40. Equation (12) and a standard normal generator are used in the case of the normal models. The "true" CDFs are also derived using the correct distribution and the exact parameters underlying each of the 21 data-generating processes.

Two main statistics related to these CDFs are also presented in tables 2, 3 and 4. AD is the average of 125 vertical percentage distances between the true and the estimated CDFs. Distances are computed for yield values ranging from 25% to 150% of the average yields at equal 1% intervals (CDF values beyond that range are negligible in all cases). MD represents the maximum of those 125 vertical distances.

In the case of farm B (table 2), for example, the data-generating process is S_B and, therefore, the CDF corresponding to the estimated S_B model is extremely close to the true CDF (AD=0.02%, MD=0.22%). Interestingly, the CDF derived from the estimated S_U model (AD=0.07%, MD=0.31%) is also very close to the true CDF. Note that this closeness is reflected in a minimal (0.01 unit) MLLFV difference between the S_U and the S_B models. The outstanding performance of the S_U model in this case might be explained by the fact that the skewness-kurtosis mixture of the S_B is relatively close to the S_B-S_U boundary. A similarly accurate approximation of the S_B by the S_U model is observed in the case of farm J, which is almost at the boundary (figure 2).

In contrast, with an AD of 0.95% and a MD if 4.58%, the CDF associated with the estimated Beta model for farm B is not that close to the true S_B-based CDF. This is consistent with the relative large, 1.33 unit difference, between the S_B and the Beta model MLLFVs. This relatively poor performance of the Beta model could be related to the fact that the skewness-kurtosis mixture of the S_B is outside of the region allowed by the Beta distribution (figure 2). With an AD of 3.54% and a MD of 14.28%, the normal model's performance is abysmal in this case, which is reflected on its much lower MLLFV.

Farm O (table 2) is an example of a case where the estimated Beta model (AD=0.53%, MD=1.82%) does a fairly good job of approximating the S_B data-generating process. The relatively small (0.36 unit) MLLFV difference between the Beta and the S_B models is again a consistent signal of a good fit. In addition, the skewness and kurtosis values implied by the estimated S_B (-0.81 and -0.06) and Beta (-0.75 and 0.01) models are very close to each other. With an AD of 1.60%, a MD of 4.23%, and a MLLFV difference of 1.79 units, the S_U model does not provide for a very good fit of the S_B data-generating process in this case. The normal model is again the worse fitting.

Of the eight cases in which the data-generating process is Beta (table 3), farm V is the one where the estimated S_B model does worse on being able to replicate the Beta-generated CDF. Even in this case, an AD of 0.66%, a MD of 3.23% and a MLLFV difference of 0.50 indicate a fairly decent fit. At -2.05 and 5.30, the skewness and kurtosis values implied by the S_B model are very close to those implied by the estimated Beta model (-1.85 and 5.00) and to the true underlying values (-1.87 and 5.08). With an AD of 1.11%, a MD of 5.01%, and a MLLFV difference of 1.20 units the estimated S_U

model's fit is noticeably worse. The normal model's AD, MD, and MLLFV difference (3.09%, 14.11%, and 14.95 units, respectively) are by far the largest.

With an AD of 0.58%, a MD of 1.95%, and a MLLFV difference of 0.47 units, farm Y is the most typical of the eight cases in representing the S_B model's capacity to replicate a Beta-generated CDF. Figure 3 provides a visual cue of the closeness with which the estimated S_B model approximates the true CDF. All vertical differences in the lower one-third of the CDF are in fact less than 1.1%. That is, the S_B model can predict cumulative probability at any point within the lower third of the true CDF with a margin of error of 1.1% or less. This is particularly significant because the lower (left) tail is the relevant segment of the CDF for the purposes of risk analyses.

Of the six cases where the data-generating process is S_U (table 4), farm D is the only one in which the estimated S_B and, to a lesser extent, the Beta model, do a relatively good job at approximating the true CDF (ADs of 0.95% and 1.42%, MDs of 3.11% and 4.61%, and MLLFV differences of 0.65 and 1.36, respectively). This might be related to the fact that the S_U skewness and kurtosis values are not too far form the S-K regions that can be accommodated by the S_B and the Beta distributions (figure 2). Both the S_B and the Beta approximations are progressively worse in the case of farm S and R.

In the case of farms I and X (S_U model skewness and kurtosis not shown in figure 2 due to scale limitations), in contrast, the MLLFVs of the S_B and Beta models approach that of the normal model as their implied skewness and kurtosis near zero. That is, because of the high kurtosis and kurtosis/skewness ratios associated with these two S_U data-generating processes, the normal models turn out to be better in approximating the true CDFs that any possible S_B or Beta models. With an AD of 5.04% and a MD of

15.60% (farm I), and an AD of 9.25% and a MD of 24.85% (farm X), however, the approximations are far from acceptable.

