"The Expanded Johnson System:

A Highly Flexible Crop Yield Distribution Model"

by

Octavio A. Ramirez[1]

and

Tanya U. McDonald[2]

Selected Paper presented at the American Agricultural Economics Association Annual Meeting, Long Beach, California, July 23-26, 2006.

1 Corresponding author, 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, fax: (505) 646-3808.

2 Research Specialist, Department of Agricultural Economics and Agricultural Business, 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.

Agricultural economists have long 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. Anderson (1974) was first to emphasize the importance of accounting for non-normality in crop yield distributions, changes in crop yield variation over time and location, and the interdependence between crop yields and prices for the purpose of economic risk analyses.

Since then, numerous authors have focused on this issue (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 many others). These authors have provided irrefutable statistical evidence of non-normality and heteroskedasticity in crop yield distributions, specifically, of the existence of kurtosis and negative skewness in a variety of cases. The possibility of positive skewness has been documented as well (Ramirez, Misra and Field 2003).

The three general types of statistical procedures that have been used for the modeling and simulation crop yield distributions are the parametric, the non-parametric and semi-parametric approaches; all of which 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, the main weakness of this method is the potential error resulting from assuming a probability distribution that is not flexible enough to properly represent the yield data. Since crop yield data sets do not often span over long time periods, especially in the case of individual farms and specialty crops, the main advantage of this method is that it performs relatively well in small

sample applications. Distributions that have been used as a basis for this method include the Normal, the Log-normal, the Logistic, the Weibull, the Beta, the Gamma and the IHS.

The non-parametric approach has opposite advantages and disadvantages. Since this method is free of any functional form assumption, it is generally more flexible and exhibits a lower model specification error risk. A main disadvantage of this technique is that it is not very precise in small sample applications or when the model includes several explanatory variables. Also, non-parametric methods don't allow for the prediction or simulation of crop yields beyond of the observed sample time frame and can't measure the impacts of explanatory factors on yield levels (Horowitz and Lee 2002). A final disadvantage of the non-parametric approach is its theoretical complexity and intensive computational requirements (Yatchew 1998).

Semi-parametric methods such as single-index models, partially linear models, non-parametric additive models, and non-parametric additive models with interactions, have become increasingly popular in the econometrics literature because they combine the advantages while eliminating some of the problems of the parametric and the nonparametric approaches (Horowitz and Lee 2002; Ker and Coble 2003; Norwood, Roberts and Lusk 2004). Semi-parametric models are more flexible than parametric models and provide for a more precise estimation than non-parametric methods in small sample applications although at expense of a greater specification error risk.

Extensive efforts have been devoted to the issue of what is the most appropriate probability distribution to be used as a basis for parametric or semi-parametric methods. Gallagher (1987) used the well-known Gamma density as a parametric model for the soybean yields distribution. Nelson and Preckel (1989) proposed a conditional Beta

distribution to model corn yields. Taylor (1990) considered the issue of estimating 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 in the statistics literature as the S_U distribution), as a possible multivariate non-normal and heteroskedastic crop yield distribution model. More recently, Ker and Coble (2003) proposed a semi-parametric model based on the Normal and the Beta densities.

Empirical comparisons of leading parametric models have been attempted in the recent literature (Norwood, Roberts and Lusk 2004). These comparisons, however, have overlooked key theoretical considerations and, therefore, not been able to elucidate the superiority of any of the existing models (Ramirez and McDonald 2006).

Statistically, any particular probability distribution can only accommodate a limited subset of the theoretically feasible mean-variance-skewness-kurtosis (MVSK) hyperspace and, therefore, it is only capable of adequately modeling underlying data distributions which moments happen to be contained within that subset. Individually, or even as a collective, the parametric models previously discussed in the literature only span a relatively small region of the MVSK hyperspace (Ramirez and McDonald 2006).

This research contributes to the yield and price distribution modeling literature by introducing a system of three parametric distributions that is capable of accommodating the basic characteristics of the stochastic behavior of a random variable, i.e. all theoretically possible MVSK combinations, addressing the main disadvantage of parametric models that has been cited in the literature to date. The proposed system

substantially reduces the lack of flexibility concern and the resulting risk of model specification error previously associated with the parametric approach.

Other important advantages of the proposed system is that it can jointly model non-normality, heteroscedasticity, and/or autocorrelation, and it can be expressed in a multivariate form in order to model the joint distribution of two or more crop yield and/or price variables of interest. It is therefore hypothesized that this system is generally superior and can supersede all currently used parametric distribution models.

The Proposed Parametric Distribution Modeling System

Statistical theory suggests that the inherent capacity of a parametric probability distribution model to adequately represent most crop yield distributions that could be encountered in practice is mainly determined by the range of each of the first four central moments (mean, variance, skewness and kurtosis) that can be accommodated by the assumed probability distribution model.

Ramirez and McDonald (2006) outline a re-parameterization technique that expands any probability distribution by two parameters which specifically and uniquely control the distributional mean and variance without affecting the range of skewness and kurtosis values that can be accommodated by that distribution. The expanded distribution obtained through this re-parameterization can therefore model any conceivable mean and variance in conjunction with the set of skewness-kurtosis combinations that were allowed by the original distribution. In addition, the mean, variance, skewness and kurtosis of the resulting parametric model can be specified as linear or non-linear functions of exogenous variables so that these four distributional moments are allowed to change across observations as those variables take different values.

In this study, Ramirez and McDonald's (2006) re-parameterization is applied to the Johnson system, which includes the S_U, the S_B and the S_L or Log-Normal distributions. Unlike other frequently assumed distributions such as the Beta and the Gamma, the Johnson system exhibits the key property of being able to accommodate any and all theoretically feasible skewness-kurtosis combinations (figure 1). However, each of those combinations is inherently associated with a fixed set of mean-variance values. This re-parameterization enhances the flexibility of the Johnson system to where it can model all theoretically feasible MVSK combinations. The re-parameterization begins with the original two-parameter system, which is defined as follows:

(1) $Z = \gamma + \delta \times \sinh^{-1}(Y)$ for the S_U,

(2) $Z = \gamma + \delta \times \ln(Y)$ for the S_L, and

(3) $Z = \gamma + \delta \times \ln[Y/(1-Y)]$ for the S_B distribution,

where Y is a non-normally distributed random variable based on a standard normal variable (Z).

