Proceedings of the 1983
Winter Simulation Conference
S. Roberts, J. Banks, B. Schmeiser (eds.)

RANDOM SAMPLES WITH KNOWN SAMPLE STATISTICS : WITH APPLICATION TO VARIANCE REDUCTION

Russell C.H. Cheng
Department of Mathematics
UWIST
Colum Drive,
Cardiff, CF1 3EU
U.K.

A number of sampling schemes are described which aim to improve the accuracy of estimators in simulation experiments. The schemes negatively correlate key sample statistics in pairs of runs or blocks of runs. Use is made of generators which produce random samples from certain distributions when sample statistics like the mean and dispersion are prescribed. Special cases considered include normal, inverse Gaussian and gamma generators. These can be used in their own right or as the basis of generators of other distributions. Guide lines are given which indicate the conditions under which the schemes might be effective.

1. INTRODUCTION

The use of antithetic variates for improving the accuracy of estimators in computer simulation experiments can lead to large increases in efficiency if applied properly. Against this must be set the additional work a user is put to in order to implement such procedures and the possibility that, if the antithetic sampling scheme has been ill-chosen, there may be little gain in efficiency at the end of the day, rendering the additonal work fruitless.

To assist the potential investigator in making use of antithetic techniques, we describe some sampling schemes which can be tried in commonly occurring situations. The general layout of the schemes and the conditions when they are effective will be indicated.

The schemes make use of methods of generating random samples from certain distributions, like the normal and inverse Gaussian (IG), conditional on certain sample statistics, such as the mean and variance, being known. Some are already described in the literature, others are not so well known or are new. These generators can either be used directly or they can be modified to provide samples from more general distributions.

2. LAYOUT OF SAMPLING SCHEMES

The basic simulation run is illustrated in Figure 1. Rectangles represent programs or subroutines that output random variates. Thus in Fig. 1 we have a random variate generator, A, which produces a random sample $\underline{X} = (X_1, X_2, \ldots, X_n)$ which is used in the simulation program, B, to produce a response, Y. The $\underline{X}$-generator is indicated as separate from the simulation program though, of course, in many simulations it can be regarded as being embedded within it.

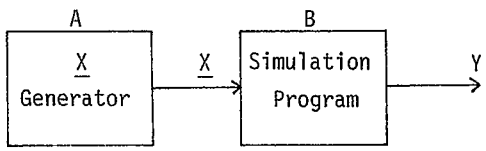

Figure 1.

Sampling in a Single Run

We shall consider two cases. The first is where it is solely the mean of Y, θ, that is of interest. The second is where other characteristics like the variance of Y, σ^2, or certain percentiles, or its cdf that are of interest. Sampling schemes

which are good for estimating θ can be bad for estimating σ^2 or the cdf of Y.

Consider first the case where it is just the mean of Y that is of interest. The central idea is to make a pair of runs in which the responses, Y,Y' are negatively correlated. Then θ can be estimated by the average

$$\theta = \tfrac{1}{2}(Y + Y').$$

This has variance

$$\text{var}(\theta) = \tfrac{1}{2}(1 + \rho)\text{var } Y$$

where ρ is the correlation between Y and Y'. This correlation can be obtained indirectly by use of a control variate, T, which can be regarded as a regressor variable on which Y depends. If the sampling of the control variate is under our control we can then arrange to correlate the variates T and T' in the two runs and thereby induce a correlation in Y and Y'.

T should be chosen as far as is possible to match the form of Y. Thus if Y is a mean, T should be a mean. Typically T will be a sample statistic of the sample $\underline{X}$. Examples are

$$T = \bar{X} \quad \text{(sample mean)}$$
$$= s^2 \quad \text{(sample variance)}$$
$$= \sum(\frac{1}{X} - \frac{1}{\bar{X}}) \quad \text{(harmonic dispersion measure)}$$
$$= X_{(1)}, X_{(n)} \quad \text{(max. and min. order statistics)}$$

Fig. 2 illustrates the sampling process. The first run is represented by blocks A and B. The value of T is calculated from the sample $\underline{X}$. We can then calculate the antithetic value T' from T (block C). To do this it is necessary to know the cdf, F(.), of T. T' can then be obtained as

$$T' = F^{-1}\{1 - F(T)\}. \qquad (2.2)$$

This is simply the inverse distribution function (IDF) transform in which a uniform variate U, here 1 - F(T), is transformed by the inverse F^{-1} into a variate T' with the same distribution as T.

