Kernal density functions to estimate parameters to simulate stochastic variables with sparse data: what is the best distribution?

J. W. Richardson[1], J. L. Outlaw[1] & K. Schumann[2]
[1]*Department of Agricultural Economics, Texas A&M University, USA*
[2]*Welch Consulting, USA*

Abstract

The purpose of this paper was to compare the goodness-of-fit for several parametric and kernal-based distributions to determine which distribution would perform well for simulating continuous random input variables whose underlying distributions were unknown. A Monte Carlo simulation procedure was developed to estimate how well some proxy distributions performed at approximating the distributions of random input variables. We conclude that without any *a priori* information on which to pick a probability distribution, the distribution for simulating a random input variable with limited specifications was a Parzen kernal distribution.

Keywords: probability distribution selection, kernal distributions, simulation, Simetar©.

1 What is the best probability distribution to simulate random input variables?

Risk analysts who use Monte Carlo simulation techniques must specify (or assume) a probability density function (PDF) for each random input variable in their models. The question of which PDF (normal, beta, gamma, Weibull, etc.) should be used is often suggested by familiarity with the data generation process or the type of problem being analyzed (Law and Kelton [1, pp. 155-216]). Alternatively, some researchers simply assume the random input variables follow a normal distribution due to the ease of parameter estimation for this distribution and rely on the Central Limit Theorem as a justification. Another option is to estimate parameters for several proxy parametric distributions and

doi:10.2495/RISK100011

select the distribution that has the "best" goodness-of-fit test statistic comparing a simulated distribution to the historical data (Palisade Corp [2]). This procedure is most appropriate when there are many independent and identically distributed (i.i.d.) observations in the data and the proxy distributions have the same or similar characteristics and support evidenced empirically by the data. However, it is often the case that there are few historical observations and therefore there exists a need to use a procedure that is reasonable in these circumstances.

An alternative is to use a kernal density function (KDF) to fit a distribution to the available observations (Parzen [3]; Silverman [4]; Chen [5]). Kernal density procedures provide flexible means to both approximate the unknown underlying distribution as well as accommodate the multi-modality that often accompanies sparse data. Silverman assesses that these procedures are sound but computationally beyond the capabilities of many analysts. There is extensive literature on kernal methods and the associated bandwidth selection; it is not the intent of this study to contribute to those discussions but to evaluate some of the more common methods under some basic specifications.

The purpose of this paper is to approximate the underlying distribution for random input variables using both parametric and kernal density functions for small samples with particular specified properties. The types of random variables presented have finite means and variances and are non-negative, such as price or production variables. A sample of 10 observations is used to represent the size of data sets often available with annual production variables of this type.

2 Methodology

A Monte Carlo simulation procedure was developed to estimate distribution functions and then systematically sample from them to determine which of the proxy distributions performed most favorably. The simulation experiment was programmed in a spreadsheet using the Simetar© add-in because it provides functions to estimate probability distribution parameters for stochastic samples and then simulate the distributions all in one pass (Richardson et al. [6]). Simulation and Econometrics to Analyze Risk: Simetar© is an Excel add-in for probability distribution parameter estimation, econometrics, forecasting, simulation, validation of simulation results, and ranking risky alternatives. For more details see www.simetar.com. Twelve parametric distributions were tested: beta, gamma, double exponential, exponential, logistic, log-log, log-logistic, log normal, normal, Pareto, Weibull, and uniform. In addition, ten kernel density functions were tested: Cauchy, cosinus, double exponential, Epanechnikov, Gaussian, Parzen, quartic, triangle, triweight, and uniform.

A flowchart of the procedure used to test the 22 PDFs is provided in Figure 1. The flowchart shows how the process proceeds for testing one random input variable. We start with a random variable, Y, with 10 positive values and then in Box 2 we estimated the parameters using the respective parameters from normal ($\tilde{Y}_N$), beta ($\tilde{Y}_B$), gamma ($\tilde{Y}_G$), uniform ($\tilde{Y}_U$), or Weibull ($\tilde{Y}_W$) maximum likelihood estimation (MLE). Several values were sampled from each distribution. These five distributions represent commonly used distributions for variables with positive values. Because we do not know the true distribution for Y so we simulate 10 sample values using the MLE parameters for the five PDFs and refer to these as the “parent” distributions (Figure 1, Box 3). The sample values for the five parent distributions in Box 3 are stochastic and change for every iteration based on their parameters in Box 2.

