PROPERTIES OF CERTAIN DISCRETE DISTRIBUTIONS

SUITABLE FOR GENERATING APPROXIMATELY NORMAL VARIABLES

S. R. Searle

Biometrics Unit and Computing Center, Cornell University, Ithaca, N.Y.

BU-228-M October, 1966

ABSTRACT

Abscissae of the normal distribution can easily be found that divide it into as many equi-probable divisions as one desires. Medians of these divisions can be stored in a table and referenced with randomly generated indices, to generate variables that are approximately normal. Properties of these medians are considered in terms of their moments, in relation to those of the normal distribution. Improvements to using medians are suggested.

PROPERTIES OF CERTAIN DISCRETE DISTRIBUTIONS
SUITABLE FOR GENERATING APPROXIMATELY NORMAL VARIABLES

S. R. Searle

Biometrics Unit and Computing Center, Cornell University, Ithaca, N.Y.

BU-228-M

October, 1966

Introduction

Muller (1959) gives an excellent account of many of the numerous methods available for generating random normal deviates. Since then Marsaglia and co-workers have developed additional methods in a long series of papers (1964 and 1965, for example). The method discussed here is a simple, but fast and easy-to-program table look-up procedure based on a table of values that represent equi-probable divisions of the normal distribution.

Equi-probable areas

Consider dividing a normal distribution into 2N areas of equal probability, N areas on either side of the mean. The probability value attached to each area will be 1/2N. Let w_i for $i = 1,2,\ldots,N-1$, be the i'th abscissa on the positive side of the mean that separates the i'th area from the (i+1)'th. Then, for the standardized normal distribution of zero mean and unit variance w_i is defined by

$$\frac{1}{\sqrt{2\pi}}\int_0^{w_i} e^{-t^2/2}dt = \frac{i}{2N} \qquad \text{--- (1)}$$

for $i = 1,2,\ldots,N-1$, with $w_i = 0$. The 2N areas of equal probability 1/2N are then defined by the 2N-1 abscissae

$$-w_{N-1},\ -w_{N-2},\ \ldots,\ -w_2,\ -w_1,\ w_0,\ w_1,\ w_2,\ \ldots,\ w_{N-2},\ w_{N-1},$$

and for a normal distribution of mean μ and variance σ^2 the corresponding abscissae are μ and $\mu \pm w_i\sigma$ for $i = 1,2,\ldots,N-1$.

In using a finite number of abscissae to represent the normal distribution we are, in effect, approximating the probability distribution function by a histogram, the rectangles of which have area 1/2N, with those at either end of the distribution being of infinite length. Now a distribution is best represented in this manner by using abscissae that are, in some sense, a mean value for the rectangles they represent. To do this it seems natural to take the abscissa that is the mean value in probability, namely the median, of each rectangle. For the rectangles of base (w_0,w_1), (w_1,w_2), ..., (w_{N-2},w_{N-1}), (w_{N-1},∞) the median, m_i, of the i'th rectangle is given by $\frac{1}{\sqrt{2\pi}}\int_{w_i}^{m_i} e^{-\frac{1}{2}t^2}dt = \frac{1}{4N}$.

Calculating m_i is equivalent to recalculating a new set of w_i from (1) for 4N intervals rather than 2N. On the other hand, for the rectangles of base (w_0,w_2), (w_2,w_4), ..., (w_{N-4},w_{N-2}) and (w_{N-2},∞) the medians are immediately available, namely $w_1,w_3,\ldots,w_{N-3},w_{N-1}$ respectively. Using these, for initial probability areas of 1/2N, the positive half of the distribution is represented by N/2 abscissae that are medians of N/2 rectangles of probability value 1/N. This procedure demands that N be an even number. The complete distribution is then represented by N such abscissae, N/2 on each side of the mean:

$$-w_{N-1},\ -w_{N-3},\ \ldots,\ -w_3,\ -w_1,\ w_1,\ w_3,\ w_5,\ \ldots,\ w_{N-3},\ w_{N-1}\ .$$

The larger the value that is chosen for N, the better will be the approximation of the resulting histogram to the normal distribution.

