TESTING THE EFFICACY OF A
TABLE LOOK-UP GENERATOR OF NORMAL VARIABLES

S. R. Searle and D. A. Evans
Biometrics Unit and Computing Center, Cornell University, Ithaca, New York

BU-227-M December, 1966

ABSTRACT

Medians of 1000 equi-probable intervals of the standardized normal distribution were referenced by randomly generated indices to generate approximately normal random variables. Five simple statistical tests were made on a series of variables simulated by this means.

TESTING THE EFFICACY OF A
TABLE LOOK-UP GENERATOR OF NORMAL VARIABLES

S. R. Searle and D. A. Evans
Biometrics Unit and Computing Center, Cornell University, Ithaca, New York

BU-227-M December, 1966

Introduction

Suppose that w_i for $i = 1,2,\ldots,N-1$ is defined by the equation

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

with w_0 defined as zero. Then the 2N-1 values w_0 and $\pm w_i$ for $i = 1,2,\ldots,N-1$ are abscissae of the normal distribution N(0,1) having zero mean and unit variance, abscissae that divide this distribution into 2N areas of equal probability 1/2N. Furthermore, for N being an even number the N values $\pm w_1,\pm w_3,\ldots,\pm w_{N-3},\pm w_{N-1}$ are medians of N intervals of equal probability 1/N. The relationship of the first 8 moments of 1000 such medians (N = 1000) to those of the N(0,1) distribution are discussed in paper BU-228-M of the Biometrics Unit. The agreement is sufficiently close that the medians appear to be a reasonably good approximation to the standardized normal distribution. This being so, computer simulation of random normal deviates can be made from these medians by storing them in a table and referencing them via an integer index randomly generated from the first 1000 integers. This note summarizes four simple statistical tests carried out on pseudo random normal deviates simulated in this manner. Their properties do, of course, depend upon the random number generator used, namely that provided in the FORTRAN compiler for the CDC 1604 computer. This multiplicative congruential generator, of the form $u_{i+1} = au_i + c$ modulus $2^{47}-1$, is a standard type of random number generator whose properties are given in IBM (1959). Although improvements upon it have recently been suggested by MacLaren and Marsaglia (1965), it has been tested and used extensively in simulation work [Hill and Dobell (1962)].

The representation of the normal distribution by the medians of 1000 equi-probable areas was tested by carrying out 100 'experiments' of 1000 samples each, for sample sizes of $n = 4, 6, 8$ and 10 simulated normal deviates. In generating each of the 100,000 samples for the four different sample sizes a new 'seed' was used for the random number generator in order to ensure no re-cycling of the generator. For each set of 100 "experiments" (of 1000 samples) with each of the four sample sizes five tests were made.

Test 1

In this the 1000 means $\bar{x}$, based on sample size n, were computed. Since, if x is N(0,1), the mean, $\bar{x}$, is N(0,1/n), a frequency distribution was obtained over the 1000 samples, of $\bar{x}$ according into which of 10 equi-probable regions of N(0,1/n) it fell. A χ^2 goodness-of-fit value on 9 degrees of freedom was then calculated to test the deviations of these frequencies from their expected value of 100 (= 1000/10). The number of occasions on which this computed χ^2_9 value was significant at the 5% level in the 100 experiments is shown in the first section of Table 1. It is seen that significance was observed about as frequently as one would expect. Such a test is, of course, not very sensitive to normality because of the Central Limit theorem.

Test 2

On the hypothesis that x is N(0,1) then $\sum_1^n x^2 - n\bar{x}^2$ has a χ^2-distribution with n-1 degrees of freedom. For each sample $\sum_1^n x^2 - n\bar{x}^2$ was therefore computed and a frequency distribution obtained of its values relative to 4 equi-probable regions of the χ^2_{n-1} distribution. A χ^2_3 goodness-of-fit statistic was then computed for these 4 regions to test for deviations from an expected value of 250 sample χ^2_{n-1}-values in each region. The number of experiments in which this χ^2_3 value was significant at the 5% level is shown in the second section of Table 1.

Test 3

The distribution of $\Sigma x^2 - n\bar{x}^2$ is χ^2_{n-1} regardless of the mean of x. But Σx^2 has a χ^2_n distribution only if x is N(0,1). A third test therefore consisted of testing Σx^2 against χ^2_n in exactly the same manner as test 2 was carried out for $\Sigma x^2 - n\bar{x}^2$. The results are shown in the last part of Table 1.