Averages of the previously discussed statistics for the seven S_B, eight Beta and six S_U data-generating processes are also presented in tables 2, 3 and 4. These averages provide additional support towards the following conclusions: a) The S_B can approximate the Beta distribution with a relatively low margin of error; b) The S_B distribution is more precise in approximating the Beta than the Beta distribution is at approximating the S_B; c) In general, the S_U can not approximate the S_B or the Beta distributions as well as these two are able to approximate each other; d) Some of the S_B and Beta approximations of the S_U distribution are subject to very large error; and e) In most cases the normal approximations of any of these three non-normal models are by far the least accurate.

These conclusions are consistent with theoretical expectations based on the S-K regions that are covered by these distributions (figure 1) and support Ramirez and McDonald's hypothesis that the Johnson system (i.e. a combination of the S_U and the S_B distributions) is sufficient to approximate any probability distribution that might be encountered in practice. At the very least, the results suggest that the Johnson system is a superior alternative to the Beta for the modeling of crop yield distributions.

Concluding Remarks

This research demonstrates that a comprehensive coverage of the theoretically feasible region of the S-K plane by a parametric probability distribution model is an essential condition for the model to provide for an acceptable approximation of the "unknown" probability distribution underlying a data-generating process. For instance, it is shown that if the "unknown" distribution is a S_U with a skewness-kurtosis combination much

beyond the S_B (or the Beta) S-K coverage area, the accuracy of a S_B model-based (or a Beta model-based) approximation deteriorates substantially, and vice versa. Therefore, it is clear that parametric models based on a distribution such as the Beta, which leaves substantial areas of the theoretically feasible S-K region uncovered, might not provide for an acceptable approximation of the true underlying distribution in some applications.

The choice of the S_B versus the Beta as a complement to the S_U distribution is justified on the following basis: a) The S_B's ability to accommodate all S-K combinations allowed by the Beta plus an additional, non-negligible, area of the theoretically feasible S-K region that is not covered by the S_U; b) This research's empirical finding that the S_B distribution is more precise in approximating the Beta than the Beta distribution is at approximating the S_B; and 3) The empirically valuable fact that a multivariate density involving S_B and S_U distributions can be specified, estimated and used as a basis for joint simulation (as exemplified in Ramirez and McDonald 2006).

The Beta was selected for this comparative evaluation of the Johnson system because its spanning of the theoretically feasible S-K space is much more comprehensive than the coverage afforded by the other distributions that have been used for the parametric modeling and simulation of crop yields. The fact the S_B can approximate the Beta with a relatively low margin of error does not necessarily imply that it can approximate all other possible alternative distributions with skewness-kurtosis values on its S-K coverage area with similar accuracy. Likewise, it does not ensure that the S_U can approximate all alternative distributions on its S-K coverage area with comparable levels of precision. However, the results from this S_B-Beta comparison are likely indicative of how well the Johnson system may be able to approximate other distributions.

Undoubtedly, using parametric models that are based on the Johnson system instead of on distributions such as the Beta or the Gamma would substantially reduce the specification error risk that has long been associated with these models, perhaps to a level that is acceptable in most applications. Although it is not possible to prove a negative, additional comparative evaluations including lesser-known alternative distributions such as normal mixtures could provide further support to this claim

References

Coble, K.H., T.O. Knight, R.D. Pope, and J.R. Williams. 1996. "Modeling Farm-level Crop Insurance Demand with Panel Data." *American Journal of Agricultural Economics* 78(2):439-447.

Gallagher, P. 1987. "U.S. Soybean Yields: Estimation and Forecasting with Nonsymmetric Disturbances." *American Journal of Agricultural Economics* 69(4):796-803.

Harri, A., K.H. Coble, C. Erdem, and T.O. Knight. 2005. "Crop Yield Normality: A Reconciliation of Previous Research." Working Paper, Department of Agricultural Economics, Mississippi State University, Starkville, Mississippi.

Ker, A.P. and K. Coble. 2003. "Modeling Conditional Yield Densities." *American Journal of Agricultural Economics* 85(2):291-304.

Mood, A.M., F.A. Graybill, and D.C. Boes. *Introduction to the Theory of Statistics*. 3rd ed. New York: McGraw-Hill, 1974.

