Distribution Choice Under Null Priors and Small Sample Size

Paul A. Feldman, James W. Richardson, and Keith D. Schumann
Agricultural and Food Policy Center
Department of Agricultural Economics
Texas A&M University
College Station, Texas 77843-2124
Phone: (979) 845-8014
Fax: (979) 845-3140
Email: paulf@tamu.edu

Selected Paper prepared for presentation at the American Agricultural Economics Association Annual Meeting, Denver, Colorado, August 1-4, 2004

Introduction

Modeling economic systems often involves making assumptions about how data are distributed. Given the nature of economic problems the task of determining how the salient data are distributed is often a difficult one. There is a wealth of literature on how key economic variables are distributed and the findings have conflicted and been at best inconclusive. One of the most notable discussions in Agricultural Economics is the question of how crop yields are distributed. For the past forty years empirical studies have been published in various journal discussing the distributions of crop yields, including contributions from Atwood, Shaik, and Watts; Day; Gallagher; Goodwin and Kerr; Just and Weninger; Moss and Shonkwiler; and Ramirez, Misra, and Field.

For the discussion on yield distributions there is a wealth of data; however, as with most economic problems the issues surrounding crop yields are complex making the analysis difficult. Given the complexity of economic issues and the difficulty of making distributional assumptions when dealing with reasonable sample sizes, it is an even greater challenge to estimate distributions for economic variables when the data is scarce. The problem of small sample sizes is a common one when dealing with economic data, which creates problems when making statistical inferences. D'Agostino and Stephens suggest that to achieve a reasonable power with a goodness-of-fit test samples sizes should not be less than twenty observations. This is often a luxury that economists do not have.

Quite often an economist's only tool in choosing the appropriate distribution for data is their knowledge about the system from which the data comes. This paper looks at the problem of distribution selection from the viewpoint of total naiveté. We would like to know which distribution performs the best when there is no knowledge of how the data were generated. In a Bayesian sense we would like to see which distribution performs the best when our priors on distributional assumptions are the null set.

Methodology

To answer the question of which distribution performs the best we have set up a Monte Carlo experiment. This experiment evaluates how robust a set of seven distributions are in estimating the true distribution of a random sample of data. The set of distributions $\underline{g(x)}$ that are evaluated are the Beta, $B(\alpha,\beta)$; Gamma, $Gm(\alpha,\beta)$; Logistic, $L(\mu,\sigma)$; Log-Log, $LL(\mu,\sigma)$; Lognormal, $Ln(\mu,\sigma)$; Normal, $N(\mu,\sigma)$; and Weibull, $W(\alpha,\beta)$. These distributions were chosen because they have been widely used to simulate economic data. Each of the distributions in $\underline{g(x)}$ are used to estimate the true distribution $f(x)$ taken from the set of distributions $\underline{f(x)}$ where the distributions in $\underline{f(x)}$ are the same set as in $\underline{g(x)}$.

A flowchart in Figure 1 maps out the steps of the experiment. The first step of the experiment is to select a distribution $f(x)$, such as Beta, to be evaluated from the set of distributions $\underline{f(x)}$. The distribution $f(x)$ is pre-specified with known parameters by the analyst with the only limitation being that the $P(x<0)<0.05$, Such as Beta(3,5). This limitation is in place to emulate the fact that most economic data is nonnegative. The second step is to generate a random sample $\underline{x_k}=\{x_1,x_2,\ldots,x_{10}\}$ of ten data points from the