In general, for those T which are means, the cdf, under the central limit theorem, tends to normality. This has two effects. Firstly, the correlation of T and T' tends to -1, the best possible. Secondly, symmetry of the normal distribution means that (2.2) is approximately $T' = 2E(T)-T$. It will often be sufficiently accurate to replace (2.2) by some piecewise polynomial in T. If a very accurate value of T' is needed, this can be obtained by Newton-Raphson iterations on (2.2) written in the form :

$$F(T') - U = 0,$$

where U = 1 - F(T). Computationally this is slow compared with generation of a single T, but in overall terms this is unimportant if variate generation is inexpensive compared with the total simulation.

The next step, once T' has been obtained, is to produce the antithetic sample $\underline{X}' = (X_1', X_2', \ldots, X_n')$

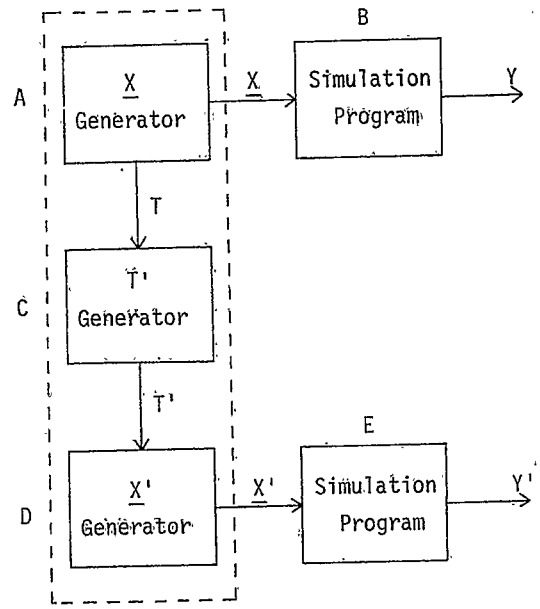

Figure 2.

Sampling for a Pair of Runs

to be used in the second run (Fig. 2, block D). The problem is to generate an X-sample, $\underline{X}'$, from the conditional distribution of $\underline{X}$ given T. This step is, statistically speaking, the most interesting one in the whole process. It is tractable only for certain X-distributions in combination with appropriately chosen T statistics. We outline a number of specific cases in the next section.

Finally, once $\underline{X}'$ has been obtained, the second run (Fig. 2 block E) can be made yielding the antithetic response Y'.

3. SAMPLING SCHEMES FOR PARTICULAR DISTRIBUTIONS

In this section we give particular examples of those combinations of distributions and sample statistics T for which it is possible to generate antithetic samples by first generating an antithetic T' and then generating $\underline{X}'$ conditional on the value of T'.

Example 1 $X \sim G(\alpha,\mu)$, a gamma variable with pdf

$$f(x) = \alpha^{\mu}x^{\mu-1}e^{-\alpha x}/\Gamma(\mu), \quad x > 0.$$

Let $T = \sum X \sim G(\alpha,n\mu)$. T fits into the format of the previous discussion in the sense that T' can be generated by (2.2). The following algorithm shows how a sample $\underline{X}'$ can be obtained with given T' value.

Algorithm RSGA : Given $T' \sim G(\alpha,n\mu)$

Step 1. Generate $W_1, W_2, \ldots, W_n$ independent $G(\alpha;\mu)$

Step 2. Set $a = T'/\sum W_j$

Return with $X'_i = aW_i \quad i = 1,2,\ldots,n$.
The X_i' are independent $G(\alpha,\mu)$ with $\sum X' = T$.
This algorithm has been mentioned by Cheng (1981, 1983a). It is essentially the result described by Aitchison (1963) as Property 4.