Moss, C.B., and J.S. Shonkwiler. 1993. "Estimating Yield Distributions Using a Stochastic Trend Model and Non-Normal Errors." *American Journal of Agricultural Economics* 75(4):1056-1062.

Nelson, C.H. and P.V. Preckel. 1989. "The Conditional Beta Distribution as a Stochastic Production Function." *American Journal of Agricultural Economics* 71(2):370-378.

Norwood, B., M.C. Roberts and J.L. Lusk. 2004. "Ranking Crop Yield Models Using Out-of-Sample Likelihood Functions." *American Journal of Agricultural Economics* 86(4):1032-1043.

Ramirez, O.A. 1997. "Estimation and Use of a Multivariate Parametric Model for Simulating Heteroskedastic, Correlated, Non-Normal Random Variables: The Case of Corn-Belt Corn, Soybeans and Wheat Yields." *American Journal of Agricultural Economics* 79(1):191-205.

Ramirez, O.A. and T. McDonald 2006a. "The Expanded Johnson System: A Highly Flexible Crop Yield Distribution Model." Paper presented at the 2006 meeting of the American Agricultural Economics Association, Long Beach, California, July 23-26 2006 and posted in AgEcon Search (http://agecon.lib.umn.edu/).

Ramirez, O.A. and T.U. McDonald. 2006b. "Ranking Crop Yield Models: A Comment." *American Journal of Agricultural Economics* 88(4):1105-1110.

Ramírez, O.A., S.K. Misra, and J.E. Field. 2003. "Crop Yield Distributions Revisited." *American Journal of Agricultural Economics* 85(1):108-120.

Ramirez, O.A., C.B. Moss, and W.G. Boggess. 1994. "Estimation and Use of the Inverse Hyperbolic Sine Transformation to Model Non-Normal Correlated Random Variables." *Journal of Applied Statistics* 21(4):289-305.

Taylor, C.R. 1990. "Two Practical Procedures for Estimating Multivariate Non-Normal Probability Density Functions." *American Journal of Agricultural Economics* 72(1):210-217.

Table 1. Select Statistics for Illinois Farm-level Corn Yield Models Based on the S_U, the S_B, the Beta and the Normal Distributions

Farm Label	Sample Size	S_U MLLFV	S_B MLLFV	Beta MLLFV	Normal MLLFV	LRTS	Final Model
A	44	-183.62	-186.67	-187.24	-191.64	16.03^{3}	S_U
B	32	-123.81	-123.81	-126.39	-134.94	22.27^{3}	S_B
C	44	-186.38	-182.15	-185.00	-187.61	10.91^{3}	S_B
D	43	-189.23	-189.39	-189.54	-192.55	6.63^{2}	S_U
E	25	-108.09	-108.00	-107.72	-112.23	9.01^{2}	Beta
F	27	-128.31	-127.08	-127.55	-128.98	3.81^{0}	N
G	31	-133.58	-133.57	-133.26	-140.68	14.83^{3}	Beta
H	34	-161.15	-160.20	-160.93	-161.80	3.20^{0}	N
I	43	-181.27	-184.84	-184.94	-185.62	8.71^{2}	S_U
J	32	-145.96	-145.94	-146.56	-149.20	6.53^{2}	S_B
K	27	-120.75	-118.66	-118.98	-126.11	14.90^{3}	S_B
L	29	-132.56	-132.49	-132.55	-132.56	0.13^{0}	N
M	37	-169.08	-169.00	-168.95	-171.97	6.02^{2}	Beta
N	45	-197.46	-195.15	-196.37	-197.47	4.64^{1}	S_B
O	42	-189.54	-188.40	-188.55	-194.36	11.92^{3}	S_B
P	42	-195.34	-195.28	-195.31	-197.77	4.97^{1}	S_B
Q	40	-174.07	-173.55	-172.74	-178.18	10.88^{3}	Beta
R	33	-145.36	-145.47	-145.67	-150.09	9.46^{3}	S_U
S	40	-181.77	-182.35	-182.50	-184.12	4.70^{1}	S_U
T	29	-131.07	-131.05	-129.44	-133.79	8.69^{2}	Beta
U	44	-201.83	-201.21	-200.55	-204.01	6.91^{2}	Beta
V	29	-127.78	-126.34	-125.69	-131.64	11.91^{3}	Beta
W	29	-131.22	-131.24	-131.20	-132.56	2.71^{0}	N
X	20	-93.45	-93.96	-94.00	-98.42	9.94^{3}	S_U
Y	29	-135.14	-135.00	-134.35	-136.90	5.08^{3}	Beta
Z	30	-143.92	-143.26	-143.37	-144.92	3.32^{0}	N

Notes: MLLFV stands for the maximum log-likelihood function value and LRTS indicates the likelihood ratio test statistic, which compares the non-normal model with the highest MLLFV with the normal model. The superscripts 1, 2 and 3 denote rejection of the null hypothesis of normality and the 10, 5 and 1% levels, respectively, according to the likelihood ratio test, while 0 indicates non rejection at the 10% level. If the null hypothesis of normality is rejected at the 10% level the final model is the one with the highest MLLFV, otherwise the final model is the normal.