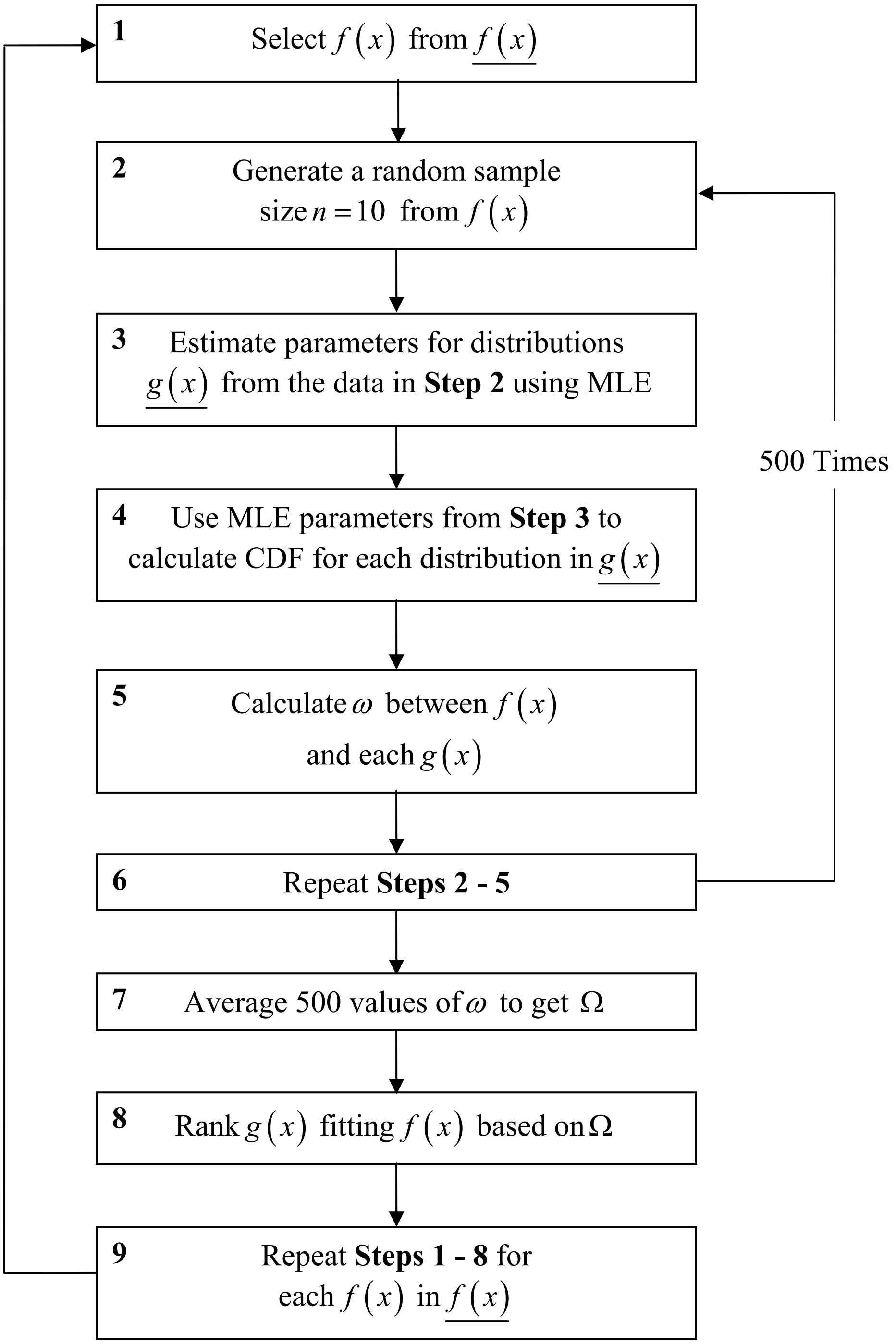
1 Select $f(x)$ from $\underline{f(x)}$
2 Generate a random sample size $n = 10$ from $f(x)$
3 Estimate parameters for distributions $\underline{g(x)}$ from the data in **Step 2** using MLE
4 Use MLE parameters from **Step 3** to calculate CDF for each distribution in $\underline{g(x)}$
5 Calculate ω between $f(x)$ and each $g(x)$
6 Repeat **Steps 2 - 5**
500 Times
7 Average 500 values of ω to get Ω
8 Rank $g(x)$ fitting $f(x)$ based on Ω
9 Repeat **Steps 1 - 8** for each $f(x)$ in $\underline{f(x)}$

Figure 1. Steps in the Procedure to Rank Distributions

distribution $f(x)$. Step three is to calculate the parameters for each distribution in $\underline{g(x)}$ given the sample $\underline{x_k}$ using the MLE method.

In step four the CDF's for each distribution in $\underline{g(x)}$ are calculated using the MLE's of their parameters. The CDF's are defined by 100 data points calculated using probabilities ranging from 0.05 to 0.95 at equal intervals evaluated with the inverse transform method.

