Proceedings of the 1986 Winter Simulation Conference
J. Wilson, J. Henriksen, S. Roberts (eds.)

DISTRIBUTION SELECTION IN STATISTICAL SIMULATION STUDIES

Mark E. Johnson
Statistics and Applied Mathematics
Los Alamos National Laboratory, MS F600
Los Alamos, NM 87545, U.S.A.

ABSTRACT

The statistics profession has been remiss in exploiting the numerous advances in simulation methodology. The purpose of this article is to outline progress in variate generation relevant to the conduct of statistical simulation studies. The emphasis is on multivariate distributions, a thriving area of research.

1. INTRODUCTION

Once upon a time (circa 1960's), statisticians would report simulation studies that used whatever probability distributions they happened to know how to generate. Since the list of generators was not particularly extensive, studies generally consisted of isolated results pertaining to particular distributions. To their credit, researchers of this era tended to admit the limitations of their studies and to acknowledge that extrapolations to other distributions or classes of distributions would require additional work or tenuous assumptions. As time passed and simulation methodology evolved, capabilities in conducting simulation studies advanced to the point where researchers had to make decisions as to what to simulate. The decision making process of selecting distributions in statistical simulation studies is the topic of this paper. As motivation for the relevance of this problem, let me cite some particular reasons given by various authors for the distributions included in their studies:

1. Legal precedence. The distributions selected were the same as those included by earlier authors. As a case in point, current Hotelling's T^2 simulations frequently resemble those run by Chase and Bulgren (1971) and Hopkins and Clay (1963). When these two studies appeared they were just fine and represented a true contribution to knowledge of the robustness of T^2 via Monte Carlo methods. Since that time, a great deal of progress has been made both in multivariate analysis and simulation methodology. Are we still interested in the the performance of T^2 (or new nonparametric alternatives) for the Marshall-Olkin bivariate exponential distribution?

2. Availability implies suitability. The distributions selected were those available in the statistical package which also does the other calculations. It is a pity that the packages cannot write the papers as well. I like having the standard packages with a simulation capability--it allows me to check these codes for situations in which I know what to expect. As versions change (both the analysis and simulation parts), it becomes increasingly difficult to explain results which cannot be reproduced.

3. Status quo. The distributions selected were not defended. It is not uncommon to see simulation studies reported with no comments at all regarding the appropriateness of the distributions selected. In one recent situation, I managed to extract the following "defense" that can be paraphrased as "I see no reason why I should have to demonstrate the usefulness or relevance of the distributions included--I never have before." This view may seem a bit extreme but ties into the old fashioned happy-go-lucky mode of operation--include distributions which can be generated rather than what ought to be generated.

4. Novelty. The distributions selected had never before been published and represented a shot out of the blue. The novelty tactic is useful in providing motivation for future papers ("In a previous paper, the distributions developed here were found useful in a simulation context."). Generally speaking, information on new distributions included is useful in interpreting the results of simulation studies.

2. GENERAL GUIDELINES

Having ridiculed the above arguments for distribution selection, it seems appropriate to provide some positive guidelines for developing rational arguments. The following discussions outline schemes corresponding to viable selection criteria.

1. Data specific models. Perhaps the most compelling argument for defending the distributions selected is that those distributions selected are similar to those from which the future data sets are to arise. In other words, select distributions in the simulations which supposedly mimic the processes producing the data to be analyzed. The disadvantage of this concept is that extrapolating results to other situations which occur later may not be possible. As an example of a simulation study driven by these practical situations, see Conover et al. (1981). The notion of testing a procedure or evaluating estimators on models similar to those observed in practice is not new, the bootstrap method (Efron, 1977) exploits this notion. Sampling from parametric models conforming to data sets allows the possibility of generating new data outside the range of the existing data set.

2. Quantitative selections to establish qualitative features. In some situations it may be possible to identify certain characteristics of interest that existing parametric models can capture. In a robustness study of the correlation coefficient, Devlin et al. (1981) argued convincingly that the elliptically contoured class of distributions provide a useful framework for assessing the performance of various estimators. In a similar vein, Nachtsheim and Johnson (1986) constructed anisotropic distributions for use in investigating the small sample performance of Hotelling's T^2 statistic.