Example 2 $X \sim N(0,1)$, a standard normal variable. Let $T = (T_1,T_2)$ where

$$T_1 = \bar{X} \sim N(0, n^{-1})$$
$$T_2 = \sum X_j^2 - n\bar{X}^2 \ (=(n-1)s^2) \sim \chi^2_{n-1}.$$

Here T_1,T_2 are independent, so T' can still be generated by (2.2) which is now applied to each component separately. Once T' is calculated we can obtain $\underline{X}'$ by use of a method described by Pullin (1979) for generating a sample $\underline{X}$ given $\bar{X},s^2$. Cheng (1983a) gives a shorter version which makes use of the special case of RSGA with $\alpha = \mu = \frac{1}{2}$. The next example indicates that it is possible to extend this example to the bivariate case where all the first and second moments are correlated between the pair of samples.

Example 3 $X = (X_1,X_2) \sim N(\underline{0}',\underline{I})$ (bivariate normal with independent standard normal components). Let $T = (T_1,T_2,\ldots,T_5)'$ where

$$T_1 = \bar{X}_1,\ T_2 = \bar{X}_2 \sim N(0,n^{-1})$$
$$T_3 = \sum X_{1j}^2 - n\bar{X}_1^2,\ T_4 = \sum X_{2j}^2 - n\bar{X}_2^2 \sim \chi^2_{n-1}$$
$$T_5 = (\sum X_{1j}.X_{2j} - n\bar{X}_1\bar{X}_2)/\sqrt{T_3T_4} = r, \text{ say,}$$

where r has pdf $\{B(\frac{1}{2},\frac{1}{2}n-1)\}^{-1}(1-r^2)^{\frac{1}{2}(n-4)}, -1 \le r \le 1$. All five components are independent so (2.2) can again be applied. Once T' has been evaluated we can generate $\underline{X}'$ given T' using the method given by Cheng (1983b) which is based on similar ideas to that used in the univariate case.

Examples 2 and 3 allow one to generate antithetic normal samples with correlated sample means and covariance structure. The next example extends this to the inverse Gaussian (IG) distribution which has a shape parameter that enables the skewness to be altered. In situations where skewed random samples are needed, this is a more flexible distribution to use than the normal.

Example 4 $X \sim I(\mu,\lambda)$, an IG variable with pdf :

$$f(x) = (\lambda/2\pi x^3)^{\frac{1}{2}}\exp\{-\lambda(x-\mu)^2/2\mu^2x\}$$

and let $T = (T_1,T_2)$ where

$$T_1 = \bar{X} \sim I(\mu,n\lambda)$$
$$T_2 = \lambda(\sum X^{-1} - \bar{X}^{-1}) \sim \chi^2_{n-1}.$$

Then T_1 and T_2 are independent so (2.2) can again be used to calculate T_1' and T_2' separately. Cheng (1983a) gives details together with an algorithm for generating $\underline{X}'$ from T'. The quantity T_2 is a measure of dispersion and is the analogue of the variance term $(n-1)s^2$ in the normal case (see Tweedie 1957, Johnson and Kotz (1970)). There is a close analogue between this algorithm and the one described by Pullin for the normal case (c.f. also Michael et al, 1976).

The final example is fairly elementary but is included as it indicates how a sample $\underline{X}$ can be generated in which prescribed order statistics are antithetically correlated.

Example 5 $X \sim U(0,1)$. Denote by $X_{(1)},X_{(2)}\ldots,X_{(n)}$ the order statistic of a random sample $\underline{X}$. Let $T = (T_1,T_2)$ where

$T_1 = X_{(1)}$ (min. order statistic)
$T_2 = X_{(n)}$ (max. order statistic)

Now T_1 and T_2 are correlated. However we can generate T' by regarding $X_{(1)}$ as generated by the IDF method and $X_{(n)}$ as generated conditional on $X_{(1)}$, in the latter case making use of the fact that $X_{(n)}$, conditional on $X_{(1)}$, is distributed as the maximum order statistics of a sample of size (n-1) uniformly distributed on $(X_{(1)}, 1)$ (see for example Pyke, 1965). Then $X_{(1)}$ and $X_{(n)}$ can be written as

$$X_{(1)} = 1 - (1 - U_1)^{1/n}$$
$$X_{(n)} = X_{(1)} + (1 - X_{(1)})U_2^{1/(n-1)}$$

where $U_1,U_2 \sim U(0,1)$. $\underline{X}'$ can then be generated by replacing U_1 and U_2 by $(1 - U_1)$ and $(1 - U_2)$:

$$X_{(1)}' = 1 - U_1^{1/n}$$
$$X_{(n)}' = X_{(1)}' + (1-X_{(1)}')(1-U_2)^{1/(n-1)}$$

This does not treat $X_{(1)}$ and $X_{(n)}$ symmetrically in the sense that $X_{(1)}$ and $X_{(1)}'$ are more negatively correlated than $X_{(n)}$ and $X_{(n)}'$. However as the sample size increases the maximum and minimum order statistics become asymptotically uncorrelated and both antithetic pairs then tend to the same correlation.