The results of Tests 2 and 3 are not quite as good as those of Test 1, although they are well within the bounds of reasonableness, especially since only four equi-probable regions of the χ^2_{n-1} and χ^2_n distributions were used. The most efficient is probably Test 3 because it is appropriate for the normal distribution of interest, N(0,1), whereas Test 2 pertains to N(μ,1) for any mean μ. Test 1 is the least effective because of its insensitivity to normality resulting from the Central Limit Theorem.

Test 4

For samples of size n the total number of random normal deviates simulated in each of the 100 experiments was 1000n. On the hypothesis that the simulated values are N(0,1) the 95% confidence interval for each experiment mean is therefore $0 \pm 1.96/\sqrt{1000n}$. Table 2 shows the numbers of experimental means that lay outside this interval. The numbers are a little less than one would expect – a total of 13 compared to expectation of 20. The reason for this undoubtedly lies in the fact that the method of simulation being tested is poor for the tails of the distribution in that no values exceedingly far from the mean can ever be obtained.

Test 5

In Table 3 are shown the first 8 moments about zero of the N(0,1) distribution and of the 1000 medians, together with the sample moments calculated from a sample of half a million deviates simulated from the 1000 medians.

With both the N(0,1) distribution and the 1000 medians the odd-order moments are, of course, zero. And with the even-order moments those of the 1000 medians are biased downward from those of the N(0,1) distribution. This arises, in part at least, from the poor representation given to the tails of the N(0,1) distribution by the 1000 medians. In this discrete approximation the point farthest from the mean on the positive side is w_{999} which has, from equation (1) with N = 1000, a value of 3.29053; and $w_{998} = 3.09023$. Hence, in the 1000 medians the value 3.29053 represents all that portion of the N(0,1) distribution from 3.09023 to infinity. Thus in sampling the discrete series no value greater than 3.29053 is ever obtained; whereas in sampling the normal distribution such values would (with small but non-zero probability) be obtained. This accounts, in part at least, for the downward bias evident in Table 3, and for the diminished number of means lying outside the 5% confidence intervals in Table 2. Methods of correcting these discrepancies to some extent are discussed in BU-228-M.

The sample moments shown in Table 3 are all reasonably close to the corresponding moments of the 1000 medians. The odd-order sample moments are, of course, not zero, although they differ from it by very little, the value of the 7'th moment being only 0.28345; and the even-order sample moments are all within 0.15 of the corresponding moments of the 1000 medians. This appears to indicate that the random number generator is performing satisfactorily.

Approximations using fewer intervals

Initial areas of equal probability 1/2N yield N medians of N areas of equal probability 1/N. The same initial areas can also be used to yield N/k medians representing areas having probability k/N. In such a formulation N/k must be an even integer in order to retain symmetry about the mean. On the positive side of the mean the N areas of equal probability 1/2N will then be grouped into N/2k areas of equal probability k/N with medians $w_k, w_{3k}, w_{5k}, \ldots, w_{N-5k}, w_{N-3k}$, and w_{N-k} . The same medians taken negatively will represent the distribution below the mean. In this way the general procedure of representing the normal distribution by a finite series of medians can be established, using any appropriate number of equi-probable areas. In all cases N must be an even number and so must N/k, with k being an integer.

All possible sets of medians of the above form that can be derived from the initial 1999 abscissae $w_0, \pm w_1, \pm w_2, \ldots, \pm w_{999}$ are indicated in Table 4. With all of them the odd-order moments are zero; the first four even-order moments are shown in Table 5. They exhibit the same evidence of downward bias as seen in Table 3, this bias increasing both as the order of the moment increases and, even more quickly, as the number of medians decreases. It is noticeable that for the second moment the downward bias is quite small, even for as few as 100 medians (a bias of 2.3%). This would seem to indicate that for problems relating only to means simulation procedures based on such a small discrete series might not be unduly inadequate. Even in the fourth moment the downward bias for 100 medians is only 8.2%; but it is 23% for the sixth moment and 44% for the eighth moment. In situations where these higher-order moments are of importance an approximation with as few as 100 medians would be inadequate. In all cases, of course, this downward bias could be reduced by using means rather than medians to represent the equi-probable areas (see BU-228-M).