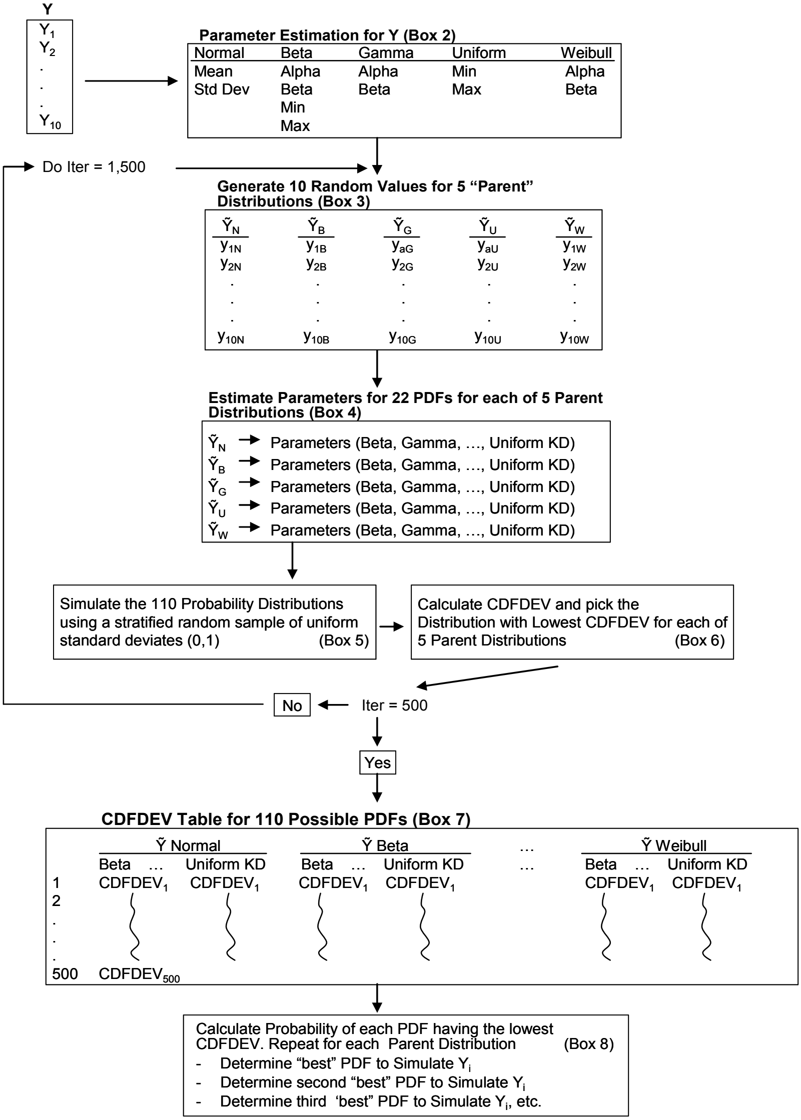
Y
Y_1
Y_2
Y_{10}
Parameter Estimation for Y (Box 2)
Normal
Mean
Std Dev
Beta
Alpha
Beta
Min
Max
Gamma
Alpha
Beta
Uniform
Min
Max
Weibull
Alpha
Beta
Do Iter = 1,500
Generate 10 Random Values for 5 “Parent” Distributions (Box 3)
$\tilde{Y}_N$ $\tilde{Y}_B$ $\tilde{Y}_G$ $\tilde{Y}_U$ $\tilde{Y}_W$
y_{1N} y_{1B} y_{aG} y_{aU} y_{1W}
y_{2N} y_{2B} y_{2G} y_{2U} y_{2W}
y_{10N} y_{10B} y_{10G} y_{10U} y_{10W}
Estimate Parameters for 22 PDFs for each of 5 Parent Distributions (Box 4)
$\tilde{Y}_N$ → Parameters (Beta, Gamma, …, Uniform KD)
$\tilde{Y}_B$ → Parameters (Beta, Gamma, …, Uniform KD)
$\tilde{Y}_G$ → Parameters (Beta, Gamma, …, Uniform KD)
$\tilde{Y}_U$ → Parameters (Beta, Gamma, …, Uniform KD)
$\tilde{Y}_W$ → Parameters (Beta, Gamma, …, Uniform KD)
Simulate the 110 Probability Distributions using a stratified random sample of uniform standard deviates (0,1) (Box 5)
Calculate CDFDEV and pick the Distribution with Lowest CDFDEV for each of 5 Parent Distributions (Box 6)
No
Iter = 500
Yes
CDFDEV Table for 110 Possible PDFs (Box 7)
$\tilde{Y}$ Normal
$\tilde{Y}$ Beta
…
$\tilde{Y}$ Weibull
Beta … Uniform KD
Beta … Uniform KD
…
Beta … Uniform KD
1 $CDFDEV_1$ $CDFDEV_1$ $CDFDEV_1$ $CDFDEV_1$ $CDFDEV_1$ $CDFDEV_1$
2
500 $CDFDEV_{500}$
Calculate Probability of each PDF having the lowest CDFDEV. Repeat for each Parent Distribution (Box 8)
- Determine “best” PDF to Simulate Y_i
- Determine second “best” PDF to Simulate Y_i
- Determine third ‘best” PDF to Simulate Y_i, etc.