3. Philosophy of quantitative departures from a baseline model. An approach which is intuitively appealing but at present practically untested is to couch the performance under varying distributions problem as an optimization problem. That is, set up the search space as a class of probability distributions constrained in some fashion (e.g., univariate, continuous, unimodal, finite second moment, etc.) and try to find the extreme performance of a procedure over this class. The advantage of a solution to this problem is that the results apply to the class rather than just the distributions actually selected.

4. Caveat emptor (and its closely related cousin, the fishing expedition). A couple of situations come to mind for which a less than ideal simulation study can be conducted and justified. Sometimes a very small simulation study relates to computing budgets, in which case it is difficult to condemn an author for something beyond his control. In another situation, a simulation study may be given as merely an indication of how a procedure works in a few selected cases. The main contribution of an article may be the genesis of a procedure rather than its performance under diverse circumstances. Whenever limited studies are conducted, it is advisable to avoid sweeping recommendations.

3. UNIVARIATE CONTINUOUS DISTRIBUTIONS

There are many univariate distributions that are easy to generate and that can fit into the guidelines provided in Section 2. Devroye's (1986) comprehensive text on non-uniform random variate generation describes algorithms for the following distributions: normal, exponential, gamma, beta, t, stable, Bessel function, logistic, hyperbolic, von Mises, Burr, generalized inverse Gaussian, and many scattered in the exercises. The slash distribution (a normal variate divided by an independent uniform) is popular in statistical simulations, since it allows highly precise calculations of estimator performance (I look forward to someday seeing a data set that the slash

distribution satisfactorily models.) The point of this list is that virtually any continuous, univariate distribution can be generated reasonably efficiently. The difficulty is in culling the list to manageable yet meaningful proportions. The guidelines given in Section 2 provide general rules for selecting distributions to include in a study.

4. CONTINUOUS MULTIVARIATE DISTRIBUTIONS

The natural baseline case for multivariate simulation studies is the multivariate normal distribution. With mean vector $\underline{0}$ and identity covariance matrix, the multivariate normal distribution's density function is proportional to $\exp[-(\underline{x}'\underline{x})/2]$, where $\underline{x}$ is a $p \times 1$ vector. In this section, we explore reasonable departures from the baseline multivariate normal. By "reasonable", we mean that the alternative distributions are

1. Easy to generate.
2. Have an identifiable property inherently and quantifiably distinct from the multivariate normal distribution.

We turn now to a brief description of some of the possible alternative distributions. A much more detailed treatment can be found in Johnson (1987).

The easiest departure to consider is to vary the marginal distributions from normal, preserving independent components. This strategy will reveal effects due to the marginal distributions alone and provide a reference point for subsequent departures from dependence.

For considering symmetric alternatives to the normal, two approaches are appealing. First, consider scale-contamination to the normal variates to get t-variates. This is accomplished easily. If X_i is a standard normal variate from $\underline{X} \sim N_p(\underline{0}, I)$, then take

$$Y_i = X_i/\sqrt{(Z_i/n)}, \quad i = 1, \ldots, p$$

where Z_i is an independent χ^2-variate with n degrees of freedom. Notice that we can use the same set of generated $\underline{X}$ variates with different Z_i's to effect a variance reduction (via correlated random samples). Using a different Z for each X_i leads to independent Y_i's. If the same Z used for each X_i, then $\underline{Y}$ would have a spherically symmetric t-distribution with uncorrelated (if the moments exist) but dependent components. The parameter n need not be an integer. A candidate set of values for n is {30, 10, 5, 2, 1, .5}. The results obtained using n=30 ought to agree pretty well with the baseline normal case unless the procedure is very sensitive to minor departures from normality. The value of n=1 corresponds to the heavy-tailed Cauchy distribution. The Cauchy distribution and in fact even heavier-tailed distribution have been observed in some experiments (Beckman and Johnson, 1987), so these distributions are not as far out as some might think (on the basis of non-existent moments).

The t-distributions noted above have tails heavier than the baseline normal. The generalized exponential power distribution (Johnson, Tietjen, Beckman, 1986) is another class of symmetric univariate distributions that is easy to generate, is amenable to variance reduction designs, and brackets the baseline normal in a reasonable fashion. The genesis of the distribution is embodied in the following generation algorithm:

1. Generate $W \sim \Gamma(\alpha, 1)$, $\alpha > 0$.
2. Transform using $X = \pm W^{\tau}$ where $\tau > 0$ and "$\pm$" denotes a random sign.
3. Translate as $\sigma[3\Gamma(\alpha)/\Gamma(\alpha+2\tau)]^{1/2} X + \mu$.