In step five the CDF's for $\underline{g(x)}$, denoted $\underline{G(x)}$, are individually compared against the CDF of the true distribution $F(x)$ to evaluate how well the distributions in $\underline{g(x)}$ estimate $f(x)$. To make this comparison to we developed a goodness-of-fit criterion based on the empirical distribution function. The formula for the goodness-of-fit criterion is defined as

$$\omega = \sum_{i=1}^{n} \left[F\left(x_{(i)}\right) - G\left(x_{(i)}\right) \right]^2 w_i \tag{1}$$

where,

$F =$ CDF of the true distribution
$G =$ CDF of the estimated distribution
$i = i$th observation
$n =$ total number of observations
$x_{(i)} = i$th ordered random number
$w_i =$ tail weighted function defined as, $\dfrac{12i^2 - 12in + 3n^2}{n^3 + 2n}$

This formula is relatively straightforward in that it is based on the sum of squared differences between the true distribution and the estimated distribution. The tail weighting function w_i is used to reflect the importance of tail probabilities in economic modeling. The tail weighting function is based on the parabolic function

(2) $$w_i = a\left(i - \frac{n}{2}\right)^2$$

where a is defined such that

(3) $$\sum_{i=1}^{n} w_i = 1$$

From this w_i is derived by

(4) $$\begin{aligned}
w_i &= a\left(i - \frac{n}{2}\right)^2 \\
\sum_{i=1}^{n} w_i &= \sum_{i=1}^{n} a\left(i - \frac{n}{2}\right)^2 = 1 \\
a\sum_{i=1}^{n}&\left(i^2 - in + \frac{n^2}{4}\right) = 1 \\
a&\left[\frac{n(n+1)(2n+1)}{6} - \frac{n^2(n+1)}{2} + \frac{n^3}{4}\right] = 1 \\
a&\left(n^3 + 2n\right) = 12 \\
a &= \frac{12}{n^3 + 2n} \\
w_i &= \frac{12}{n^3 + 2n}\left(i - \frac{n}{2}\right)^2 \\
&= \frac{12i^2 - 12in + 3n^2}{n^3 + 2n}
\end{aligned}$$

Thus w_i amplifies the contribution to ω as $G(x)$ differs from $F(x)$ in the tails of the true distribution.

The CDF's of each distribution in $\underline{G(x)}$ is compared to $F(x)$ by calculating an ω for each $G(x)$ in the set $\underline{g(x)}$ as well as a linearly interpolated empirical distribution of the data. This gives a set $\underline{\omega_k} = \{\omega_E, \omega_B, \omega_{Gm}, \omega_L, \omega_{LL}, \omega_{Ln}, \omega_N, \omega_W\}$ for the *k*th sample $\underline{x_k}$ of $f(x)$. Step six is the repetition of steps two through five for $t = 500$ times giving 500 random samples for each $f(x)$. For each repetition a set of $\underline{\omega_k}$ is calculated from the random sample $\underline{x_k}$. In step seven an average is taken for each $\omega_{g(x)}$ as $\Omega_{g(x)} = \frac{1}{t}\sum_{k=1}^{t}\omega_{g(x)}$ giving a set $\underline{\Omega} = \{\Omega_E, \Omega_B, \Omega_{Gm}, \Omega_L, \Omega_{LL}, \Omega_{Ln}, \Omega_N, \Omega_W\}$. For step eight each $\Omega_{g(x)}$ in $\underline{\Omega}$ is ranked to evaluate which distribution $g(x)$ fitted $f(x)$ most accurately. The ninth and final step is to repeat this procedure for each $f(x)$ in $\underline{f(x)}$.

The selection of parameters for each $f(x)$ is done in a somewhat arbitrary manner. Three sets of parameters are chosen for each $f(x)$ in order to explore how differences in shape and scale of a distribution affect the accuracy of estimates $g(x)$ of $f(x)$. Table 1 contains the parameterization for each true distribution used in the analysis.

In economic analysis there may be the existence of data which come from a bimodal distribution. Therefore the list of distributions in $\underline{f(x)}$ was expanded to include mixture distributions. Two mixture distributions were added for each distribution

Table 1. Parameters for Distributions in $\underline{f(x)}$

	Distribution 1		Distribution 2		Distribution 3	
	Param 1	Param 2	Param 1	Param 2	Param 1	Param 2
Beta	3	3	3	5	5	3
Gamma	2	10	5	15	8	5
Logistic	140	5	140	15	140	25
Log-Log	80	10	80	25	80	40
Lognormal	3	0.5	4	0.5	5	0.5
Normal	100	15	100	25	100	35
Weibull	2	100	3	100	4	100

in $f(x)$. There was no cross distributional mixtures, i.e. each mixture distribution contained only two parameterized distributions of the same type. Each mixture distribution was sampled and estimated using the distributions in $g(x)$ using the procedure described above. Parameters were chosen for each distribution in a manner that would make it difficult if not impossible to determine from a small sample the modality of the true distribution. This ensured that the estimation of the mixture distributions using the unimodal distributions in $g(x)$ would be a reasonable procedure. Table 2 contains the parameterization for each true mixture distribution used in the analysis.