Suppose the N medians used to represent the N areas of equal probability are stored in a vector Z(1), Z(2), ..., Z(N). Then random generation of an index between 1 and N inclusive yields a series of Z-values that are approximately normally distributed. The efficiency of such a procedure lies in its speed; its disadvantage is the need for N storage locations; and its value as an approximator to the normal distribution depends both on the task for which it is to be used, (namely the kind of problem being simulated), and on its inherent properties. The latter are now investigated.

Computer time

First it is interesting to consider the computer time involved in using the above method, which we may logically call the abscissae method. Table 1 shows comparative times for computing random normal deviates from (i) the subroutine generator provided in the FORTRAN compiler (which is based on averaging sixteen uniform random numbers) and (ii) the abscissae generator being discussed here, with 1000 abscissae. These times are based on computing 10,000 such deviates in a CDC 1604 computer. It is seen that the abscissae method, programmed in FORTRAN, with 1000 abscissae is nearly three times as fast as the subroutine provided by FORTRAN. For comparative purposes similar figures for two routines developed by Marsaglia and Bray (1964) are also shown. These routines are described as 'super' and 'small', the times published for these having been recorded on an IBM 7090 computer. Since this machine is something in the order of two to two-and-a-half times as fast as a CDC 1604 the times given by Marsaglia and Bray have been discounted by a factor of 2.25 to make them approximately comparable with CDC 1604 times. It is encouraging to see that the FORTRAN programmed abscissae method is about the same order of magnitude of speed as Marsaglia's 'small' routine. Programming the abscissae routine in machine language, without the use of subroutines, would undoubtedly speed it up still further.

Moments of 1000 medians

Properties of the 1000 points being used to represent N(0,1), the medians of 1000 areas of probability 1/1000, will be considered in terms of their moments, in relation to those of the N(0,1) distribution.

The odd-order moments of N(0,1) are zero, and the even-order moments are $(2n)!/2^n r!$ for $r = 1,2,\ldots$. In particular the 2'nd, 4'th, 6'th and 8'th moments are 1, 3, 15 and 105 respectively. And the moments of the discrete distribution are $(.001) \sum_{i=1}^{500} [(w_{2i-1})^p + (-w_{2i-1})^p]$ for $p = 1,2,3,\ldots$, with the odd-order moments being zero-similar to those of the N(0,1) distribution. The even-order moments are $(.002) \sum_{i=1}^{500} w_{2i-1}^p$ for $p = 2,4,6,\ldots$.

Comparison of the first four even-order moments is shown in Table 2, in the first three columns. The most noticeable result is that the moments of the discrete approximation of 1000 medians are consistently less than those of the N(0,1) distribution; i.e. the moments of the approximatia are biased downward. The extent of the bias is shown in Table 2. This bias is undoubtedly due to the absence from the discrete approximation of values representing points far out in the tails of the normal distribution. Their absence reduces the moments, and this reduction increases as the order of the moments increases (see Table 2).

Means instead of medians

The absence of points far out in the tails of the normal distribution can never be completely overcome in this abscissae method of simulation but it can be reduced. Each point in the discrete series is, by the process used for determining it, the median of the area it represents. Representation of these areas by their means reduces the downward bias in the moments. That this is so can be demonstrated by considering the equi-probable areas of the normal distribution as trapeziums (whose parallel sides are almost equal in length). The mean of a trapezoidal distribution of this shape exceeds its median, and hence using means produces less bias in the moments than does the use of medians.