One notes that when, in Table 5, the normal distribution is approximated to by only 2 medians they are ±0.6745, and the non-zero moments are the first four even powers of 0.6745. These are the values shown in the last line of Table 5.

Table 1

100 Experiments Each of 1000 Samples

Sample Size	Number of experiments (out of 100) in which the χ^2 goodness-of-fit value was significant at the 5% level		
n	Test 1 Distribution of $\bar{x}$ against $N(0,1/n)$. χ^2_9 – using 10 regions	Test 2 Distribution of $\sum_1^n x^2 - n\bar{x}^2$ against χ^2_{n-1}. χ^2_3 – using 4 regions	Test 3 Distribution of $\sum_1^n x^2$ against χ^2_n. χ^2_3 – using 4 regions
4	5	9	6
6	7	6	9
8	5	6	4
10	3	4	6

Table 2

Test 4: Confidence intervals for experiment means

Sample Size	Confidence Interval $= \pm 1.96/\sqrt{1000n}$	Number of experiments (out of 100) for which experiment mean lay outside confidence interval
4	± .0309901	3
6	± .0253034	1
8	± .0219133	6
10	± .019600	3

Table 3

Test 5: Moments about zero

Distribution	Moments							
	1	2	3	4	5	6	7	8
N(0,1)	0	1.0	0	3.0	0	15.0	0	105.0
Discrete approximation of 1000 points	0	.99869	0	2.96454	0	14.2663	0	91.2445
Sample moments of 500,000 taken from discrete approximation	-.00030	1.00108	.00525	2.97418	.03786	14.30140	.28345	91.38290

Table 4

Discrete approximations to the standardized normal distribution using symmetric abscissae that are medians of equi-probable areas. (Each approximation is derived from 1999 abscissae $w_0 = 0, \pm w_1, \pm w_2, \dots, \pm w_{999}$, that define 2000 areas of equal probability 1/2000)

	Equi-probable areas		Positive half of distribution	
k	Number representing whole distribution 1000/k	Probability attached to each area k/1000	Number of equi-probable areas 500/k	Subscripts of abscissae w_i representing medians of equi-probable areas k,3k,5k,...,100-5k,1000-3k,1000-k
1	1000	1/1000	500	1, 3, 5,...,995,997,999
2	500	1/500	250	2, 6, 10,...,990,994,998
4	250	1/250	125	4, 12, 20,...,980,988,996
5	200	1/200	100	5, 15, 25,...,975,985,995
10	100	1/100	50	10, 30, 50,...,950,970,990
20	50	1/50	25	20, 60,100,...,900,940,980
25	40	1/40	20	25, 75,125,...,875,925,975
50	20	1/20	10	50,150,250,...,750,850,950
100	10	1/10	5	100,300,500,700,900
125	8	1/8	4	125,375,625,875
250	4	1/4	2	250,750
500	2	1/2	1	500

Table 5

Moments about zero
of the N(0,1) distribution and
discrete approximations
thereto (see Table 5)

Distribution	Moments			
	2	4	6	8
N(0,1)	1.0	3.0	15.0	105.0
Approximations				
No. of medians				
1000	.9987	2.9645	14.2663	91.2445
500	.9974	2.9362	13.8023	84.5306
250	.9948	2.8866	13.0852	75.3944
200	.9936	2.8639	12.7843	71.8885
100	.9873	2.7626	11.5782	59.2593
50	.9749	2.5940	9.8900	44.4530
40	.9688	2.5199	9.2353	39.4196
20	.9385	2.2072	6.8811	23.9973
10	.8798	1.7406	4.2282	10.9911
8	.8510	1.5540	3.3844	7.7670
4	.7124	.8807	1.1591	1.5332
2	.4549	.2070	.0947	.0428

References

MacLaren, M. D., and Marsaglia, G. (1965). Uniform random number generators. J. Assoc. Comp. Mach. 12, 83-89.

IBM (1959). Random number generation and testing. Reference Manual C20-8011, International Business Machines, White Plains, N.Y.

Hull, T. E., and Dobell, A. R. (1962). Random number generators. SIAM Review 4, 230-254.

Searle, S. R. (1966). Properties of certain discrete distributions suitable for generating approximately normal variables. BU-228-M of the Biometrics Unit, Cornell University, Ithaca, N.Y.