Once $X_{(1)}'$ and $X_{(n)}'$ are generated, the full sample $\underline{X}'$ is given by generating $U_2,U_3,\ldots,U_{n-1} \sim UID(0, 1)$, setting $X_{(j)}' = U_{(j)}(X_{(n)}'-X_{(1)}')$ $(j = 2,3,\ldots,n-1)$ and returning with $\{X_{(j)}'\}$ randomly permuted.

This last example has obvious extensions to other prescribed combinations of order statistics.

4. SAMPLING SCHEMES FOR GENERAL DISTRIBUTIONS

It might appear that the previous discussion restricts the distribution of the X-sample to just a few special examples. This is not the case. Suppose that the X's have cdf $F_X(.)$ different from any of the ones considered. We now start by generating a Z-version of X where Z is one of the special cases, and then convert the Z-version to X for use in the simulation. The procedure is indicated in Fig. 3.

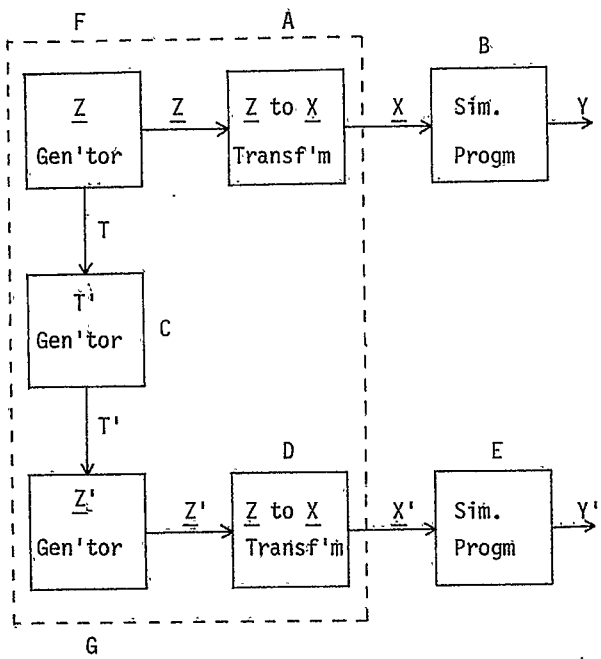

Figure 3.

Sampling a Pair of Runs with Sample Conversion

In inducing the negative correlation between runs, Z now plays the role of X in Fig. 2. The only modification is that Z must be converted to X for use in the simulation. This can again be done by the IDF transform :

$$X_i = F_X^{-1}\{\Psi(Z_i)\} \qquad (i = 1,2,\ldots,n) \qquad (4.1)$$

where $\Psi(.)$ is the cdf of Z. The transform is monotone in Z. This means that sample measures of location and dispersion of X will be correlated with corresponding quantities in the Z-sample. The correlation will be especially strong if the cdf of X is similar to Ψ as then (4.1) will be close to the identity transform X = Z. Thus any antithetic scheme which we apply to Z will have much the same effect on the X-sample. If Ψ is dissimilar to F_X this only lessens the correlation between runs; it does not alter the distributional properties of the X's which remain independent with exact cdf $F_X(.)$.

An added advantage of making Ψ resemble F_X is that the transformation (4.1) can be replaced by some simple piecewise polynomial approximation, obviating the need to evaluate Ψ and F^{-1} explicitly. Cheng (1983a) gives an example in which the IG distribution is used to approximate the extreme value distribution.

5. SAMPLING SCHEMES FOR ESTIMATING A CDF

We now examine the modification required when additional features of Y such as its variance, or selected percentiles, or its cdf are of interest. Making a set of paired runs is not effective. Instead we need to make two blocks of runs. An estimate of variance or percentile can then be obtained from the sample of Y's obtained from each block. Variance reduction is again achieved by introducing correlation between the two blocks. This paired block scheme is illustrated in Fig. 4.