Table 2. Key Statistics about the S_U, S_B, Beta and Normal Models Estimated on the Basis of the Seven S_B-Simulated Datasets

	FARM B (DGP=S_B)				FARM C (DGP=S_B)				FARM J (DGP=S_B)			
Model	**S_U**	**S_B**	**Beta**	**Norm**	**S_U**	**S_B**	**Beta**	**Norm**	**S_U**	**S_B**	**Beta**	**Norm**
MLLFV	-156.70	-156.69	-158.02	-172.58	-168.81	-163.41	-164.79	-173.64	-181.44	-181.44	-181.97	-190.54
Skew	-3.25	-2.77	-1.46	0	-3.24	-0.63	-0.71	0	-2.01	-1.93	-1.19	0
Kurt	23.27	14.25	3.04	0	23.15	-0.74	-0.30	0	7.92	7.04	2.08	0
AD	0.07%	0.02%	0.95%	3.54%	2.65%	0.02%	1.13%	2.67%	0.04%	0.01%	0.78%	3.44%
MD	0.31%	0.22%	4.58%	14.28%	7.67%	0.25%	4.83%	10.21%	0.14%	0.07%	2.77%	10.78%
	FARM K (DGP=S_B)				**FARM N (DGP=S_B)**				**FARM O (DGP=S_B)**			
Model	**S_U**	**S_B**	**Beta**	**Norm**	**S_U**	**S_B**	**Beta**	**Norm**	**S_U**	**S_B**	**Beta**	**Norm**
MLLFV	-180.46	-176.62	-178.09	-192.06	-177.98	-171.56	-173.61	-178.09	-180.70	-178.92	-179.29	-185.33
Skew	-7.42	-1.05	-1.50	0	0.36	-0.09	-0.16	0	-2.13	-0.81	-0.75	0
Kurt	176.36	0.17	2.62	0	0.23	-1.21	-1.09	0	9.02	-0.06	0.01	0
AD	4.20%	0.04%	1.36%	5.04%	2.19%	0.04%	0.74%	2.18%	1.60%	0.02%	0.53%	2.94%
MD	11.26%	0.29%	3.80%	14.06%	6.41%	0.14%	2.51%	6.70%	4.23%	0.13%	1.82%	8.92%
	FARM P (DGP=S_B)				**AVERAGES (DGP=S_B)**			
Model	**S_U**	**S_B**	**Beta**	**Norm**	**S_U**	**S_B**	**Beta**	**Norm**
MLLFV	-185.85	-185.56	-185.59	-187.83	-175.99	-173.46	-174.48	-182.87
Skew	-0.96	-0.66	-0.67	0	-2.66	-1.13	-0.92	0.00
Kurt	1.70	0.19	0.32	0	34.52	2.81	0.95	0.00
AD	0.62%	0.04%	0.21%	2.13%	1.62%	0.03%	0.81%	3.13%
MD	1.49%	0.07%	0.57%	5.56%	4.50%	0.17%	2.98%	10.07%

Notes: DGP stands for data-generating process; MLLFV, Skew and Kurt refer to the maximum log-likelihood function, skewness and kurtosis values; AD is the average of 125 vertical percentage distances between the true and the estimated CDFs and MD represents the maximum of those 125 vertical distances. Distances are computed for yield values ranging from 25% to 150% of the mean yields at equal 1% intervals (CDF values beyond that range are negligible in all cases).