Using medians we have w_{2i-1} to represent the area (of probability 1/1000) between w_{2i-2} and w_{2i} for $i = 1,2,\ldots,499$; and w_{999} represents the area from w_{998} to infinity. With means, the area between w_{2i-2} and w_{2i} is represented by

$$v_i = \frac{1000}{\sqrt{2\pi}} \int_{w_{2i-2}}^{w_{2i}} te^{-\frac{1}{2}t^2} dt$$

$$= 398.9423 \left(e^{-\frac{1}{2}w^2_{2i-2}} - e^{-\frac{1}{2}w^2_{2i}}\right) \quad --- (2)$$

for $i = 1,2,\ldots,499$; and for the area from w_{998} to infinity v_{500} would be used:

$$v_{500} = \frac{1000}{\sqrt{2\pi}} \int_{w_{998}}^{\infty} te^{-\frac{1}{2}t^2} dt = 398.9423e^{-\frac{1}{2}w^2_{998}}. \quad --- (3)$$

The first 4 even-order moments of these means, $.002 \sum_{i=1}^{500} v_i^p$ for p = 2,4,6 and 8 are shown in Table 2 under the heading "1000 Means". These moments are still less than those of the N(0,1) distribution, but by considerably smaller amounts than the moments of the 1000 medians. Indeed, the downward bias is less than half for the 2'nd, 4'th and 6'th moments and almost half for the 8'th moment.

Changing from medians to means

It is interesting to see that the tails of the distribution is where changing from medians to means accounts for a large portion of the reduced bias evident in Table 2. Thus in the positive tail the median is $w_{999} = 3.29053$ and with $w_{998} = 3.09023$ we have from (3) that the mean of the tail is

$$v_{500} = 398.9423/(2.7183)^{4.7747} = 3.36732 .$$

Hence replacing w_{999} by v_{500} has the effect of increasing the p'th even-order moment of the 1000 medians by an amount $.002(v_{500}^p - w_{999}^p)$. Clearly this is also the amount by which the downward biases of the moments is reduced. These biases are shown in Table 3. Shown there too are the biases for 996 medians and 4 means, 2 in each tail. For these the median of the interval penultimate to each tail, $w_{997} = 2.96773$, is replaced by the mean in the same interval which, by (2) is

$$\begin{aligned} v_{499} &= 398.9423(e^{-\frac{1}{2}w_{996}^2} - e^{-\frac{1}{2}w_{998}^2}) \\ &= 398.9423e^{-\frac{1}{2}w_{996}^2} - v_{500} , \end{aligned}$$

and with $w_{996} = 2.87816$ this reduces to $v_{499} = 2.97271$. Hence the moments for 996 medians and 4 means are those for 998 medians and 2 means with $.002[(2.97271)^p - (2.96773)^p]$ added, the downward biases being reduced by like amounts. These downward biases, together with those from Table 2, are shown in Table 3. It is clear that replacing the median by the mean in just one pair of intervals, the tails, accounts for nearly all of the reduction in bias achieved by replacing all the medians by their means.

Improvement through the method of moments

Further improvement, in terms of reducing the bias in the moments of the discrete distribution, could be made on the means by replacing some of those pertaining to intervals at the tail by abscissae that lead to no bias at all in some of the moments. This is, in effect, the method of moments applied to intervals at the tail. For example, the means of the tail intervals themselves are $\pm v_{500} = \pm 3.36732$. If these are replaced by $\pm x$ such that the second moment has no bias, x would be the solution to

$$.002(x^2 - 3.36732^2) = .0001 \ ,$$

the right-hand side being the bias in the second moment of the 1000 means (Table 2). Solutions are $x = \pm 3.47474$. Using this value in place of $\pm v_{500} = \pm 3.36732$ the discrete distribution has biases in its moments as shown in Table 4 under the heading "998 Means and 2 M.O.M. tail points". These biases differ little from those of the 1000 means, as one would expect since the only change is replacing $v_{500} = 3.36732$ by $x = 3.47474$ and the difference between the two is minor.