The resulting variate from step 3 has mean μ, variance σ^2 and is symmetric, unimodal about μ. Moreover, for $\alpha = 3/2$ and $\tau = 1/2$, the distribution is normal (μ, σ^2); for α-t=1, the ordinary exponential power distribution occurs. The general exponential power distribution includes the normal as an intermediate special case, rather than a limiting case as in the t-distribution ($n \to \infty$). Heavier than normal tails occur for $\alpha < 3/2$. A variance reduction design is possible by reuse of the gamma variates in step 1 for various τ values. For additional details, see Johnson (1987, Section 2.4).

For asymmetric alternatives from the normal, a nice progression away from normality can be obtained using the lognormal distribution. If X is normal $(0, \sigma^2)$ then $Y = \lambda\exp(X) + \xi$ is lognormal with shape parameter σ. The lognormal tends to the normal as $\sigma \to 0$. A candidate set of shape parameter values is {.01, .5, 1., 2.}. The parameters λ and ξ can be selected to achieve a specified mean and variance.

Implicitly it has been assumed in the above distributions that each marginal distribution is the same. There is nothing wrong with exploring effects

due to non-normal marginals in a subset of the components. It is conceivable that some statistical procedures can tolerate one or two highly non-normal marginal components if the remaining components are normal.

More intellectually satisfying departures from baseline multivariate normality involve the introduction of dependence among the component distribution. This type of departure is the primary theme of Multivariate Statistical Simulation and has been a major area of the author's research for the past ten years. The presentation here will focus on various representations of the multivariate normal and then examine adjustments to these constructions.

Consider first the general representation (Cambanis, Huang, and Simons, 1977)

$$\underline{X} = R\,\underline{U}^{(n)} \qquad (3.1)$$

where R is a scalar random variable independent of $\underline{U}^{(n)}$ that is uniform on the n-dimensional unit hypersphere. For R distributed as $\sqrt{\chi^2_{(n)}}$, $\underline{X}$ is multivariate normal with mean vector $\underline{0}$ and identity covariance matrix. Distributions having representation (3.1) are spherically symmetric and differ from the multivariate normal through appropriate choice of R. Two reasonable choices for the distribution of R^2 are the Pearson Type VI and the beta distribution. In the former case, a Pearson Type VI has density function:

$$g(z) \propto z^{n/z - 1}\,(1 + z)^{-m},\quad z > 0.$$

The corresponding distribution of $\underline{X}$ is multivariate Pearson Type VII having density function:

$$h(\underline{x}) \propto \frac{1}{(1 + \underline{x}'\underline{x})^m}$$

For the other case consider R^2 having a beta distribution with parameters p/2 and m + 1. Using this distribution in the basic representation leads to the multivariate Pearson Type II distribution having density function

$$f(\underline{x}) \propto (1 - \underline{x}'\underline{x})^m$$

on the finite support $\underline{x}'\underline{x} \leq 1$. In the bivariate setting, a reasonable set of cases for these Pearson Types is, as follows:

Type II: m = -.5, 0, .5, 1., 2., 5.
Type VII: m = 1.1, 1.5, 2., 3., 10.

In both sets, large values of m correspond to near bivariate normality.

As noted earlier, the Pearson Type VII can be generated using the alternate representation

$$Y_i = X_i/\sqrt{(\chi^2_{(n)}/n)}$$

where a single χ^2 variate is used for each X_i.

Another possible approach to generalizing the basic construction (3.1) is to consider non-uniform distributions for the vector $U^{(n)}$. This approach has been explored by Nachtsheim and Johnson (1986). A nice feature here is that there are many existing, well-studied distributions on the circle or sphere to employ in (3.1). In particular, they consider cardioid, triangular, Von Mises, offset normal, wrapped Cauchy, and wrapped normal in two dimensions and Bingham and Fisher in three dimensions. The basic, general density form in two dimensions is:

$$f(x,y) = h[\tan^{-1}(y/x)]\,\exp[-(x^2+y^2)/2],$$

where $h(\theta)$ has support $(0, 2\pi)$. Of course, in relaxing uniformity of $U^{(n)}$, the resulting marginal distributions vary as well. However, this is the price to be paid to accomplish directional asymmetry.