Table 3. Key Statistics about the S_U, S_B, Beta and Normal Models Estimated on the Basis of the Eight Beta-Simulated Datasets

	FARM E (DGP=Beta)				FARM G (DGP=Beta)				FARM M (DGP=Beta)			
Model	**S_U**	**S_B**	**Beta**	**Norm**	**S_U**	**S_B**	**Beta**	**Norm**	**S_U**	**S_B**	**Beta**	**Norm**
MLLFV	-181.38	-181.00	-180.71	-190.63	-172.60	-172.16	-171.79	-182.69	-183.60	-183.20	-183.14	-186.53
Skew	-2.78	-1.71	-1.56	0	-3.16	-1.83	-1.59	0	-1.24	-0.84	-0.81	0
Kurt	16.32	3.94	3.56	0	21.83	4.55	3.62	0	2.86	0.51	0.54	0
AD	0.83%	0.49%	0.05%	3.44%	0.86%	0.49%	0.00%	2.91%	0.58%	0.16%	0.02%	1.93%
MD	2.88%	1.79%	0.24%	10.73%	3.46%	2.11%	0.04%	11.22%	1.75%	0.62%	0.08%	6.36%
	FARM Q (DGP=Beta)				**FARM T (DGP=Beta)**				**FARM U (DGP=Beta)**			
Model	**S_U**	**S_B**	**Beta**	**Norm**	**S_U**	**S_B**	**Beta**	**Norm**	**S_U**	**S_B**	**Beta**	**Norm**
MLLFV	-173.72	-173.09	-172.41	-186.76	-178.86	-178.21	-177.71	-191.80	-183.79	-181.38	-181.12	-186.05
Skew	-3.98	-2.09	-1.80	0	4.45	-2.00	-1.79	0	-1.41	-0.61	-0.55	0
Kurt	37.64	5.75	4.64	0	49.11	5.09	4.66	0	3.72	-0.50	-0.53	0
AD	0.98%	0.71%	0.01%	3.46%	1.52%	0.97%	0.13%	4.56%	2.02%	0.51%	0.01%	2.83%
MD	3.85%	2.97%	0.04%	14.16%	4.99%	3.15%	0.55%	13.74%	4.34%	1.23%	0.05%	6.89%
	FARM V (DGP=Beta)				**FARM Y (DGP=Beta)**				**AVERAGES (DGP=Beta)**			
Model	**S_U**	**S_B**	**Beta**	**Norm**	**S_U**	**S_B**	**Beta**	**Norm**	**S_U**	**S_B**	**Beta**	**Norm**
MLLFV	-173.22	-172.52	-172.02	-186.98	-186.69	-184.53	-184.06	-191.69	-179.23	-178.26	-177.87	-187.89
Skew	-4.74	-2.05	-1.85	0	-2.41	-0.95	-0.84	0	-1.91	-1.51	-1.35	0.00
Kurt	57.24	5.30	5.00	0	11.79	0.22	0.07	0	25.06	3.11	2.70	0.00
AD	1.11%	0.66%	0.03%	3.09%	1.76%	0.58%	0.05%	2.80%	1.21%	0.57%	0.04%	3.13%
MD	5.01%	3.23%	0.18%	14.11%	5.00%	1.95%	0.30%	9.14%	3.91%	2.13%	0.19%	10.79%

Notes: DGP stands for data-generating process; MLLFV, Skew and Kurt refer to the maximum log-likelihood function, skewness and kurtosis values; AD is the average of 125 vertical percentage distances between the true and the estimated CDFs and MD represents the maximum of those 125 vertical distances. Distances are computed for yield values ranging from 25% to 150% of the mean yields at equal 1% intervals (CDF values beyond that range are negligible in all cases).

Table 4. Key Statistics about the S_U, S_B, Beta and Normal Models Estimated on the Basis of the Six S_U-Simulated Datasets

	FARM A (DGP=S_U)				FARM D (DGP=S_U)				FARM I (DGP=S_U)			
Model	**S_U**	**S_B**	**Beta**	**Norm**	**S_U**	**S_B**	**Beta**	**Norm**	**S_U**	**S_B**	**Beta**	**Norm**
MLLFV	-165.79	-174.33	-176.38	-177.93	-175.40	-176.05	-176.76	-178.60	-167.28	-185.23	-185.23	-185.23
Skew	-3.94	-0.31	-0.21	0	-1.18	-0.58	-0.35	0	-0.83	0	0	0
Kurt	58.33	0.14	0.03	0	3.44	0.60	0.15	0	369.63	0	0	0
AD	0.06%	2.96%	3.74%	3.91%	0.02%	0.95%	1.42%	2.03%	0.08%	5.04%	5.04%	5.04%
MD	0.25%	11.34%	12.92%	14.02%	0.09%	3.11%	4.61%	6.73%	0.18%	15.60%	15.60%	15.60%
	FARM R (DGP=S_U)				**FARM S (DGP=S_U)**				**FARM X (DGP=S_U)**			
Model	**S_U**	**S_B**	**Beta**	**Norm**	**S_U**	**S_B**	**Beta**	**Norm**	**S_U**	**S_B**	**Beta**	**Norm**
MLLFV	-176.92	-178.85	-180.59	-184.67	-181.43	-183.89	-184.68	-185.28	-186.37	-225.13	-225.13	-225.13
Skew	-2.11	-0.73	-0.45	0	-1.39	-0.25	-0.13	0	-29.46	0	0	0
Kurt	10.14	0.95	0.27	0	5.71	0.09	-0.01	0	10369.3	0	0	0
AD	0.10%	1.86%	2.57%	3.80%	0.06%	1.56%	1.93%	2.45%	0.04%	9.25%	9.25%	9.25%
MD	0.35%	5.75%	7.56%	10.69%	0.17%	5.42%	6.47%	7.68%	0.18%	24.85%	24.85%	24.85%

