1981 Winter Simulation Conference Proceedings
T.I. Ören, C.M. Delfosse, C.M. Shub (Eds.)

FITTING "STANDARD" DISTRIBUTIONS TO DATA IS NECESSARILY GOOD: DOGMA OR MYTH?

Bennett L. Fox
Département d'informatique et
de recherche opérationnelle
Université de Montréal
C.P. 6128, Succursale "A"
Montréal, P.Q. Canada H3C 3J7

This paper questions the conventional wisdom that "standard" distributions such as the normal, gamma, or beta families should necessarily be used to model inputs for simulation and proposes an alternative. Relevant criteria are the quality of the fits and the impact of the input distributions on effective use of variance reduction techniques and on variate-generation speed. The proposed alternative, a "quasi-empirical" distribution, looks good on all these measures.

A "standard" distribution is for our purposes any theoretical distribution commonly found in statistics textbooks. Examples are gamma, beta, normal, and Weibull distributions. These distributions share a key feature: they have only a few parameters. Why fit such distributions to data or use them to capture subjective assessments? One reason is smoothness. Another is compactness: remembering only a few parameters is easier than remembering all the data, and summarizing subjective assessments by a few parameters makes them appear less vague. Sometimes there is a better reason: postulating certain distributions, such as the exponential, may be natural if they simplify probabilistic or statistical analysis and if an underlying limit theorem, such as Poisson superposition, seems to apply. We will argue that this last reason is the only good reason. Even with that reason, validation of the postulated distribution together with the estimates of its parameters is needed.

A fit may be spurious. Unless there is a huge amount of data, goodness-of-fit tests have notoriously low power--even less when applied to a gamut of distributions to select the best fit out of the set. Especially when several parameters are estimated from data, cross-validation gives more power. Fit using half the data, test the fit with the other half, then interchange roles.

Particularly for shape parameters, estimation often is difficult or non-robust, involving nonlinear equations; outliers may well result in grossly-wrong estimates. In addition, it is quite possible that the distribution from which the data come

81CH1709-5/81/0000-0305$00.75

differs significantly from the fitted distribution no matter what parameter values are chosen. In such cases, fitting loses and distorts information. This risk makes us skeptical whether the first two reasons for fitting, namely smoothness and compactness, are sufficient.

The above remarks apply in any setting. In the simulation context, there is another, perhaps more important, consideration.

Except for the exponential distribution, none of the "standard" distributions can be sampled quickly and easily by inversion (see below). Only inversion is compatible with variance reduction techniques based on inducing correlation (e.g., common random numbers, antithetic variates, and control variates). Reason: only inversion transforms uniform random numbers to nonuniform random numbers in a monotone, one-to-one way. This allows synchronization of simulation experiments as discussed in detail by Bratley, Fox, and Schrage (in preparation). If rejection is used to generate nonuniform random numbers, in general we eventually get a rejection in one experiment corresponding to an acceptance in the other; this misalignment persists, destroying synchronization.

How can one generate random numbers from "standard" distributions by inversion? The Weibull distribution has a closed-form cumulative, but generating Weibull variates is not necessarily fast: proceeding directly, we take a log and a k-th root every time. The gamma, beta, and normal distributions are more typical in that they do not have closed-from cumulatives. If standard methods, such as continued fractions, are used to approximate the inverses, variate generation is slow. Recently Ahrens and Kohrt (1981) presented what appears to be a method for generating variates by inversion from "largely arbitrary" distributions that is "almost as fast" as tailored rejection algorithms. However, their method is complicated and they do not indicate its accuracy.

If for a particular distribution variate generation by accurate inversion significantly lengthens simulation runs, we have two alternatives. The first is to go a relatively fast rejection method, thus giving up the chance to reduce the number of runs to achieve a given statistical precision by such variance reduction techniques as common random numbers, antithetic variates, and control variates. The second alternative is more attractive to us: choose a different distribution

(i) that models the (generally unknown) real distribution about as well as any other, and

(ii) for which accurate inversion is fast and easy.

This distribution may be "nonstandard", but that does not bother us at all. We have already argued that the putative benefits of fitting with "standard" distributions may well be illusory.

Our candidate "nonstandard" distribution is a "quasi-empirical" distribution that blends a continuous piecewise-linear component with a shifted exponential tail. The first part closely mimics the data, interpolating the usual empirical distribution up to a certain breakpoint. The exponential tail has some theoretical justification, as detailed in Bratley, Fox, and Schrage (in preparation); for sensitivity analysis, replace it by Weibull tails of various shapes. With a reasonable number of breakpoints, say 20 or more, the quasi-

empirical distribution will look quite smooth. The piecewise linear interpolation tends to shift the mean of the empirical distribution slightly to the left, but this is exactly offset by selecting the parameter of the exponential tail suitably; see Bratley, Fox, and Schrage (in preparation) for details or carry out the routine calculus and algebra yourself. We can also use a quasi-empirical distribution to approximate a "standard" distribution, whether the latter arises by fitting data or simply reflects a subjective assessment. It seems to us more natural to postulate a quasi-empirical distribution directly. In view of all this, it seems fair to say that the quasi-empirical distribution satisfies (i) above. We now show that it also satisfies (ii).

The key observation is that the vertical spacings between successive breakpoints are all the same, say 1/n. Let u be a uniform random number. Let j be the integer part of nu and let w be the fractional part (w=nu-j). Thus, j tells us what piece of the quasi-empirical distribution to work with. If the piece is linear, use w to do (inverse) linear interpolation. If the piece is the exponential tail, inversion is again simple; it is easy to figure out how to do this. For the record, Bratley, Fox, and Schrage (in preparation) give the answer. Because the tail will normally be sampled rarely, variate generation is fast and only a few lines of FORTRAN or other high-level language are needed.

To close, we answer the question posed in the title: that fitting "standard" distributions to data is necessarily good is both a dogma and a myth. Such fitting some accept uncritically as a principle, but the belief that it is universally applicable is false.

Paul Bratley's comments on an earlier version of this paper improved its style. This paper was written while the author was a visiting professor at Yale University.

REFERENCES

Ahrens, J.H. and K.D. Kohrt (1981), Computer Methods for Efficient Sampling from Largely Arbitrary Statistical Distributions, Computing, Vol. 26, pp. 19-31.

Bratley, P., B.L. Fox, and L.E. Schrage (in preparation), A Guide to Simulation.