Figure 1: Flow chart of methodology used to determine the best PDF for simulating a random input variable.

The next step is to calculate the parameters for the proxy parametric distributions and estimate the kernal distributions to be tested using the 10 stochastic iterations for each of the five assumed distributions (Box 4). As the sample values in Box 3 change for each iteration, the parameters for the 110 PDFs in Box 4 are automatically updated by Simetar©. The 110 distributions (22 PDFs * 5 parent distributions) were simulated using a stratified random sample of uniform standard deviates (Box 5).

A goodness-of-fit criteria value is calculated (Box 6) for each of the 110 PDFs by comparing the simulated distributions in Box 5 to their parent distributions in Box 3. In other words, the 22 simulated distributions that simulated the $\tilde{Y}_N$ sample in Box 5 are compared to the stochastic sample of 10 values in $\tilde{Y}_N$ in Box 3 to test how closely they matched the values for the parent distribution. This step is repeated to compare the 22 PDFs simulating the random sample values for the remaining parent distributions ($\tilde{Y}_B$, then $\tilde{Y}_G$, and $\tilde{Y}_U$ and $\tilde{Y}_W$).

The goodness-of-fit criteria selected for testing how closely the simulated PDFs compare to the parent distribution is a weighted cumulative distribution comparison function (CDFDEV) available in Simetar©. The CDFDEV criteria is calculated as the sum of the squared distance between two distribution functions with penalty weights increasing in value as the observations move away from the mean. If a simulated PDF is identical to the parent distribution, the CDFDEV value equals zero. When comparing two or more distributions as to their goodness-of-fit, the distribution with the smallest CDFDEV is the “best.”

As indicated in Figure 1, the simulation procedure repeats the steps in Boxes 3-6 for 500 iterations or trials. At the end of the simulation we have 500 CDFDEVs for each of the 110 PDFs tested (Box 7). The 22 PDFs of CDFDEVs for the $\tilde{Y}_N$ sample are compared to one another to see which one is lowest for each iteration and across all 500 iterations (Box 8). The best way to compare how closely a distribution simulated its parent distribution is to count how many times out of 500 it had the lowest CDFDEV, i.e., calculate the probability that a particular PDF will have the lowest CDFDEV. The counting process was repeated a second time to estimate the probability of a distribution being the second most preferred, and a third time to determine which PDF is the third most preferred, etc.

The purpose of using five parent distributions, (Boxes 2 and 3) is to minimize the bias of generating the stochastic sample values with only a single distribution. By testing five parent distributions we can be more confident that the PDF with the highest probability of the lowest CDFDEV is adequate enough to be a first choice when we have to assume a distribution for a sparse data random input variable of the specified type. Additionally, this allowed for testing whether the type of distribution that generated the random sample biased the selection of the “best” PDF.