	AVERAGES (DGP=S_U)			
Model	**S_U**	**S_B**	**Beta**	**Norm**
MLLFV	-175.53	-187.24	-188.13	-189.47
Skew	-6.49	-0.31	-0.19	0.00
Kurt	1802.76	0.30	0.07	0.00
AD	0.06%	3.60%	3.99%	4.41%
MD	0.20%	11.01%	12.00%	13.26%

Notes: DGP stands for data-generating process; MLLFV, Skew and Kurt refer to the maximum log-likelihood function, skewness and kurtosis values; AD is the average of 125 vertical percentage distances between the true and the estimated CDFs and MD represents the maximum of those 125 vertical distances. Distances are computed for yield values ranging from 25% to 150% of the mean yields at equal 1% intervals (CDF values beyond that range are negligible in all cases).

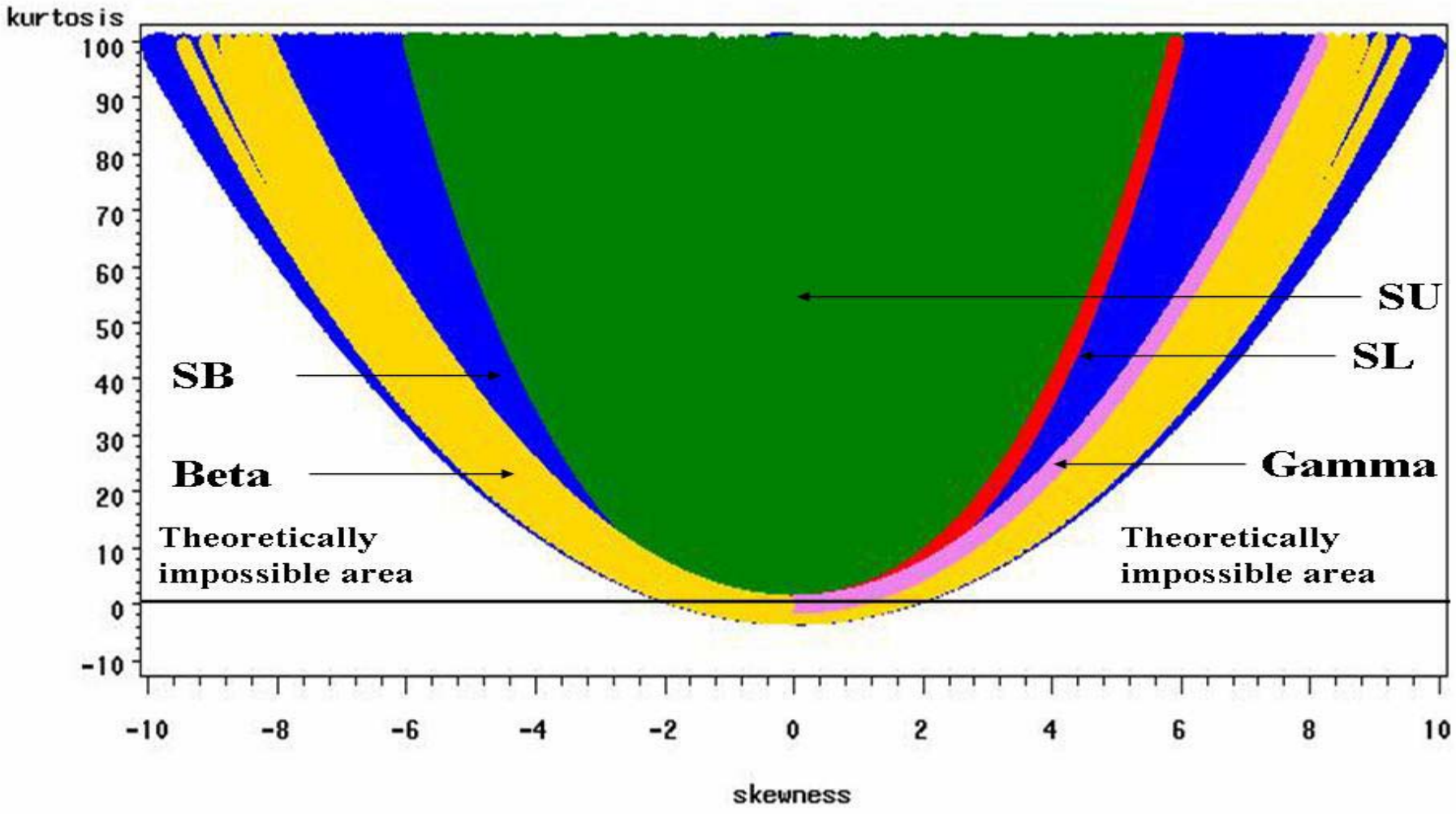

Figure 1. S_U, S_L, S_B, Beta and Gamma distributions in the S-K plane

Note: The S_B distribution allows all S-K combinations in the blue as well as in the yellow (Beta) and pink (Gamma) areas.