From (1), (2) and (3) it follows that:

(4) $Y = \sinh(\frac{Z-\gamma}{\delta}) = \sinh(N)$ for the S_U,

(5) $Y = \exp(\frac{Z-\gamma}{\delta}) = e^N$ for the S_L, and

(6) $Y = \frac{\exp(\frac{Z-\gamma}{\delta})}{(1+\exp(\frac{Z-\gamma}{\delta}))} = \frac{e^N}{(1+e^N)}$ for the S_B distribution,

where N is a normal random variable with mean of $-\frac{\gamma}{\delta}$ and variance of $\frac{1}{\delta^2}$.

Following Johnson (1949):

(7) $\mathrm{E[Y]} = \exp(0.5\delta^{-2}) \times \sinh(\gamma / \delta) = F_{SU}$

$\mathrm{V[Y]} = 0.5\{\exp(\delta^{-2}) - 1\} \times \{\exp(\delta^{-2}) \times \cosh(2\gamma\delta^{-1}) + 1\} = G_{SU}$,

(8) $\mathrm{E[Y]} = \exp(0.5\delta^{-2} - \gamma\delta^{-1}) = F_{SL}$

$\mathrm{V[Y]} = \exp(2\delta^{-2} - 2\gamma\delta^{-1}) - \{\exp(0.5\delta^{-2} - \gamma\delta^{-1})\}^2 = G_{SL}$, and

(9) $\mathrm{E[Y]} = F_{SB}$

$\mathrm{V[Y]} = G_{SB}$;

for the S_U, S_L and S_B distributions, respectively; where F_{SB} and G_{SB} are lengthier formulas of the parameters γ and δ. The skewness and kurtosis coefficients of the S_U, S_L and S_B distributions are lengthy functions of γ and δ as well. All of these formulas, an a Gauss 6.0 program to compute the first four central moments of these distributions given values for γ and δ are available from the authors.

The random variables (Y) corresponding to each of the three distributions are then standardized to have a mean of zero and a variance of one, as follows:

(10) $\mathrm{Y}^S = \dfrac{\sinh(N) - F_{SU}}{G_{SU}^{1/2}}$ for the S_U,

(11) $\mathrm{Y}^S = \dfrac{e^N - F_{SL}}{G_{SL}^{1/2}}$ for the S_L, and

(12) $\mathrm{Y}^S = \dfrac{\dfrac{e^N}{(1+e^N)} - F_{SB}}{G_{SB}^{1/2}}$ for the S_B distribution.

Note that after standardization, the parameters γ and δ no longer affect the mean and the variance of the distributions, rather, they are focused on controlling distributional

skewness and kurtosis only. Yet, since standardization only involves subtracting from and dividing the original random variables (Y) by constants, the distributions corresponding to these standardized variables (Y^S) can still accommodate the same sets of skewness- kurtosis combinations allowed by the S_U, S_L and S_B distributions in the original Johnson system.

The final step in the re-parameterization process is to expand the Y^S distributions so that, instead of being zero and one, their means and variances can be controlled by parameters or by parametric functions of explanatory variables as follows:

(13) $$Y_t^F = \sigma_t Y^S - \mathrm{M}_t = (Z_t\sigma)Y^S - (X_t\beta),$$

where Y^S is as defined in equations (10), (11) and (12) for the S_U, the S_L and the S_B distributions; t = 1,…,T denotes the observations; and Y_t^F represents the final random variables of interest. From (13), note that for all three re-parameterized variables:

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

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

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. Note that $\mathrm{M_t}$ and σ_t could also be specified as non-linear functions of X_t and Z_t.

Further, note that the three probability distributions corresponding to this final set of non-normal random variables (Y_t^F) maintains identical shape (i.e. skewness-kurtosis) characteristics as the original families in the Johnson system. Therefore, the re-parameterized system can accommodate any theoretically possible mean-variance-skewness-kurtosis combination. As a result, it is believed to be sufficient to accurately

model any crop yield distribution that could be encountered in practice. That is, the proposed system addresses the previously discussed lack of flexibility and model misspecification risk concerns that have been consistently cited as the main disadvantage of parametric probability distribution models.

Figure 1 illustrates the different skewness-kurtosis (S-K) regions covered by each of the three distributions in the Johnson system, as well as by the Beta and the Gamma. 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 extremely limited. Also note that this 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. Therefore, it is likely that the S_B is at least equally suitable as the Beta in a particular application. Note, however, that the region spanned by the Beta is quite narrower than the S_B's, i.e. the Beta only covers a subset of the S-K area spanned by the S_B. Therefore, it is possible that the Beta is not as suitable as the S_B in some applications.

In short, even if re-parameterized according to the previously discussed (Ramirez and McDonald 2006) procedure, because the Gamma and the Beta distributions can not accommodate a substantial subset the empirically possible S-K set, they can not be expected to be nearly as flexible and generally applicable as the Johnson system.

Estimation of the Expanded Johnson System

Estimation of the proposed system can be accomplished by maximum likelihood procedures. Since all three distributions originate from one-to-one transformations to a normal random variable (N), the transformation technique (Mood, Graybill and Boes 1974) can be applied to derive their corresponding probability distribution functions (pdf). According to this technique the pdf of the transformed random variable (Y_t^F) is given by:

$$(15) \qquad P(\mathrm{Y}_t^F) = \left|\frac{\partial(q^{-1}(Y_t^F))}{\partial \mathrm{Y}_t^F}\right| \times P(q^{-1}(Y_t^F)) = J(Y_t^F) \times P(q^{-1}(Y_t^F)),$$

where $q^{-1}(Y_t^F)$ is the inverse of the transformation of N into Y_t^F (i.e. the function relating N to Y_t^F), $P(q^{-1}(Y_t^F))$ is the pdf of an independently and identically distributed normal random variable N with mean $(-\frac{\gamma}{\delta})$ and variance δ^{-2} evaluated at $q^{-1}(Y_t^F)$, and the term $J(Y_t^F)$ is known as the Jacobian of the transformation.

Specifically, from equation (13) and equations (10) to (12) it follows that:

$$(16) \qquad N = q_{SU}^{-1}(Y_t^F) = \sinh^{-1}\{R_{SU\,t}\} \text{ for the } \mathrm{S_U},$$

$$(17) \qquad N = q_{SL}^{-1}(Y_t^F) = \ln\{R_{SL\,t}\} \text{ for the } \mathrm{S_L}, \text{ and}$$

$$(18) \qquad N = q_{SB}^{-1}(Y_t^F) = \ln\{\frac{R_{SB\,t}}{1 - R_{SB\,t}}\} \text{ for the } \mathrm{S_B} \text{ random variable,}$$

where:

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

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

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

and F_{SU}, F_{SL}, F_{SB}, G_{SU}, G_{SL} and G_{SB} are as defined in equations (7) to (9).