To further test the procedure we simulated eight different independent variables $\tilde{Y}_1$, ..., $\tilde{Y}_8$ (Table 1). Additionally, the procedure was used to test two multivariate distributions under the assumption that they were linear

Table 1: Historical data for eight random input variables used for testing alternative PDFs for simulating random variables that have only positive values.

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8
1	212.09	307.00	35.72	32.00	45.87	40.70	9.73	32.90
2	409.35	395.00	30.02	31.60	84.53	60.70	64.47	85.20
3	269.27	358.00	29.58	26.80	87.89	59.60	66.25	86.30
4	359.48	339.00	32.39	29.20	73.56	45.30	35.00	77.20
5	712.73	364.00	42.92	26.10	78.42	45.50	57.58	64.80
6	531.43	360.00	35.31	23.10	86.36	49.00	67.59	73.30
7	729.00	558.00	35.43	26.90	87.85	59.00	68.57	92.50
8	618.18	554.00	21.88	27.90	65.43	54.60	54.26	70.60
9	436.36	393.00	17.78	19.00	41.96	40.00	13.76	72.00
10	475.18	403.11	23.90	33.00	57.61	60.00	70.74	99.00
Mean	475.31	403.11	30.49	27.56	70.95	51.44	50.79	75.38
Minimum	212.09	307.00	17.78	19.00	41.96	40.00	9.73	32.90
Maximum	729.00	558.00	42.92	33.00	87.89	60.70	70.74	99.00

dependencies between $\tilde{Y}_1, \ldots, \tilde{Y}_4$ as well as $\tilde{Y}_5, \ldots, \tilde{Y}_8$. The linear dependencies were modeled for the parent distributions using the procedure reported by Richardson et al. [7]. The results for the preferred PDFs were identical regardless of assuming a multivariate distribution or the assumption of independent random variables.

3 Results

The average CDFDEV values for simulating variables $\tilde{Y}_1, \ldots, \tilde{Y}_4$ are summarized in Table 2 to show how the goodness-of-fit values work. For variable $\tilde{Y}_1$, the PDF that generated the lowest average CDFDEV over the 500 iterations is the Parzen KD with a value of 219.73. The next best fit is the triweight KD with an average CDFDEV of 271.38. The ranking of PDFs with the Parzen KD ranked first followed by the triweight KD is consistent across all four variables. Excluding the Cauchy KD, the kernal distributions outperformed all of the parametric distributions. The beta distribution was the preferred parametric distribution for three cases variables ($\tilde{Y}_2, \ldots, \tilde{Y}_4$). It is interesting that the normal density and Gaussian kernal functions did not perform well, even though the parent distributions for $\tilde{Y}_1, \ldots, \tilde{Y}_4$ were generated using a normal distribution. This result suggests that the Parzen or triweight kernel distributions are more suitable with small samples when the true underlying distribution for the random input variable is unknown.

Average CDFDEVs are useful to suggest a ranking of PDFs but they do not provide any statistical inference to the selection of the "best" PDF. To provide a

Table 2: Average CDFDEV values simulated for variables Y1-Y4 to demonstrate the range of values observed from simulating the variables with parameters for 22 alternative assumed distributions.

Distribution	YN1	YN2	YN3	YN4
Beta	>	579.43	4.46	1.20
Double Exponential	>	>	788.41	256.18
Exponential	>	>	>	1,183.06
Gamma	>	>	266.85	62.36
Logistic	>	>	396.21	129.83
Log-Log	>	>	1,016.92	335.25
Log-Logistic	>	>	>	499.85
Lognormal	>	>	>	93.95
Normal	>	>	113.89	37.40
Pareto	>	>	>	>
Uniform	>	789.52	6.07	1.95
Weibull	>	>	74.65	25.45
Cauchy KD	>	1,173.50	8.96	2.82
Cosinus KD	422.15	103.56	0.78	0.25
Double Exp KD	1,932.73	482.92	3.68	1.17
Epanechnikov KD	440.43	108.07	0.81	0.26
Gaussian KD	1,225.90	305.36	2.34	0.75
Parzen KD	**219.73**	**53.49**	**0.40**	**0.13**
Quartic KD	342.16	83.73	0.63	0.20
Triangle KD	377.94	92.81	0.70	0.23
Triweight KD	281.39	68.73	0.51	0.17
Uniform KD	637.90	157.02	1.20	0.39
A value greater than 2,000 is indicated by a ">" sign.				