The above schemes allow scrutinizing statistical procedures under reasonable departures from normality. In a sense, among the many multivariate distributions available these set-ups allow the most controlled experiments in exploring non-normality. There are other distributions which particular circumstances might dictate consideration. The following distributions are further considered by Johnson (1987):

- Johnson translation system
  - lognormal
  - $\sinh^{-1}$-normal
  - logit normal
- Khintchine distributions
- Burr-Pareto-logistic family
- Miscellaneous
  - Morgenstern (a/k/a Gumbel-Farlie-Eyraud)
  - Plackett
  - Wishart
  - Ali-Mikhail-Haq

5. PLOTS.

A useful device for studying bivariate distributions (especially bivariate) is contour and three-dimensional plots. Some sample plots similar to those found in Johnson (1987) are given to stimulate the interest of the reader.

6. SUMMARY.

The topic of distribution selection in statistical simulation is one of increasing importance, as the selection process is critical to the eventual success of a study. Treating distributions as factors or controlled variables in a design context is an essential first step to conducting meaningful simulations.

REFERENCES.

Beckman, R. J. and M. E. Johnson (1987). "Fitting Student's t Distribution to Grouped Data, With Application to a Particle Scattering Experiment," Technometrics, to appear.

Cambanis, S., S. Huang, and G. Simons (1981). "On the Theory of Elliptically Contoured Distributions," Journal of Multivariate Analysis, 11, 368-385.

Chase, G. R. and W. G. Bulgren (1971). "A Monte Carlo Investigation of the Robustness of Hofelling's T^2," Journal of the American Statistical Association, 66, 499-502.

Conover, W. J., M. E. Johnson and M. M. Johnson (1981). "A Comparative Study of Tests for Homogeneity of Variances, with Applications to the Outer Continential Shelf Bidding Data," Technometrics, 23, 351-361.

Devlin, S. J., R. Gnanadesikan, and J. R. Kettenring (1981). "Robust Estimation of Dispersion Matrices and Principle Components," Journal of the American Statistical Association, 76, 354-362.

Devroye, L. (1986). Non-Uniform Random Variate Generation. Springer-Verlag, New York.

Efron, B. (1977). "Bootstrap Methods: Another Look at the Jacknife," Annals of Statistics, 7, 1-26.

Hopkins, J. W. and P. P. F. Clay (1963). "Some Empirical Distributions of Bivariate T^2 and Homoscedasticity Criterior M Under Unegual Variance and Leptokurtosis," Journal of the American Statictical Association, 58, 1048-1053.

Johnson, M. E. (1987). Multivariate Statistical Simulation, John Wiley, New York.

Johnson, M. E., G. L. Tietjen, and R. J. Beckman (1980). "A New Family of Probability Distributions and Applications to Monte Carlo Studies," Journal of the American Statistical Association, 75, 276-279.

Nachtsheim, C. J. and M. E. Johnson (1986). "Some New Anisotropic Distributions Useful in Simulation," in preparation.

AUTHOR'S BIOGRAPHY

MARK E. JOHNSON is a staff memeber in the Statistics and Applied Mathematics Group at Los Alamos National Laboratory. He received a B.A. ('73) in mathematics and M.S. ('74) and Ph.D. ('76) degrees in Industrial and Management Engineering from the University of Iowa. He has been a visiting professor at the University of Arizona (Systems and Industrial Engineering) and the University of Minnesota (Applied Statistics). He is an associate editor of Technometrics, JSCS and AJMMS. He has been awarded the Jack Youden prize ('81) from Technometrics and the Thomas L. Saaty prize ('84) from AJMMS. He is a member of ASA, IMS, WNAR, and RSS.