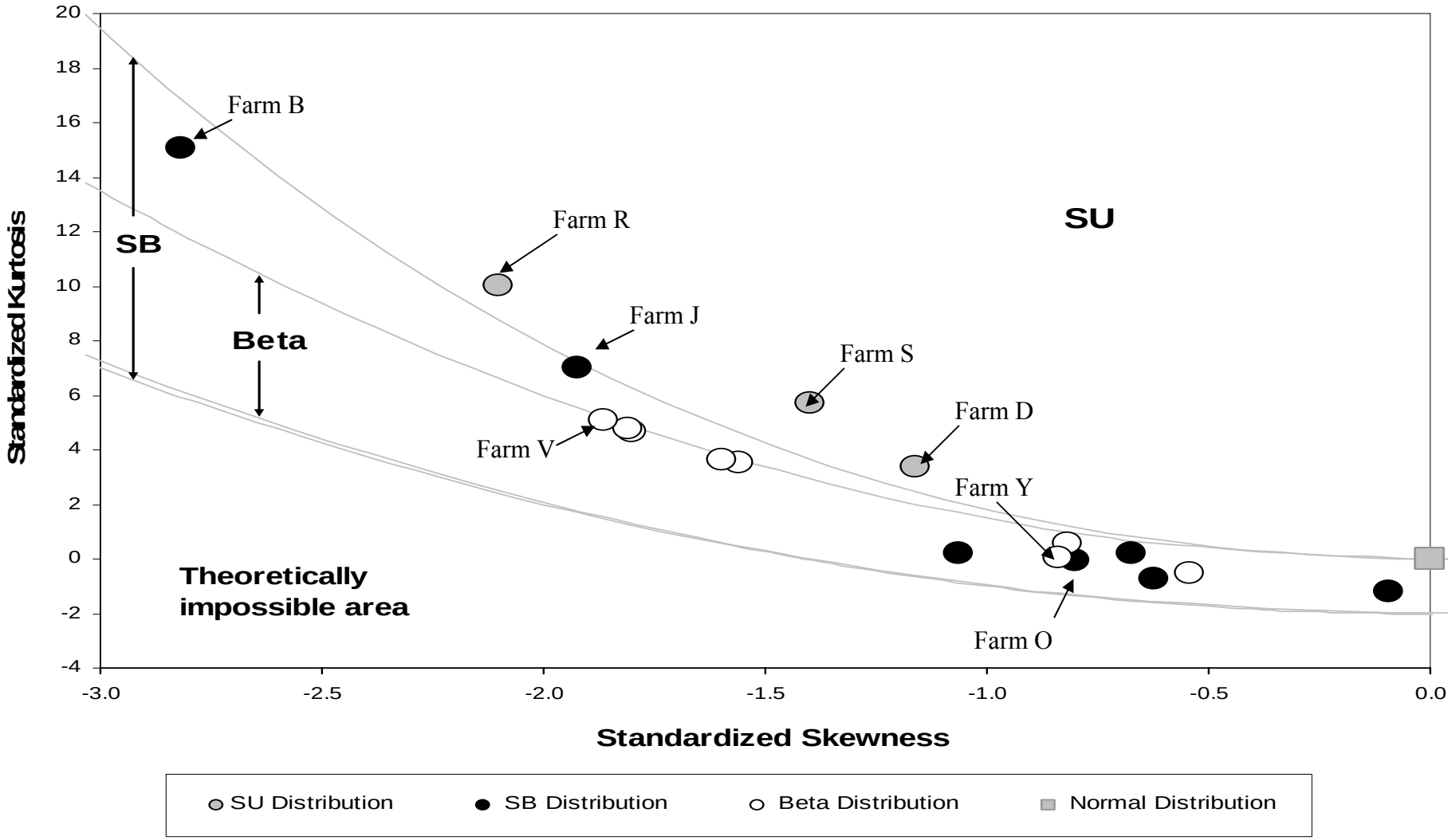

Figure 2. Skewness-kurtosis combinations of estimated non-normal models