probabilistic rigor for picking one PDF over another we turn to the probability statistics. In this case we calculated the probability out of 500 draws that a PDF would have the lowest CDFDEV, or the second lowest or the third lowest, etc.

Based on the small set of eight random input variables in Table 1, the "best" PDF to use for simulating a random input variable, if we do not have a *priori* information, is the Parzen kernal (Table 3). The "best" PDF for simulating the eight random variables is the Parzen kernal (Table 3). The probability of the Parzen kernal being "best" is greater than 90% for the 40 combinations of variables ($\tilde{Y}_1$, ..., $\tilde{Y}_8$) and parent distributions (normal, beta, gamma, uniform, and Weibull) reported in Table 3 and greater than 95% for 31 of the 40 combinations. The assumption regarding the distribution used to simulate the observations for the parent distributions did not affect the outcome for any of the eight random variables, as there is no pattern showing that the Parzen KD was less accurate in simulating a normal, beta, gamma, uniform, or Weibull

Table 3: Probability distribution rankings for simulating eight random variables having few observations.

PDF Ranking	Y1						Y2				
	Normal	Beta	Gamma	Uniform	Weibull		Normal	Beta	Gamma	Uniform	Weibull
P(Parzen First)	97.4%	97.2%	98.6%	98.2%	97.6%		97.4%	92.8%	96.8%	93.2%	96.8%
P(Triweight First if Parzen Ignored)	96.2%	94.4%	97.2%	96.2%	96.2%		94.2%	88.4%	93.4%	90.0%	94.8%
P(Quartic First if Ignore above Dist)	93.6%	86.4%	92.0%	87.4%	92.8%		86.8%	76.6%	87.8%	80.0%	87.2%
P(Triangle of Ignore above Dist)	93.0%	85.6%	91.4%	87.6%	92.2%		86.4%	76.8%	87.8%	79.0%	86.4%
P(Cosinus if Ignore above Dist)	94.2%	91.2%	94.8%	91.6%	94.0%		89.2%	80.6%	90.0%	80.6%	88.2%
	Y3						Y4				
	Normal	Beta	Gamma	Uniform	Weibull		Normal	Beta	Gamma	Uniform	Weibull
P(Parzen First)	97.0%	94.2%	95.4%	94.8%	97.2%		95.8%	94.6%	96.0%	93.4%	95.4%
P(Triweight First if Parzen Ignored)	94.2%	89.8%	94.2%	90.4%	94.2%		93.4%	90.6%	92.6%	89.0%	92.6%
P(Quartic First if Ignore above Dist)	88.6%	79.8%	86.8%	79.8%	88.6%		84.4%	79.6%	85.8%	78.6%	85.4%
P(Triangle of Ignore above Dist)	87.8%	78.0%	86.8%	78.4%	87.6%		83.6%	79.6%	85.6%	77.2%	86.0%
P(Cosinus if Ignore above Dist)	89.4%	79.8%	89.0%	80.2%	90.4%		86.4%	81.6%	86.8%	79.8%	87.4%

Table 3: Continued.