But if this method is applied simultaneously to the 2 intervals at each tail the reduction in bias of the moments is more noticeable. Suppose the means in the tails and the intervals penultimate thereto, $\pm v_{500} = \pm 3.36732$ and $\pm v_{499} = \pm 2.97271$ respectively, are replaced by $\pm x$ and $\pm y$ such that the 2'nd and 4'th moments of the discrete series are unbiased. Then x and y are determined by the equations

$$.002(x^2 - 3.36732^2 + y^2 - 2.97271^2) = .0001$$

and

$$.002(x^4 - 3.36732^4 + y^4 - 2.97271^4) = .0107 \ .$$

Solutions to these are

$$x = \pm 3.47066 \text{ and } y = 2.86013 \ .$$

Moments of the discrete series comprised of these and 996 means have the biases shown in Table 4. They are appreciably less than those derived earlier; indeed, to two places of decimals the only bias in the first 8 moments is that of 2.13 in the 8'th moment. The considerable reduction in bias that occurs here is undoubtedly due to the tail now being represented by $x = 3.47066$ instead of $v_{500} = 3.36732$. This is a much greater change than when only one M.O.M. tail point was fitted. One notices a further consequence of this: that the M.O.M.

tail point for the interval penultimate to the tail is $y = 2.86013$, which is less than the mean for that interval, $v_{499} = 2.97271$; and indeed it is less than the median, $w_{997} = 2.96773$. It is, in fact, less than $w_{996} = 2.87816$ and so does not even lie inside the interval it represents! Whilst this is somewhat unsatisfying there is nothing to prevent the use of this value in Z(999) in order to have the moments of the Z's as indicated in Table 4. As already mentioned, this procedure does not eliminate the problem of having no chance at all of generating values a long way from the mean – but it does make the values that are used have moments more nearly equal to those of the exact normal distribution. Increasing N, and applying the procedure for more than 2 M.O.M. tail points would increase the accuracy still further. And for the new generations of computers the large amount of storage required, N locations, may be only a small disadvantage. The speed of the table look-up procedure will remain.

References

Marsaglia, G., MacLaren, M. D., and Bray, T. A. (1964). A fast procedure for generating normal random variables. Comm. Assoc. Comp. Mach. 7, 4-10.

Marsaglia, G., and Bray, T. A. (1965). A convenient method for generating normal random variables. SIAM Review 6, 260-264.

Muller, Mervin E. (1959). A comparison of methods for generating normal deviates on digital computers. J. Assoc. Comp. Mach. 6, 376-383.

Table 1

Comparative computer times for four
methods of generating random normal deviates

Computer	Routine	Time for each call to subroutine: Recorded	Time for each call to subroutine: CDC 1604 equivalent	Rate (No. of values per sec): Recorded	Rate (No. of values per sec): CDC 1604 equivalent
CDC 1604	FORTRAN	1420 μsec	1420 μsec	700/sec	700/sec
	Abscissae	525 "	525 "	1900 "	1900 "
IBM 7090	Super*	109 "	245 "	9200 "	4100 "
	Small*	174 "	392 "	5700 "	2600 "

* Taken from Marsaglia and Bray (1964)

Table 2

Moments of order 2, 4, 6 and 8 about zero

Moments	Distributions: N(0,1)	Discrete distributions of 1000 points: 1000 Medians: Moments	1000 Medians: Downward bias from N(0,1)	1000 Means: Moments	1000 Means: Downward bias from N(0,1)
2	1	.9987	.0013	.9999	.0001
4	3	2.9645	.0355	2.9893	.0107
6	15	14.2663	.7337	14.6668	.3332
8	105	91.2445	13.7555	97.0663	7.9337

Table 3

Downward bias from N(0,1) of moments

of 4 discrete distributions of 1000 points

Moments	N(0,1)	Downward Bias from N(0,1)			
		Discrete Distribution of 1000 points			
		1000 Medians (Table 2)	998 Medians and 2 Means	996 Medians and 4 Means	1000 Means (Table 2)
2	1	.0013	.0003	.0002	.0001
4	3	.0355	.0123	.0113	.0107
6	15	.7337	.3510	.3372	.3332
8	105	13.7555	8.3745	8.2120	7.9337