Results

The experiment for this paper was conducted in Microsoft Excel using the Simetar software tool (Richardson, Schumann, and Feldman). Random samples were generated using the Monte Carlo methods available in Excel. Parameter estimates were calculated using Simetar's MLE functions. Both Excel and Simetar functions were used to calculate the CDF's using the inverse transform method. The Simetar function

Table 2. Parameters for Mixture Distributions in $\boldsymbol{f(x)}$

	Mixture 1				Mixture 2			
	Distribution 1		Distribution 2		Distribution 1		Distribution 2	
	Param 1	Param 2	Param 1	Param 2	Param 1	Param 2	Param 1	Param 2
Beta	3	10	10	3	3	3	7	7
Gamma	5	5	10	10	5	10	10	5
Logistic	10	5	20	5	200	35	100	15
Log-Log	60	10	100	10	100	10	80	70
Lognormal	3	0.3	4	0.2	2	0.5	4	0.5
Normal	70	20	150	30	150	55	200	15
Weibull	2	5	5	1,000	2	1	5	10,000

CDFDev was used to calculate ω and the simulation engine in Simetar was used to generate the 500 repetitions.

Table 3 shows the rankings of $\underline{g\left(x\right)}$ when the true distribution $f\left(x\right)$ is Normal(μ,σ). The distributions $\underline{g\left(x\right)}$ are listed at the top of the table and the parameters selected when $f\left(x\right)$ is Normal are listed on the left side of the table. Each row is associated with a given set of parameters and contains the ranking of how well the distribution $g\left(x\right)$ fit the parameterized $f\left(x\right)$. The last row in the table is the overall ranking of the distributions in $\underline{g\left(x\right)}$ for all parameterizations of the Normal distribution. For the first and third parameterizations of a Normal distribution, $f\left(x\right)$, the Normal distribution, $g\left(x\right)$, was ranked number one and ranked second for one parameterization. Overall the Normal distribution was ranked number one for fitting data that were truly distributed Normal.

An overview of Table 3 shows the rankings of $\underline{g\left(x\right)}$ did not significantly change when the scale of $f\left(x\right)$ changed. Changes in rank of more than one position occurred only for the Empirical, Lognormal, and Weibull distributions when $f\left(x\right)$ was Normal. The Empirical ranged from being ranked fifth to eighth, the range for the Lognormal was fifth to ninth, and the Weibull ranged from first to fourth. The insensitivity in the ranking

Table 3. Rankings of $\underline{g\left(x\right)}$ for $f\left(x\right) = Normal\left(\mu,\sigma\right)$

Dist / Params	Empirical	Beta	Gamma	Logistic	Log-Log	Lognormal	Normal	Weibull
Normal(100,15)	8	7	3	2	9	5	1	4
Normal(100,25)	5	6	4	3	9	8	2	1
Normal(100,35)	5	7	4	3	8	9	1	2
Overall Rank	5	6	4	3	8	7	1	2

to changes in the scale of the distribution $f(x)$ was consistent for all distributions in $f(x)$ and was also the case for changes in the shape of the distribution $f(x)$.

Table 4 shows the overall rankings of $g(x)$ for all of the distributions in $f(x)$ when $f(x)$ is unimodal. In most cases the best estimator of $f(x)$ was the same distribution in $g(x)$, as indicated by the bold values in Table 4. Exceptions occurred for $f(x)=Gm(\alpha,\beta)$ where the rank of $g(x)=Gm(\alpha,\beta)$ was second, $f(x)=Ln(\mu,\sigma)$ where the rank of $g(x)=Ln(\mu,\sigma)$ was third, and $f(x)=B(\alpha,\beta)$ where the rank of $g(x)=B(\alpha,\beta)$ was fifth.

The overall ranking of distributions in $g(x)$ for all distributions in $f(x)$ show that the Weibull distribution fit $f(x)$ the best for all distributions in $f(x)$ (Table 4). The next best distribution was the Normal with the Beta distribution performing the worst.

Table 5 summarizes the rankings of $g(x)$ when the true distribution $f(x)$ is a mixture distribution. The shape, scale, and location of the distributions had a greater affect on the rankings of the mixture distributions than on the unimodal distributions.