	Y5					Y6				
	Normal	Beta	Gamma	Uniform	Weibull	Normal	Beta	Gamma	Uniform	Weibull
P(Parzen First)	97.6%	96.8%	97.8%	96.4%	97.6%	96.2%	93.2%	96.8%	90.6%	96.2%
P(Triweight First if Parzen Ignored)	95.8%	95.2%	95.2%	94.2%	96.4%	94.0%	88.6%	93.8%	86.4%	94.8%
P(Quartic First if Ignore above Dist)	90.8%	84.2%	90.0%	85.8%	90.8%	89.4%	77.6%	89.0%	77.8%	91.6%
P(Triangle of Ignore above Dist)	90.4%	84.6%	89.8%	84.6%	90.2%	89.6%	76.0%	88.4%	77.6%	90.4%
P(Cosinus if Ignore above Dist)	94.8%	92.8%	94.6%	90.8%	94.8%	90.4%	81.0%	89.2%	76.4%	89.4%
	Y7					Y8				
	Normal	Beta	Gamma	Uniform	Weibull	Normal	Beta	Gamma	Uniform	Weibull
P(Parzen First)	97.8%	96.6%	97.2%	97.0%	97.4%	97.2%	97.4%	97.4%	96.4%	94.8%
P(Triweight First if Parzen Ignored)	95.8%	94.6%	95.2%	93.2%	95.6%	93.2%	94.8%	94.8%	93.8%	93.0%
P(Quartic First if Ignore above Dist)	90.6%	84.0%	90.8%	84.0%	89.4%	86.8%	84.4%	87.8%	82.8%	87.0%
P(Triangle of Ignore above Dist)	90.4%	83.6%	89.6%	82.8%	87.8%	86.6%	83.4%	87.4%	81.0%	86.4%
P(Cosinus if Ignore above Dist)	91.4%	89.2%	91.6%	88.2%	91.8%	89.8%	86.2%	90.4%	84.8%	87.8%

distribution. Even if the random values used to produce the parent distribution sample were generated by a normal distribution, the CDFDEV statistic was never lowest for the normal distribution.

The second best distribution was the Triweight kernal across all variables and distributions (Table 3). Other PDFs outperform the Triweight kernal 3% to 16% for the 40 distribution variables tested suggesting the Triweight is a not a good second choice. The third, fourth, and fifth best distributions were quartic, triangle, and cosines kernals, respectively.

None of the 12 frequently used parametric distributions outperformed the Parzen, triweight, quartic, triangle, and cosines kernals with a significant probability. This outcome should be expected because the 12 parametric distributions tested force the data to conform to standard forms while the kernal distributions are more flexible to accommodate irregularities associated with small samples.

4 Summary

The purpose of this paper was to compare the goodness-of-fit for 12 different parametric distributions and 10 kernal distributions to determine which distribution would perform best for simulating random variables with positive values. A Monte Carlo simulation procedure was developed to estimate how well 22 PDFs performed at reproducing random input distributions. The procedure sampled eight variables five different ways, calculated parameters for 22 PDFs and simulated the process 500 times to assign probabilities to each PDF being the "best" for simulating random input variables.

The procedure ranked the Parzen kernal "best" more than 90% of the time. The triweight kernal was ranked second. The top two kernal distributions significantly outperformed 12 common parametric distributions and nine other kernel distributions.

The conclusion based on these results is that if we do not have any sound information on which to pick a probability distribution for simulating a random input variable with sparse positive values, we should use a Parzen kernal distribution.

References

[1] Law, A.M. & Kelton, W.D., *Simulation Modeling and Analysis*, McGraw-Hill Book Co.: New York, 1982.

[2] Palisade Corp. User's Guide: Best Fit. Newfield, NY: Palisade Corporation, 1995.

[3] Parzen, E., An estimation of a probability density function and mode. *The Ann. of Mathematical Statistics*. **33**, pp. 1065-1076, 1962.

[4] Silverman, B.W., *Density Estimation for Statistics and Data Analysis*, Chapman & Hall/CRC: Boca Raton, FL, 1998.

[5] Chen, S.X., Probability density function estimation using gamma kernals. *The Ann. of Inst. Statistics and Mathematics*, **52**, pp. 471-480, 2000.

[6] Richardson, J.W., Schumann, K., & Feldman, P., *Simetar: Simulation for Excel to Analyze Risk.* Department of Agricultural Economics, Texas A&M University, College Station, Texas, January 2008.

[7] Richardson, J.W., Klose, S.L., & Gray, A.W., An applied procedure for estimating and simulating multivariate empirical (MVE) probability distributions in farm-level risk assessment and policy analysis. *Journal of Agricultural and Applied Economics*, **32(2)**, pp. 299-315, August 2000.