Table 4

Downward bias from N(0,1) of moments of discrete distributions

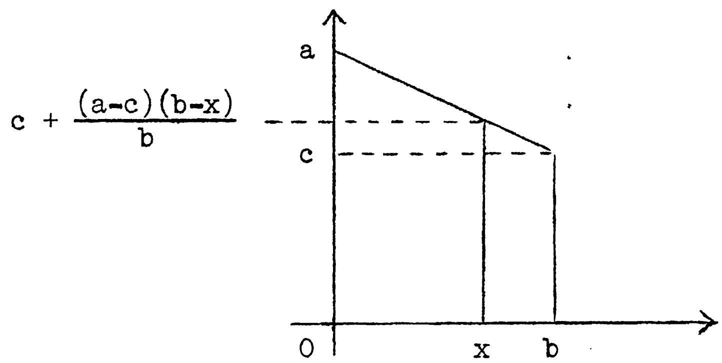
of means and M.O.M.* tail points

Moments	N(0,1)	Downward Bias from N(0,1)		
		Discrete Distribution of 1000 points		
		1000 Means (Table 2)	998 Means and 2 M.O.M. tail points	996 Means and 4 M.O.M. tail points
2	1	.0001	0.0	0.0
4	3	.0107	.0084	0.0
6	15	.3332	.2965	.0388
8	105	7.9337	7.8750	2.1305

* By M.O.M. tail points is meant points in the intervals at the tails estimated by application of the method of moments (M.O.M.).

Appendix

Table 2 illustrates the reduction in the bias of the moments of the discrete series brought about by using means instead of medians. Evidence for this reduction can be gained by considering each equi-probable and (other than the two tails) of the normal distribution as being represented by a trapezium. Consider then, the mean and median of a trapezoidal distribution of base b and parallel sides a and c:



Its mean is

$$\mu = \frac{2}{b(a+c)} \int_0^b \left[c + \frac{(a-c)(b-x)}{b}\right] x\,dx = \frac{b(a+2c)}{3(a+c)}$$

and its median is the solution for m to

$$\frac{m}{2}\left[a + c + \frac{(a-c)(b-m)}{b}\right] = \frac{b(a+c)}{4}$$

which reduces to

$$2(a-c)m^2 - 4abm + b^2(a+c) = 0 .$$

Solutions are

$$m = \frac{b\left[a \pm \sqrt{\tfrac{1}{2}(a^2 + c^2)}\right]}{a - c}$$

and because the median cannot exceed b it must be

$$m = \frac{b[a - \sqrt{\frac{1}{2}(a^2 + c^2)}]}{a - c} .$$

Now the trapeziums used in representing the N(0,1) distribution will have their parallel sides a and c differing by only a small amount relative to a. Thus

$$\frac{a-c}{a} \equiv \epsilon$$

where ϵ is small, and this gives

$$\mu = \frac{b}{3}\frac{(3-2\epsilon)}{(2-\epsilon)}$$

which, to first order terms in ϵ, is

$$\mu = b(3-2\epsilon)(1+\tfrac{1}{2}\epsilon)/6$$

$$= b(\tfrac{1}{2} - \frac{\epsilon}{12}) .$$

Likewise the median is

$$m = \frac{b\{1 - \sqrt{\frac{1}{2}[1 + (1-\epsilon)^2]}\}}{1 - (1-\epsilon)}$$

and, to first order terms in ϵ, this is

$$m = b\{1 - [1 - (\epsilon-\tfrac{1}{2}\epsilon^2)]^{\frac{1}{2}}\}/\epsilon$$

$$= b(\tfrac{1}{2} - \frac{\epsilon}{4}) .$$

Hence $\mu > m$;

i.e. the mean of a trapezoidal distribution is greater than its median. This will be true for every equi-probable area represented by the medians in our approximation to N(0,1), and hence using means rather than medians to represent these areas will reduce the downward bias in the moments of the discrete distribution.

Furthermore, the greatest difference between mean and median will be in the tail of the N(0,1) distribution, for then c/a will tend to zero and so ϵ will be at its greatest, so maximizing $\mu - m = b\epsilon/6$.