The Jacobians are obtained by taking the absolute value of the derivatives of the above inverse transformation functions (equations (16, 17 and 18) with respect to Y_t^F :

(22) $$J_{SU}(Y_t^F) = \frac{G_{SU}^{1/2}}{Z_t\sigma(1 + R_{SU\,t}^{\,2})^{1/2}}$$ for the $\mathrm{S_U}$,

(23) $$J_{SL}(Y_t^F) = \frac{G_{SL}^{1/2}}{Z_t\sigma R_{SL\,t}}$$ for the $\mathrm{S_L}$, and

(24) $$J_{SB}(Y_t^F) = \frac{G_{SB}^{1/2}}{Z_t\sigma R_{SB_t}(1 - R_{SB_t})}$$ for the $\mathrm{S_B}$ distribution.

The pdfs for the $\mathrm{S_U}$, the $\mathrm{S_L}$ and the $\mathrm{S_B}$ variables are hence obtained by substituting (16) and (22), (17) and (23), and (18) and (24), into equation (15), respectively.

Following standard procedure, the log-likelihood functions to be maximized in order to estimate the parameters of each of these three distributions are obtained by taking the natural logarithms of the corresponding pdfs and adding over the t = 1,…,T observations:

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

where:

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

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

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

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

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

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

for the S_U, S_L and S_B distributions, respectively; $G_t > 0$; and G_{SU}, R_{SUt}, G_{SL}, R_{SLt}, G_{SB} and R_{SBt} are as defined in equations (7) to (9) and (19) to (21).

An adjustment that facilitates estimation and interpretation is re-defining the distributional shape parameters as follows: for the S_U $\gamma=-\mu$, for the $S_{\mathbf{B}}$ $\gamma=\mu$, and for all three families $\delta=1/\theta$. 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 $\mu<0$, $\mu=0$ and $\mu>0$ 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: $\theta=\mu=0$. Finally, for the purposes of estimation, the following parameter range restrictions are recommended: for the S_U $0<\theta<1.5$ and $-15<\mu<15$; for the S_B $0<\theta<100$ and $-7.5<\mu<7.5$; and for the S_L $0<\theta<1$.

A Multivariate Johnson System Model

Another advantage of the Johnson system is that, because its three distributions originate from transformations of normal random variables, a multivariate form involving one, two or all three of the Johnson system distributions can be obtained on the basis of a multivariate normal distribution. This is important because many applications involve estimating and simulating joint yield distributions from several farms, counties, regions, commodities, etc. which are correlated with each other and can not a priory be assumed to exhibit the same distributional shape parameters or even follow the same distribution.

Because of the previously discussed flexibility of the expanded Johnson system, its multivariate arrangement allows for a direct estimation of the correlations between a set of non-normal random variables even when the probability distributions associated with those variables are markedly different, i.e. each of them may exhibit any theoretically feasible MVSK combination. In addition, likelihood ratio (LR) tests can be conducted to ascertain if the means and/or variances of some or all of the variables follow the same time trends or are similarly affected by changes in other exogenous factors. LR tests can also be used to evaluate if sub-sets of these variables that are best characterized by a particular family of distributions (S_U, S_L or S_B) exhibit exactly the same shape parameters γ and δ, i.e. identical skewness and kurtosis levels.

This is particularly important when estimating yield distributions because the data available at the country, regional, and particularly at the individual county and farm levels are often fairly short time series. Therefore, precise estimation of mean and variance trends and of the shape parameters determining the distributional skewness and kurtosis, is often not possible with univariate models. As exemplified in the

following applications section, this issue can be addressed by consolidating univariate models into a parsimonious multivariate Johnson system model that is more statistically efficient in using the limited crop yield information available for estimation. The log-likelihood function needed to estimate this model is derived next.

For a model with M variables that are contemporaneously correlated the (MT*x*MT) correlation matrix is:

(29) $\Omega = \Sigma \otimes \mathrm{I}_T$,

where I_T is a T*x*T is identity matrix, $\otimes$ is the Kroenecker product operator and Σ is the following M*x*M matrix containing the correlations between the random variables corresponding to each of the M equations:

(30) $$\Sigma = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1M} \\ \rho_{21} & 1 & \cdots & \rho_{2M} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{M1} & \rho_{M2} & \cdots & 1 \end{bmatrix}$$

Then, the vector of normally distributed random variables underlying the proposed non-normal pdf models can be expressed as $N_M \sim N(-\frac{\gamma_M}{\delta_M}, \delta_M^{-2}, \Sigma)$, where γ_M and δ_M are (M*x*1) vectors of parameters, ___ indicates an element-by-element vector division, and Σ is the previously discussed correlation matrix. The joint probability density function for the random vector N_M for any observation t is:

(31) $$P_N(N_M) = (2\pi)^{-M/2}(\prod_{j=1}^{M}\delta_j)|\Sigma|^{-1/2}\exp[-0.5\{(N_M + \frac{\gamma_M}{\delta_M})\delta_M\}'\Sigma^{-1}\{(N_M + \frac{\gamma_M}{\delta_M})\delta_M\}],$$

Following the multivariate transformation technique (Mood, Graybill and Boes 1974), the joint probability density function for the non-normal random variable

vector Y_M^F is obtained by applying any of the three previously discussed (S_U, S_L or S_B) transformations {equation (10), (11) or (12), and (13)} to each of the elements of N_M :

$$P_M(Y_{Mt}^F) = \left|J(Y_{Mt}^F)\right| \times P_N(q_M^{-1}(Y_{Mt}^F)), \tag{32}$$

where P_M is the multivariate normal density defined in equation (34), $q_M^{-1}(Y_{Mt}^F)$ is an M*x*1 vector of inverse (S_U, S_L or S_B) transformations from the elements of N_M into the elements of Y_{Mt}^F, and t denotes the fact that, because all of the transformations involve equation (13), the resulting multivariate non-normal density will exhibit different mean and variance vectors over time.