Mark E. Johnson
Statistics and Applied
Mathematics, MS F600
Los Alamos National
Laboratory
Los Alamos, NM 87545

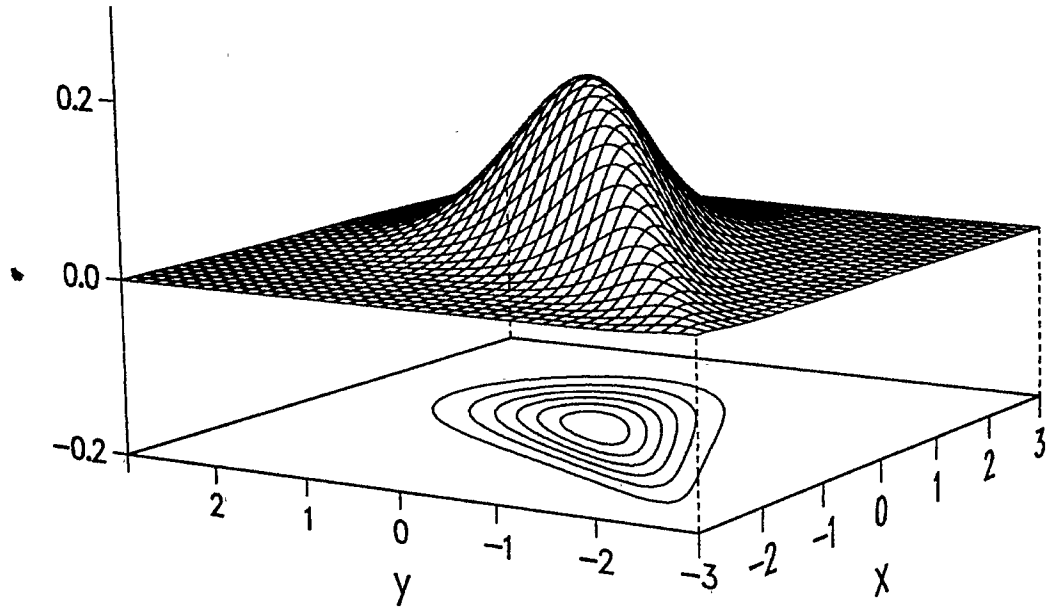

Figure 1. Burr-Pareto-logistic Density Function, Normal Marginals (α=0.75).

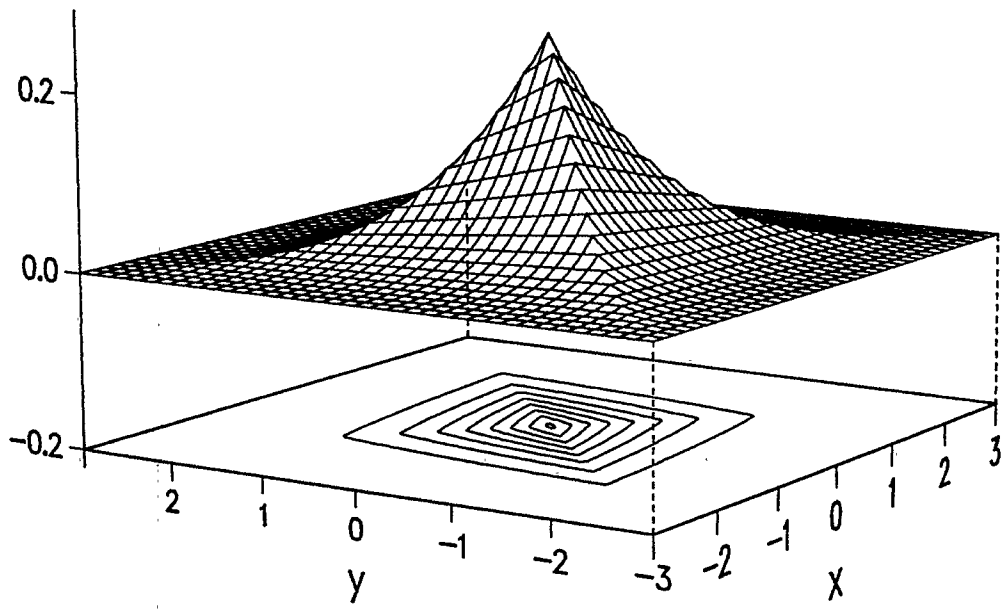

Figure 2. Khintchine Normal Density Function (α=0.5).