Relative to $\underline{X}$, the variate S has the same interpretation that T had before : it is a sample statistic of $\underline{X}$ and it is highly correlated with Y. However we do not correlate a pair of S values directly. Each block gives rise to a sample of S values : $\underline{S} = (S_1, S_2, \ldots, S_m)$ from which we can calculate some suitable chosen statistic T. The idea is now to correlate T with the corresponding statistic T' computed from sample $\underline{S}'$ obtained from the second block.

The correlation of T with T' (Fig. 4, blocks F, C,G) is computationally the same as blocks F,C,G of Fig. 3. However instead of using S_i to generate an individual X_i we use S_i to generate a whole sample of $\underline{X}_i$. This is depicted by blocks A which are thus *computationally* identical to block G of Fig. 3 or 4.

For the scheme to be effective we want S to be similar in form to Y, and T to be a quantity (or quantities) that strongly influences the distributional characteristic of Y under investigation. For example, suppose Y is dependent on the average of the X's and we wish to estimate a percentile of the distribution of Y. Then we can take $S = \bar{X}$ and we would expect an estimator of the percentile obtained from a block of runs to be influenced by the location and dispersion of the S-sample. So take

$$T = (\bar{S}, \textstyle\sum S^2 - m\bar{S}^2).$$

When X is a normal variable, the above example has an interesting extension to the case when S is bivariate normal : $S = (\zeta_1, \zeta_2)$. T can then be the five-component statistic of Example 3, Section 3. This allows all the first and second order sample moments of $\underline{S}$ and $\underline{S}'$ to be negatively correlated. We can use ζ_1 as the sample mean of the $\underline{X}$ sample and ζ_2 can be treated as the normal-version of its sample variance and so can be converted to an exact χ^2 variate using a transform of the form (4.1). This particular scheme has applications in situations where Y depends on both the location and dispersion of the sample $\underline{X}$.

6. DISCUSSION

Certain points should be borne in mind concerning the effectiveness and ease of use of the sampling schemes described above.

In assessing efficiency we must weigh the reduction in variance against the additional work that use of an antithetic sampling scheme entails. The suggested methods of generating antithetic samples may take up to several times the length of time needed to generate independent samples. However the generation of random variates usually only takes a fraction of the time to carry out the whole simulation. For example, if the generation of variates take 10%of the