Also in (32) $J(Y_{Mt}^F)$ is the M*x*M Jacobian matrix of the vector transformation:

$$J(Y_{Mt}^F) = \begin{vmatrix} \frac{\partial N_1}{\partial Y_{1t}^F} & \frac{\partial N_1}{\partial Y_{2t}^F} & \frac{\partial N_1}{\partial Y_{3t}^F} & \cdots & \frac{\partial N_1}{\partial Y_{Mt}^F} \\ \frac{\partial N_2}{\partial Y_{1t}^F} & \frac{\partial N_2}{\partial Y_{2t}^F} & \frac{\partial N_2}{\partial Y_{3t}^F} & \cdots & \frac{\partial N_2}{\partial Y_{Mt}^F} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial N_M}{\partial Y_{1t}^F} & \frac{\partial N_M}{\partial Y_{2t}^F} & \frac{\partial N_M}{\partial Y_{3t}^F} & \cdots & \frac{\partial N_M}{\partial Y_{Mt}^F} \end{vmatrix} = \begin{vmatrix} \frac{\partial N_1}{\partial Y_{1t}^F} & 0 & 0 & \cdots & 0 \\ 0 & \frac{\partial N_2}{\partial Y_{2t}^F} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{\partial N_M}{\partial Y_{Mt}^F} \end{vmatrix} = \prod_{j=1}^{M} \frac{\partial N_j}{\partial Y_{jt}^F} \tag{33}$$

Therefore, the multivariate nonnormal density function for time period t is:

$$P_M(Y_{Mt}^F) = \{\prod_{j=1}^{M} \frac{\partial N_j}{\partial Y_{jt}^F} \delta_j\} (2\pi)^{-M/2} \left|\Sigma\right|^{-1/2} \exp[-0.5(\delta_M q_M^{-1}(Y_{Mt}^F) + \gamma_M)' \Sigma^{-1} (\delta_M q_M^{-1}(Y_{Mt}^F) + \gamma_M)] \tag{34}$$

where the specific Jacobian derivative ($\frac{\partial N_j}{\partial Y_{jt}^F}$) and inverse transformation functions $\{q_M^{-1}(Y_{Mt}^F)\}$ to be substituted in will depend on which nonnormal (S_U, S_L and/or S_B) family of distributions is to be assumed as a model for each of the j=1,…,M random variables under analysis.

The likelihood function that has to be maximized in order to estimate a multivariate non-normal pdf model involving the S_U, S_L and/or S_B families is obtained by taking the natural logarithm of equation (34) and adding across all T observations.

Exemplary Applications

Gauss 6.0 programs to estimate the parameters of each of the three expanded Johnson system distributions on the basis of sample data, as well as a multivariate estimation program, have been developed and will be made available by the authors upon request. Given estimated or assumed parameter values, these programs also compute the implied mean(s), variance(s), skewness and kurtosis, and simulate draws from the estimated or assumed distribution(s), which may be used for economic risk analysis.

Parametric models of farm-level corn yields based on the expanded Johnson system (i.e. the S_U, the S_B and the S_L distributions) are estimated using those programs. The yield data, obtained from the University of Illinois Endowment Farms database, included 26 corn farms located in twelve counties across that State. Data are available from 1959 to 2003, with a sample size varying from 20 to 45. The mean and standard deviations are specified as second and first degree polynomial functions of time, i.e.:

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

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

Thus, with the exception of the S_L in which μ=0; all univariate non-normal models initially include seven parameters (β_0, β_1, β_2, σ_0, σ_1, θ and μ). Normal models with the same mean and standard deviation specifications are estimated for comparison.

Preliminary examination of the results reveals that the maximum values reached by the log-likelihood functions (MLLFV) associated with the S_L models are lower than

the MLLFV corresponding to the S_U and S_B models in all 26 cases. In fact, the null hypothesis of normality (Ho: $\theta=0$) is not rejected in any of the 26 S_L models ($\alpha=0.1$). 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 2005) and the S_L distribution only allows for positive skewness (figure 1). Therefore, the S_L results are excluded form the following discussion.

Select statistics about the estimated S_U, S_B and normal models are presented in Table 1. The S_B model shows a higher MLLFV than the S_U model in 19 of the 26 cases. Likelihood ratio (LR) tests reject the null hypothesis of normality (Ho: $\theta=\mu=0$) in 17 of the 26 S_U models and in 18 of the 26 S_B models as well ($\alpha=0.10$). However, when the S_U or the S_B model with the highest MLLFV is selected as the most suitable non-normal model, Ho: $\theta=\mu=0$ is rejected 20, 14 and nine out of 26 times at the 0.10, 0.05 and 0.01 significance levels, respectively. Also note that five of the six non-rejections of normality ($\alpha=0.10$) correspond to the smaller ($T\leq30$) sample sizes.

Out of the 20 cases that are classified as non-normal, the S_B models exhibit the highest MLLFV in 14 cases and the S_U models in six cases (table 1). The MLLFV differences between the estimated S_U and S_B models corresponding to each of the 20 non-normal yield distribution cases range from near zero up to 4.23 units, with 14 being in excess of 0.5 units, eight larger than one unit, and five exceeding two units. The cumulative distribution functions (CDFs) implied by the estimated S_U and S_B models for farms b, s, o, k, n and a are derived on the basis of simulated yield data (n=10 million) in order to assess the empirical relevance of such range of MLLFV differences.

Two statistics, AD and MD, are computed from these CDFs. AD is the average of 125 vertical percentage distances between the CDF with the highest MLLFV and the other. 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). MD refers to the maximum of those 125 vertical distances.

The S_U and the S_B models for farm b exhibit nearly identical MLLFVs. Accordingly, both the average and the maximum vertical percentage differences between the CDFs implied by these two models (AD=0.06%, MD=0.23%) are negligible. In the case of farm s, the S_U model's MLLFV is only 0.58 units higher than the S_B's. The average and maximum vertical percentage differences (AD=1.04% and MD=3.25%) are considerably higher in this case. The MLLFV for the S_B model corresponding to farm o is 1.14 units higher that the S_U's, which results in an AD of 1.58% and a MD of 4.14%. A MLLFV difference of 2.09 units (farm k) is associated with even larger (2.48% and 5.99%) average and maximum CDF differences (figure 3). Larger (2.31 and 3.05) MLLFV differences (farms n and a, respectively) produce more extreme average (2.47% and 3.64%) and maximum (8.17% and 19.58%) vertical CDF differences.

In short, the larger MLLFV differences between the S_B and the S_U models in this application do translate into substantial discrepancies in these models' probabilistic yield predictions. Also note that the four largest MLLFV differences correspond to models estimated on the basis of relatively large ($T \geq 43$) sample sizes, which suggests that additional meaningful differentials could be observed if all sample sizes were of at least this magnitude. In principle, this confirms the need to consider both the S_B and the S_U as potential crop yield distribution models in any particular application.