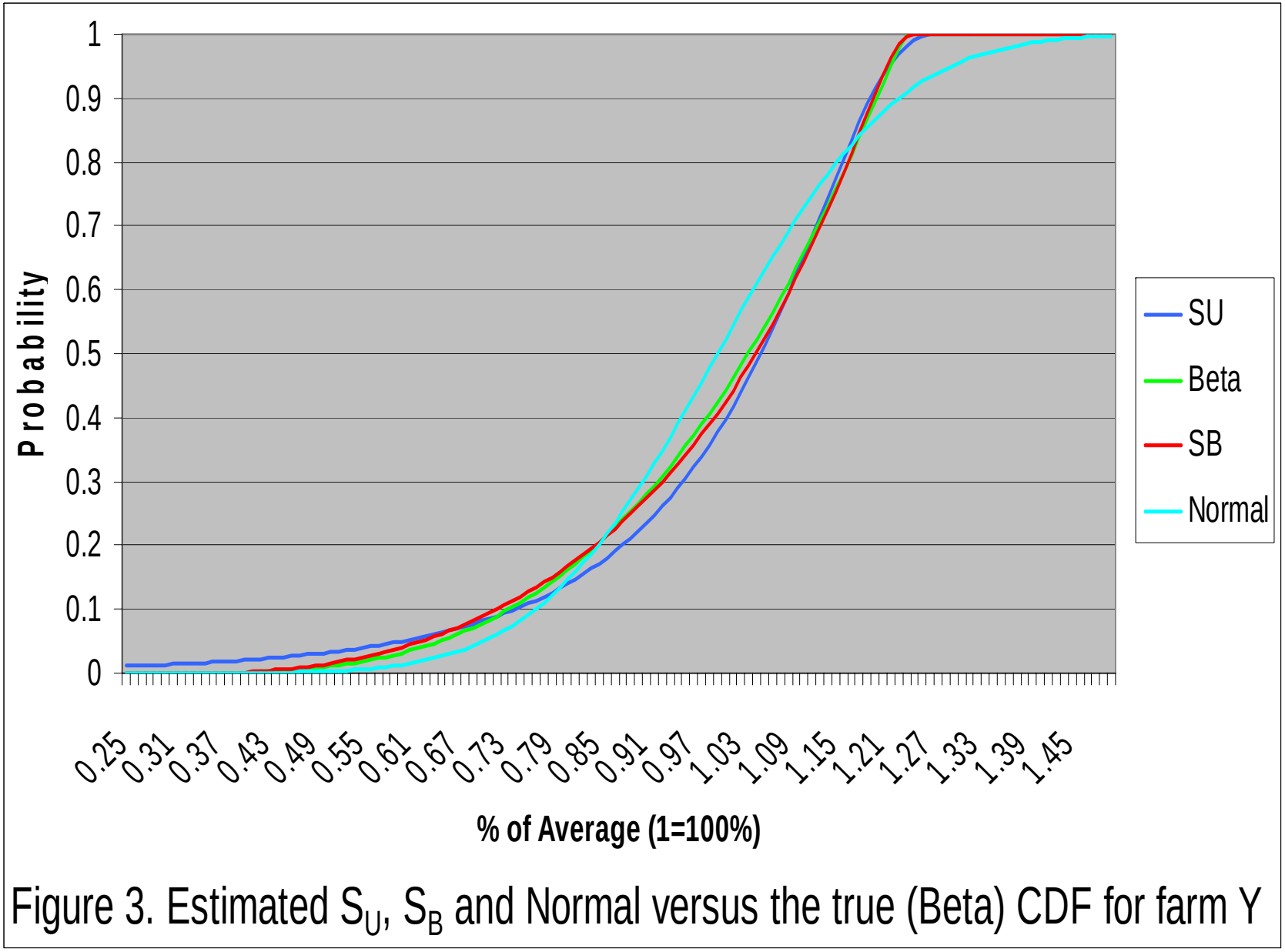

Figure 3. Estimated S_U, S_B and Normal versus the true (Beta) CDF for farm Y