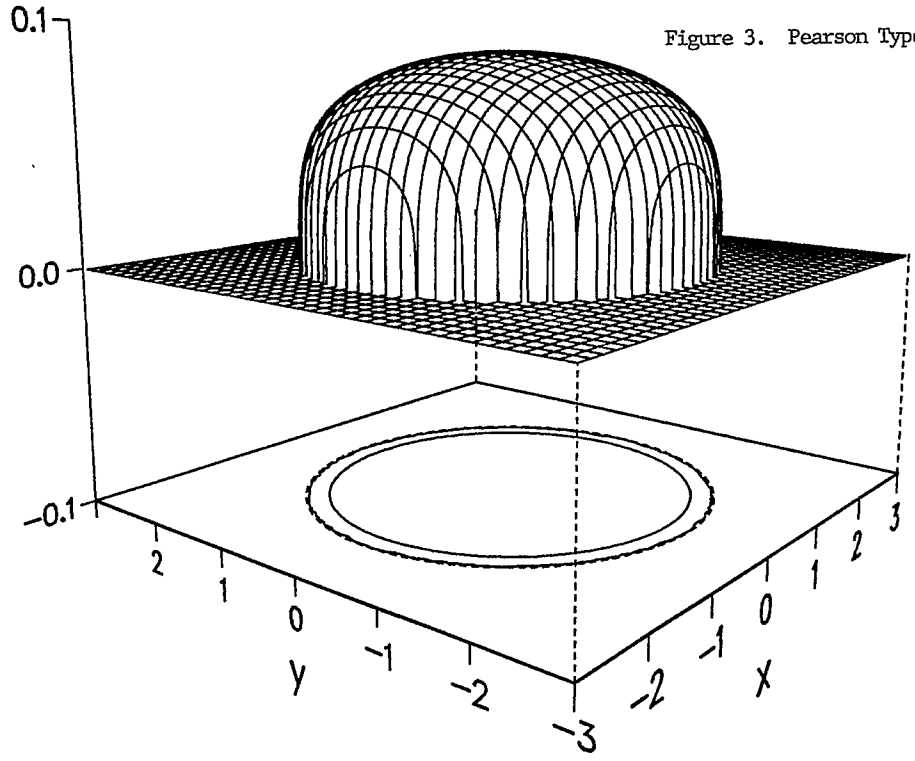

Figure 3. Pearson Type II Density Function (m=0.25).

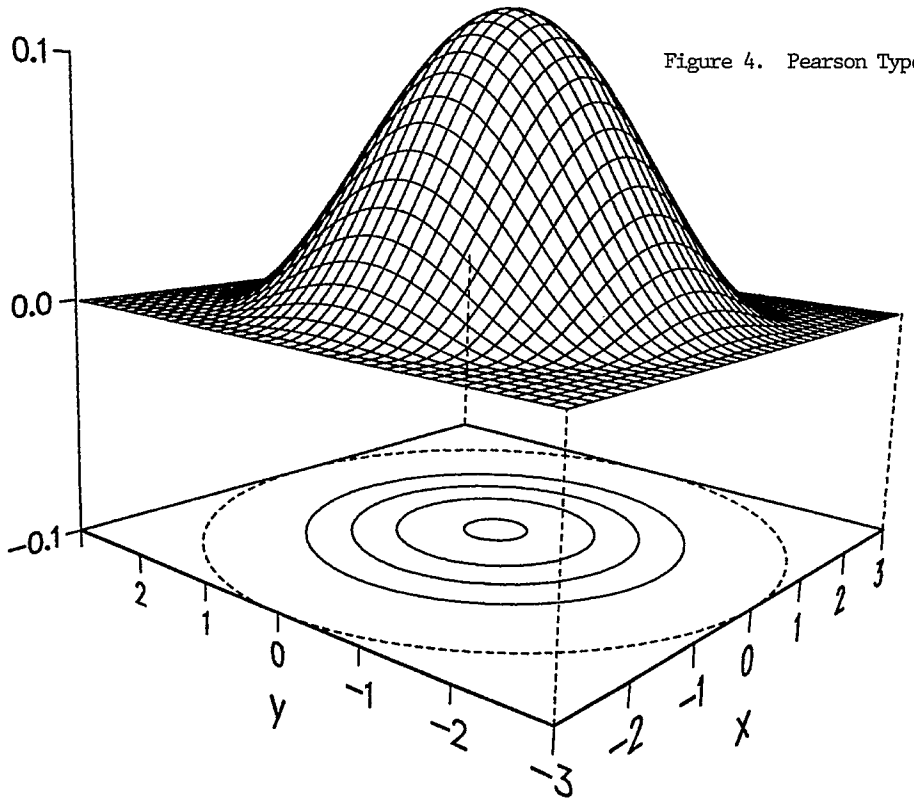

Figure 4. Pearson Type II Density Function (m=2.5).