The S-K combinations corresponding to the S_U or to the S_B model with the highest MLLFV for each of the yield samples analyzed are presented in figure 3. Three of the estimated S_U distributions (farms a, i and x) exhibit quite large (>50) kurtosis values and are thus not shown in figure 3. The S-K combinations of the remaining 17 non-normal distributions stretch from fairly low to relatively high S-K value combinations. In fact, 15 of those 17 can be grouped into three categories. Category A includes seven farm yield distributions with low or negative kurtosis and low negative skewness. Category B encompasses five distributions with moderate levels of positive kurtosis and negative skewness. Category C involves three yield distributions with higher levels of positive kurtosis and negative skewness.

The previously discussed multivariate estimation methods can be used evaluate if or to which extent the yields corresponding to the farms in each of those categories could be adequately represented by S_U and/or S_B distributions with the same shape (θ and μ) parameter values, i.e. exhibiting identical skewness and kurtosis levels. They can also be used to assess if the functions modeling the means and standard deviations of those yield distributions are identical or at least share some common parameter values. The contemporaneous correlations between the farm yield distributions in each of those categories can be simultaneously estimated in the process.

Table 2 two summarizes the results of applying these multivariate estimation methods to category B in figure 3, which includes farms e, g, q, d and s. The univariate models include seven parameters for each farm yield distribution, for a total of 35. The initial multivariate model adds 10 yield correlation coefficients and therefore has 45 parameters. The MLLFV of the multivariate model (-738.43) is 47.68 units higher than

the sum of the MLLFVs for the five univariate models (-786.11). A LTRS of 2x47.68=95.26 ($\chi^2_{(10,0.005)}$=25.2) easily rejects the null hypothesis of yield distribution independence at the 99.5% level; which is a strong argument for using multivariate estimation in this case.

Theoretically, the information that is transferred across the univariate models through the correlation matrix (equation 30) makes the multivariate model more statistically efficient. Estimation efficiency gains can also be obtained by using the multivariate estimation framework to eliminate all statistically redundant parameters. This is achieved by comparing the estimates for parameters with similar roles (i.e. the estimates for β_0, β_1, β_2, σ_0, σ_1, θ and μ across the five yield distributions, as well as the covariance parameters) and setting them equal if they are within one standard error estimate of each other. Each equality restriction is evaluated through likelihood ratio tests (α=0.20 to reduce the probability of incorrectly accepting the restriction). In addition, all statistically insignificant parameters (α=0.10) are set equal to zero.

The final multivariate model obtained through the previously described process is presented in table 2. In that model, all five distributions share the non-normality parameter (θ), while the three S_B models exhibit the same μ parameter value. μ is not statistically significant in the two S_U distributions and, therefore is set to zero. Models g and d and models q and s share intercept parameters (i.e. $\beta_{0g}=\beta_{0d}$ and $\beta_{0q}=\beta_{0s}$). Other mean equation restrictions include $\beta_{1q}=\beta_{1d}=\beta_{1s}$ and $\beta_{2q}=\beta_{2d}$; that is, the yields from farms q and d are found to follow identical time trends which, in turn, are quite similar to the trend in farm s yields.

In addition the intercepts of the standard deviation equations are found to be the same for farms g, q, d and s (i.e. $\sigma_{0g}=\sigma_{0q}=\sigma_{0d}=\sigma_{0s}$). Since the slope parameters in those equations are all statistically insignificant, it is concluded that the yield distributions for farms g, q, d and s exhibit the same constant variance over time. Finally, at 0.70, the correlations between yields from farms q and d and q and s, are found to be equally high; while the correlations between farms e and q and e and s are similarly low (0.26). At 0.46, the six remaining correlations are found to be equally moderate (table 2).

The 29 parameter restrictions imposed to the initial 45-parameter multivariate model reduces its MLLFV from -738.43 to -749.34. A likelihood ratio test (LRTS= 2x(749.34-738.43)=21.82 compared to $\chi^2_{(29,0.25)}$=33.7) does not reject the overall set of restrictions leading to the final multivariate model even at an α of 0.25. The highly parsimonious (16-parameter) final model makes the best use of the available yield information for the purposes of parameter estimation. As a result, on average, the standard error estimates in the final multivariate model are about half the size of those in the univariate models; and 20 of the 23 standard error estimates are lower in the final multivariate model than in the univariate models (table 2).

In short, the final multivariate model is substantially more reliable than the initial set of five univariate models and should provide for am improved representation and a more realistic simulation of these five yield distributions for the purposes if risk analysis.

Conclusions and Recommendations

The expanded form of the Johnson system advanced in this manuscript can model any theoretically possible combination of the first four central moments of a random variable. That is, the proposed system can accommodate any mean-variance-skewness-kurtosis

combination exhibited by a yield, price or any other distribution that may be encountered in practice. None of the probability distribution models previously discussed in the literature come close to achieving such property. The three families in this system nest the normal density which facilitates testing for non-normality. Also because these families are obtained from three alternative transformations to normality, it is possible to specify the system in a multivariate form. All of these characteristics are highly desirable for the applied modeling and simulation of probability distributions.

Estimation of the parameters of any of the three expanded Johnson system distributions, both in a univariate and a multivariate context, can be accomplished using the Gauss programs that have been developed and are available from the authors. Given parameter values, these programs also compute the means, variances, skewness and kurtosis, and simulate draws from the distribution(s) for use in economic risk analyses.

An application involving Illinois farm-level corn yields illustrates the estimation, characteristics and use of the proposed system. Normality is rejected in 20 of the 26 farm yield samples in the analysis, with non-rejection being clearly associated with the smaller sample sizes. Although the yield data analyzed is from the same state and crop, the skewness and kurtosis combinations implied by the best fitting non-normal models extend over a large region of the S-K plane, corresponding to both the S_U and the S_B families. Substantial, empirically relevant differences between the CDFs implied by the estimated S_U and S_B models are found in the several cases where their MLLFVs differ by relatively larger magnitudes. Statistically, this suggests that one of the two models is likely inferior to the other in those cases, which corroborates the need for probability distribution models that can span larger regions of the S-K space.

Theoretically, it is known that the most commonly used parametric models based on the Beta and the Gamma distributions span S-K regions that are far more restrictive than the S_B's and totally preclude the S_U's (figure 1). In addition, the application in this manuscript shows that the multivariate estimation capabilities afforded by the Johnson system can dramatically improve model quality and provide valuable information about the correlations among the variables of interest. Therefore is recommended that the expanded Johnson system is considered for use in future empirical work.

However, future research is needed to ascertain whether the expanded Johnson system's allowing for all theoretically possible mean-variance-skewness-kurtosis combinations is indeed sufficient to ensure a highly accurate representation of the stochastic behavior of any biological or economic variable of interest. If this is proven to be the case, there would be no need to consider any other probability distribution but the proposed system for the modeling and simulation of continuous random variables.