Table 4. Rankings of $g(x)$ for All Unimodal $f(x)$

	Empirical	Beta	Gamma	Logistic	Log-Log	Lognormal	Normal	Weibull
Beta	4	**5**	6	3	7	8	2	1
Gamma	6	8	**2**	5	1	7	4	3
Logistic	6	8	4	**1**	7	5	2	3
Log-Log	7	8	2	4	**1**	3	5	5
Lognormal	7	8	2	5	1	**3**	6	4
Normal	5	6	4	3	8	7	**1**	2
Weibull	5	7	4	3	6	8	2	**1**
Overall Rank	6	8	3	4	5	7	2	1

The only distribution in the set $\underline{g(x)}$ that was unaffected by shape and scale was the Lognormal distribution, which performed poorly under all parameterizations.

The overall rankings for how well the distributions in $\underline{g(x)}$ estimated the mixture distributions showed significant changes in the ranking of some of the distributions. Although the Beta performed the worst when estimating the unimodal distributions, it was the best distribution when estimating the mixture distributions. The Weibull moved from the number one ranking in unimodal estimation to being ranked fourth in estimating the mixture distributions. The Normal distribution, however, was ranked second in estimating both the unimodal distributions and the mixture distributions.

Conclusion

When modeling an economic system it is of great importance to know how the variables in the system are distributed. Often times there are limited data for the variables in the system and it is difficult to make statistical inferences about how the variables are distributed. This makes knowledge of the system an important tool for the analyst when making distributional assumptions. We examined how well a set of distributions $\underline{g(x)}$ perform when there is limited information about the variable for which

Table 5. Rankings of $\underline{\boldsymbol{g(x)}}$ for All Mixture Distributions from $\boldsymbol{f(x)}$

	Empirical	Beta	Gamma	Logistic	Log-Log	Lognormal	Normal	Weibull
Beta	4	**3**	6	4	7	8	1	1
Gamma	5	3	**5**	2	1	8	3	5
Logistic	8	3	6	**3**	1	7	1	5
Log-Log	6	4	3	2	**7**	8	4	1
Lognormal	3	2	7	3	5	**8**	1	6
Normal	3	1	6	5	7	8	**3**	1
Weibull	2	1	7	4	5	8	3	**6**
Overall Rank	5	1	7	3	6	8	2	4

a distribution is being estimated.

The results of the study indicate that from the Empirical, Beta, Gamma, Logistic, Log-Log, Lognormal, Normal, and Weibull distribution the most robust distribution is the Normal distribution. The Normal distribution had the best overall performance in estimating the true distribution for both unimodal and mixture distributions. This result is not surprising given the central limit theorem and the assumption of normality made in many statistical techniques.

One interesting finding from this study is the performance of the empirical distribution of the data in estimating the true distribution. Because the empirical distribution is bounded by the data it regularly under estimated at least one of the tails of the true distribution. However, when irregularities were added by using mixture distributions the empirical distribution often performed better than in the unimodal case.

This study was a simple experiment to determine how well certain distributions estimate the true distribution. Even when examining the seven distributions used in this research there are literally millions of permutations for the parameterization of the true distribution which time constraints did not allow us to examine. If all of these permutations were examined the findings may differ.

References

Atwood, Joseph, Saleem Shaik, and Myles Watts. "Crop Yield Distributions" *American Journal of Agricultural Economics.* 85(2003):888-901.

D'Agostino, Ralph B. and Michael A. Stephens. *Goodness-of-fit Techniques*. New York: Marcel Dekker, 1986.

Day, Richard H. “Probability Distributions of Field Crop Yields.” *Journal of Farm Economics.* 47(1965):713-41.

Gallagher, Paul “U.S. Soybean Yields: Estimation and Forecasting with Nonsymmetric Disturbances.” *American Journal of Agricultural Economics.* 71(1987):796-803.

Goodwin, Barry K. and Alan P. Ker. “Nonparametric Estimation of Crop Yield Distributions: Implications for Rating Group Risk (GRP) Crop Insurance Contracts.” *American Journal of Agricultural Economics.* 80(1998)139-53.

Just, Richard E. and Quinn Weninger. "Are Crop Yields Normally Distributed?" *American Journal of Agricultural Economics.* 81(1999):287-304.

Moss, Charles B. and J.S. Shonkwiler. “Estimating Yield Distributions with a Stochastic Trend and Nonnormal Errors.” *American Journal of Agricultural Economics.* 75(1993):1056-62.

Ramirez, Octavio A., Sukant Misra, and James Field. “Crop-Yield Distributions Revisited.” *American Journal of Agricultural Economics.* 85(2003):108-120.

Richardson, James W., Keith D. Schumann, and Paul A. Feldman. "Simetar: Simulation for Excel to Analyze Risk." Mimeo. Texas A&M University, September 2000.