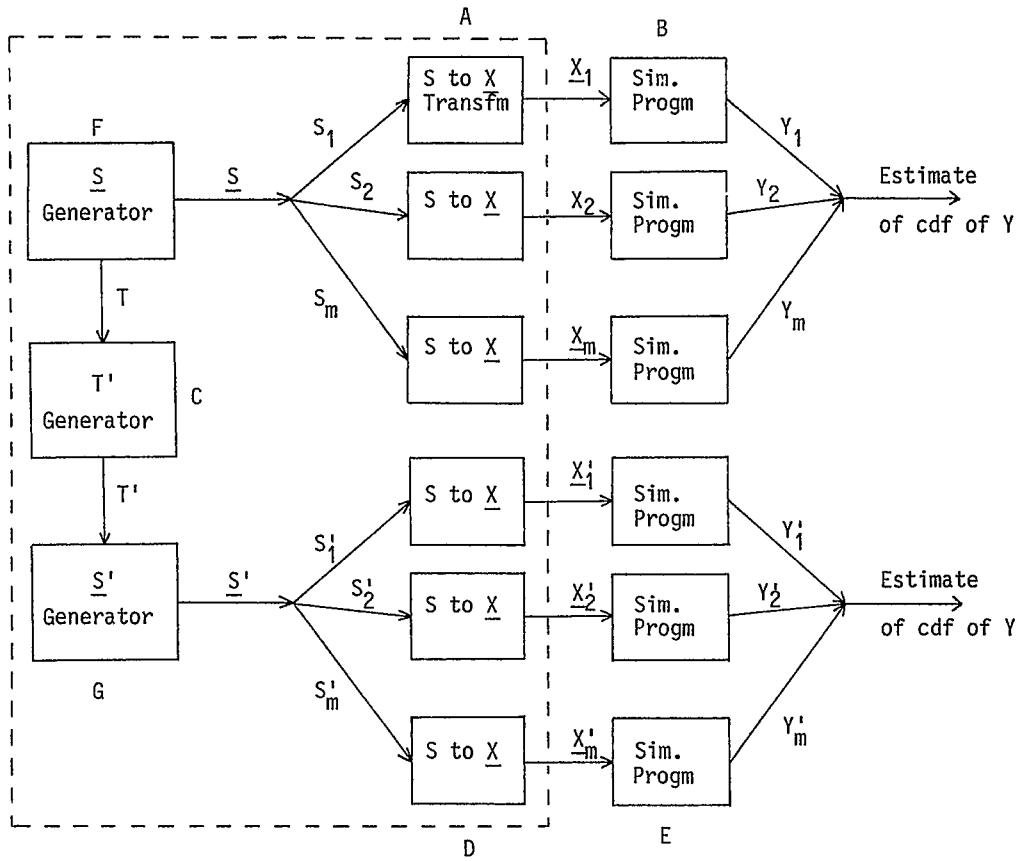

Figure 4.

Generation of a Pair of Y Samples to Estimate the cdf of Y

overall time and we take five times as long when using an antithetic scheme, this gives an overall increase in total time of 40%. The break-even point is thus obtained with a variance reduction of just 40%.

As far as variance reduction is concerned, the suggested sampling schemes are obviously more suited to some kinds of simulations than to others. An ideal simulation would be one where the stochastic input comprises a single sample $\underline{X}$ of fixed sample size. Simulations concerned with evaluating the distributional properties of test statistics such as their bias, efficiency and power comprise one important category for which our methods are particularly suited. In Figs. 2, 3 and 4, the dotted box shows that in each case the antithetic scheme can be set up as a subroutine, separate from the simulation program, outputting solely the random samples required by the simulation.

At the other end of the spectrum are those simulations connected with evaluating the performance of a complex, say queueing, system. Such simulations involve several stochastic input streams. Moreover the number of variates actually used from each stream will usually be random numbers. Our methods are not so well suited to such problems. However their use is not totally precluded. One possibility is to apply antithetic sampling to each input stream separately. The variates in each stream can be divided into groups of fixed size and antithetic sampling applied to corresponding groups in separate runs. There may be an "end effect" because different numbers of variates are used in different runs. This will result in some variates being generated at the end of runs using fewer variates, but which are not then used. The main objection to this blanket approach is the rather messy nature of the book-keeping required to keep track of the antithetic variates used in different runs. A more parsimonious method is usually desirable. For instance in bottleneck studies, though many input steams may be used, it should be possible to identify one or two streams as being critical whose sample statistics are likely to be strong candidates as good control variates. Antithetic sampling can be applied to these streams only. Successful variance reduction, as well as shortening simulation time, thus serves the added advantage of pinpointing bottleneck factors, presumably the point of such studies in the first place.

REFERENCES

Aitchison, J (1963), Inverse distributions and independent gamma distributed products of random variables, Biometrika, Vol. 50, pp. 505-508.

Cheng, RCH (1981), The use of antithetic control variates in computer simulations. WSC Proceedings-81, Vol I, Eds Ören TI, Delfosse CM and Shub CM. IEEE, New York, pp 313-318.

Cheng, RCH (1983a), Generation of inverse Gaussian variates with given sample mean and dispersion, MATH Report 83-1, Dept. of Mathematics, UWIST, Cardiff, Britain.

Cheng, RCH (1983b), Generation of bivariate normal samples with given sample mean and covariance matrix, MATH Report, 83-2, Dept. of Mathematics, UWIST, Cardiff, Britain.

Johnson NL and Kotz (1970), Continuous univariate distributions -1, Houghton-Miffin, Boston.

Michael JR, Schucany WR and Haas RW (1976), Generating random variates using transformations with multiple roots. The Amer. Stat. Vol. 30, pp. 88-90.

Pullin, DI (1979), Generation of normal variates with given sample mean and variance, J. Stat. Comput. and Simul. Vol. 9, pp. 303-309.

Pyke R (1956), Spacings (with Discussion), J.Roy. Statist. Soc. B, Vol. 27, pp. 395-449.

Tweedie MCK (1957), Statistical properties of inverse Gaussian distributions. I, Ann. Math. Statist., Vol. 28, pp. 362-377.