References

Anderson, J.R. 1974. "Simulation: Methodology and Application in Agricultural Economics." *Review of Marketing and Agricultural Economics* 42:3-55.

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.

Horowitz, J.L. and S. Lee. 2002. "Semiparametric Methods in Applied Econometrics: Do the Models Fit the Data?" *Statistical Modeling* 2(1):3-22.

Johnson, N.L. 1949. "System of Frequency Curves Generated by Method of Translation." *Biometrika* 36:149-176.

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.U. McDonald. 2006. "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.

Yatchew, A. 1998. Nonparametric Regression Techniques in Economics." *Journal of Economic Literature* Vol.XXXYI:669-721.

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

Farm Label	Sample Size	S_U MLLFV	S_B MLLFV	$\|S_U-S_B\|$ MLLFV	Normal MLLFV	LRTS	Final Model
a	44	-183.62	-186.67	3.05	-191.64	16.03^{3}	S_U
b	32	-123.81	-123.81	0.00	-134.94	22.27^{3}	S_B
c	44	-186.38	-182.15	4.22	-187.61	10.91^{3}	S_B
d	43	-189.23	-189.39	0.16	-192.55	6.63^{2}	S_U
e	25	-108.09	-108.00	0.08	-112.23	8.45^{2}	S_B
f	27	-128.31	-127.08	1.23	-128.98	3.81^{0}	N
g	31	-133.58	-133.57	0.00	-140.68	14.22^{3}	S_B
h	34	-161.15	-160.20	0.95	-161.80	3.20^{0}	N
i	43	-181.27	-184.84	3.58	-185.62	8.71^{2}	S_U
j	32	-145.96	-145.94	0.02	-149.20	6.53^{2}	S_B
k	27	-120.75	-118.66	2.09	-126.11	14.90^{3}	S_B
l	29	-132.56	-132.49	0.06	-132.56	0.13^{0}	N
m	37	-169.08	-169.00	0.09	-171.97	5.93^{1}	S_B
n	45	-197.46	-195.15	2.31	-197.47	4.64^{1}	S_B
o	42	-189.54	-188.40	1.13	-194.36	11.92^{3}	S_B
p	42	-195.34	-195.28	0.06	-197.77	4.97^{1}	S_B
q	40	-174.07	-173.55	0.51	-178.18	9.26^{3}	S_B
r	33	-145.36	-145.47	0.11	-150.09	9.46^{3}	S_U
s	40	-181.77	-182.35	0.58	-184.12	4.70^{1}	S_U
t	29	-131.07	-131.05	0.02	-133.79	5.47^{1}	S_B
u	44	-201.83	-201.21	0.61	-204.01	5.60^{1}	S_B
v	29	-127.78	-126.34	1.45	-131.64	10.61^{3}	S_B
w	29	-131.22	-131.24	0.02	-132.56	2.67^{0}	N
x	20	-93.45	-93.96	0.51	-98.42	9.94^{3}	S_U
y	29	-135.14	-135.00	0.14	-136.90	3.80^{0}	N
z	30	-143.92	-143.26	0.66	-144.92	3.32^{0}	N

Notes: MLLFV stands for the maximum log-likelihood function value, $|S_U-S_B|$ MLLFV refers to the absolute value of the S_U-S_B MLLFV difference, 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. Parameter and Standard Error Estimates for the Univariate, and for the Initial and Final Multivariate Johnson System Yield Distribution Models for Farms e, g, q, d and s

	Univariate Models		Initial Multivariate		Final Multivariate	
Parameter	**Par. Est.**	**S.E. Est.**	**Par. Est.**	**S.E. Est.**	**Par. Est.**	**S.E. Est.**
β_{0e}	63.075	5.062	67.524	6.192	66.684	3.909
β_{1e}	2.500	1.011	2.107	1.304	3.247	0.488
$100x\beta_{2e}$	6.991	4.177	6.958	5.676	0.000	.
σ_{0e}	11.832	2.186	8.030	4.045	5.580	3.054
σ_{1e}	0.833	0.171	0.958	0.335	1.397	0.392
μ_e	-1.717	3.110	-2.434	1.229	-5.067	2.316
θ_e	0.919	0.912	0.654	0.256	0.562	0.067
β_{0g}	114.918	9.817	111.459	19.039	111.162	3.169
β_{1g}	3.700	0.833	4.543	2.604	4.466	0.614
$100x\beta_{2g}$	-6.065	1.366	-9.295	6.783	-9.241	2.319
σ_{0g}	22.610	5.512	20.034	13.672	21.760	2.114
σ_{1g}	0.000	.	0.033	0.922	0.000	.
μ_g	-3.443	15.033	-2.378	4.837	-5.067	2.316
θ_g	0.718	1.182	0.794	0.821	0.562	0.067
β_{0q}	121.588	10.021	130.342	8.502	130.715	3.337
β_{1q}	1.597	1.161	0.522	0.902	0.683	0.170
$100x\beta_{2q}$	-0.744	2.925	1.444	1.825	1.029	0.288
σ_{0q}	21.487	13.018	21.310	3.215	21.760	2.114
σ_{1q}	0.018	0.485	0.000	.	0.000	.
μ_q	-1.911	0.657	-1.907	0.294	-5.067	2.316
θ_q	0.851	0.356	0.748	0.173	0.562	0.067
β_{0d}	97.446	6.364	102.540	6.960	111.162	3.169
β_{1d}	1.333	0.746	0.757	0.892	0.683	0.170
$100x\beta_{2d}$	0.906	1.873	1.884	1.998	1.029	0.288
σ_{0d}	15.302	2.573	14.883	5.766	21.760	2.114
σ_{1d}	0.273	0.055	0.279	0.243	0.000	.
μ_d	-1.889	3.767	-1.010	2.187	0.000	.
θ_d	0.446	0.398	0.416	0.393	0.562	0.067
β_{0s}	120.137	6.969	129.723	8.275	130.715	3.337
β_{1s}	1.833	0.457	0.743	1.108	0.683	0.170
$100x\beta_{2s}$	-2.208	0.551	-0.387	2.632	0.000	.
σ_{0s}	21.738	4.644	23.488	10.967	21.760	2.114
σ_{1s}	0.159	0.055	0.229	0.344	0.000	.
μ_s	-0.982	0.877	-0.683	0.381	0.000	.
θ_s	0.614	0.334	0.829	0.283	0.562	0.067

Table 2 (continued). Parameter and Standard Error Estimates for the Univariate, and for the Initial and Final Multivariate Johnson System Yield Distribution Models for Farms e, g, q, d and s

	Univariate Models		Initial Multivariate		Final Multivariate	
Parameter	**Par. Est.**	**S.E. Est.**	**Par. Est.**	**S.E. Est.**	**Par. Est.**	**S.E. Est.**
ρ_{eg}	0.000	.	0.395	0.155	0.462	0.076
ρ_{eq}	0.000	.	0.293	0.153	0.262	0.124
ρ_{ed}	0.000	.	0.465	0.132	0.462	0.076
ρ_{es}	0.000	.	0.189	0.160	0.262	0.124
ρ_{gq}	0.000	.	0.588	0.106	0.462	0.076
ρ_{gd}	0.000	.	0.557	0.114	0.462	0.076
ρ_{gs}	0.000	.	0.532	0.147	0.462	0.076
ρ_{qd}	0.000	.	0.662	0.092	0.703	0.048
ρ_{qs}	0.000	.	0.789	0.069	0.703	0.048
ρ_{ds}	0.000	.	0.487	0.126	0.462	0.076

Notes: The parameter and standard error estimates for β_{2e}, β_{2g}, β_{2q}, β_{2d} and β_{2s}, have been multiplied times 100. The dots indicate that the standard error estimates are not computed since the parameter estimates have been set equal to zero.

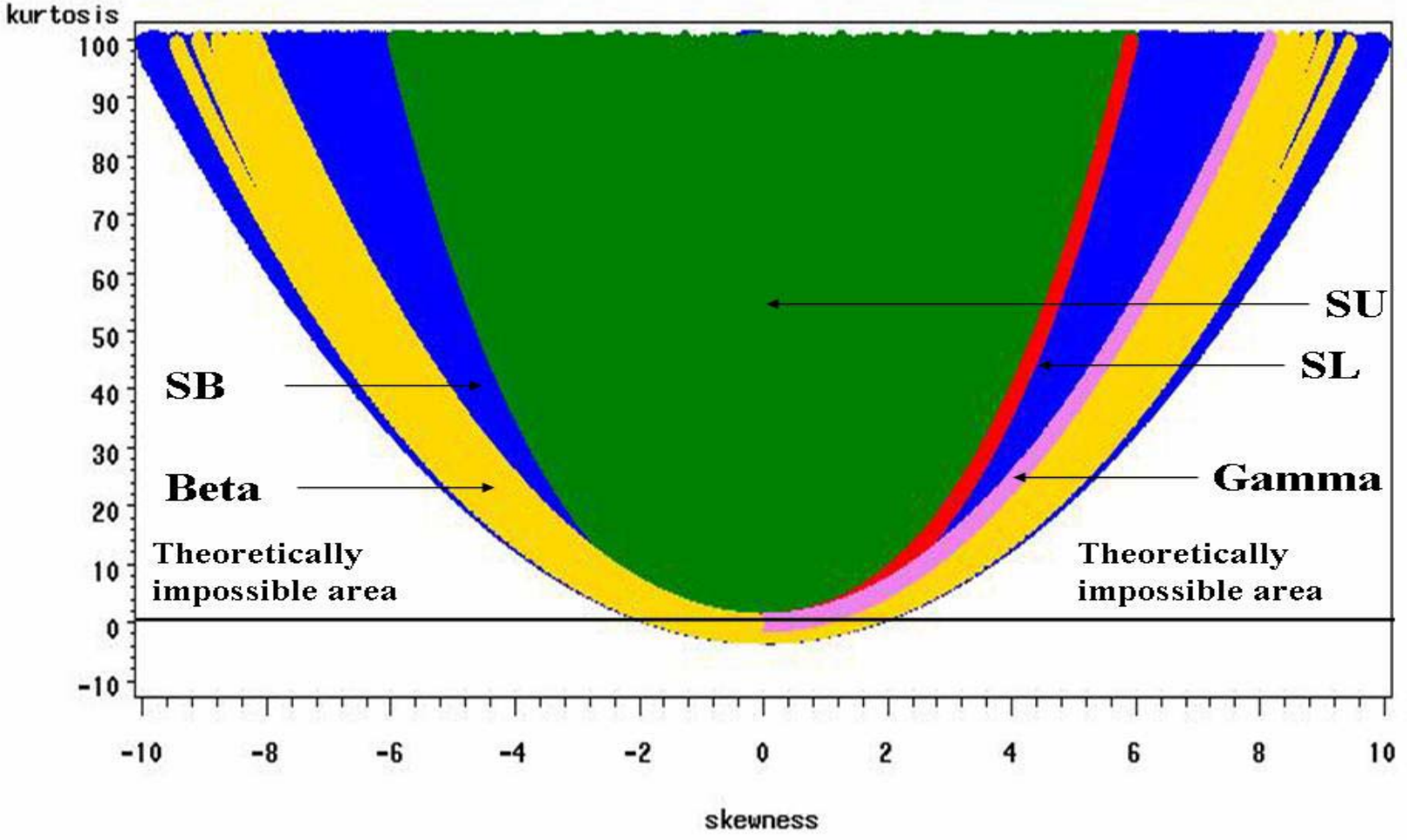

Figure 1. S_U, S_L, S_B, Beta and Gamma distributions in the S-K plane

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

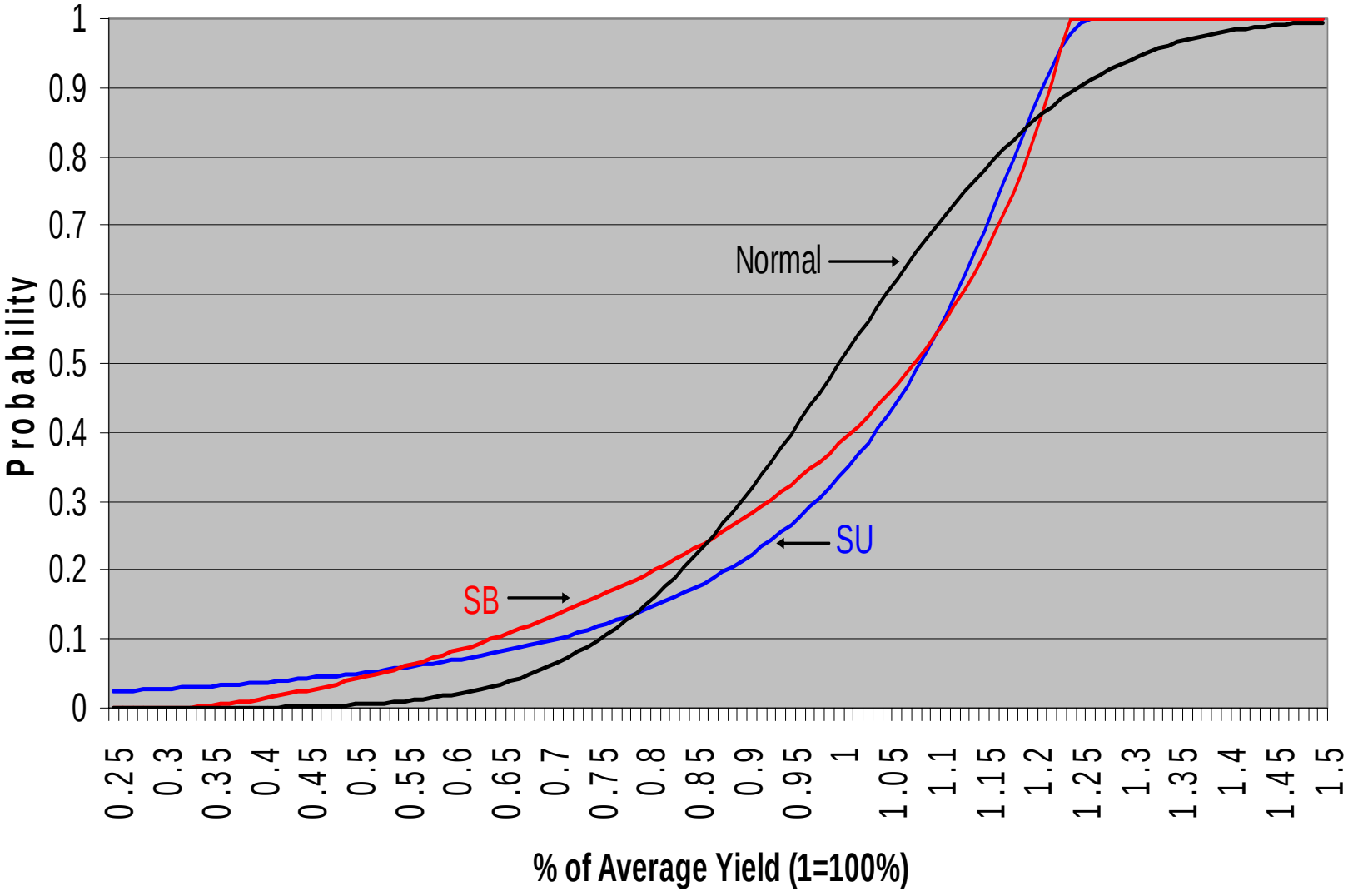

Figure 2. CDFs from SU, SB and normal yield models for farm K

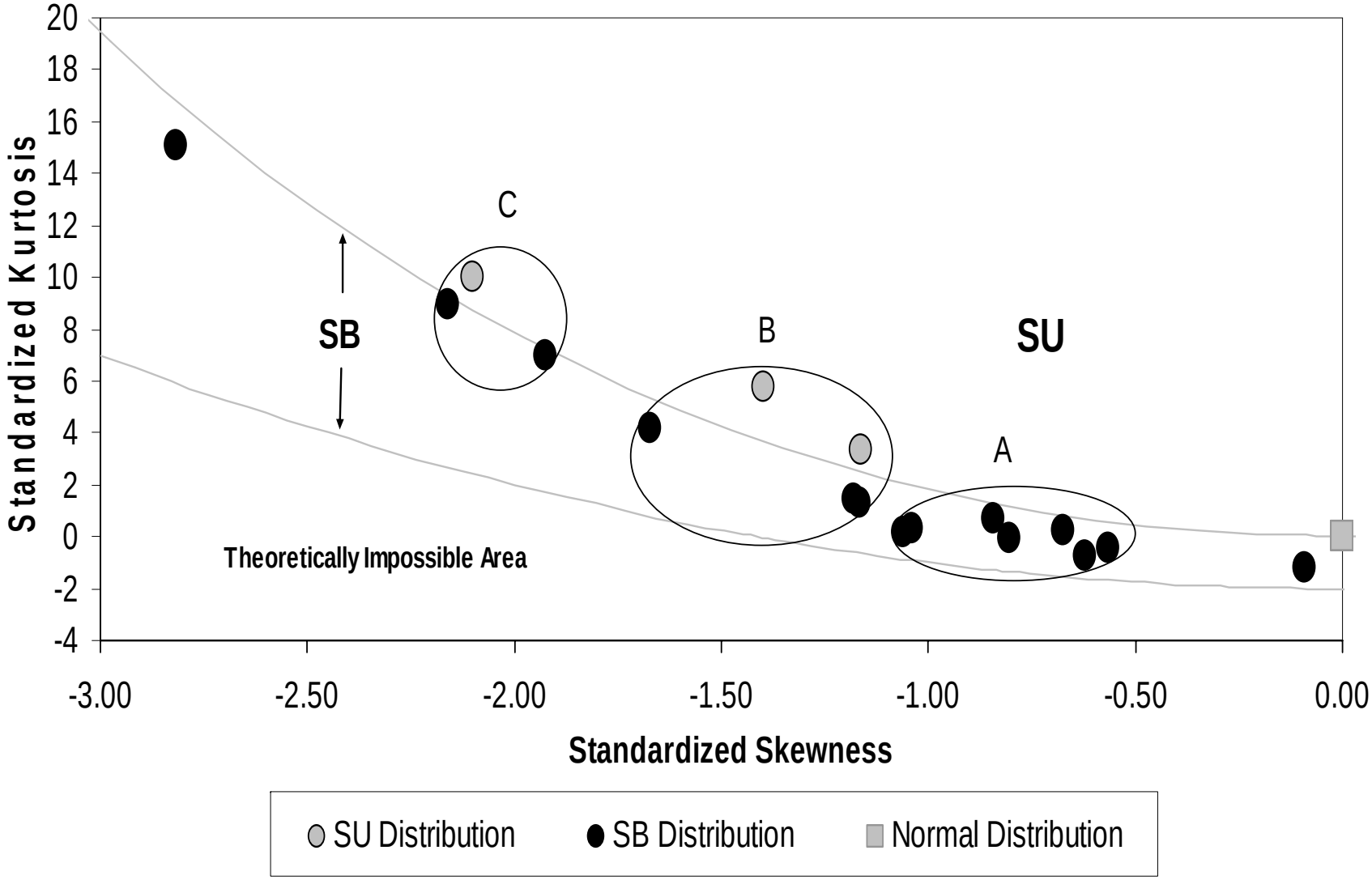

Figure 3. Skewness-kurtosis combinations of non-normal models