ANALYSIS OF THE PROPERTIES OF TWO BINOMIAL RANDOM NUMBER GENERATORS

Susan J. Kenny, University of Oklahoma Health Sciences Center
J. Paul Costiloe, University of Oklahoma Health Sciences Center
Andrew I. Cucchiara, University of Oklahoma Health Sciences Center

INTRODUCTION

Random number generators are used in simulation studies to produce sets of random variables that arise from distributions with known parameters. Simulation studies are conducted with these datasets to test models of interest to researchers. Typical examples of simulation studies include tests of the assumptions of a statistical model or formulation of equations to predict the rate of spread of a disease within a population. Since simulation studies can have important consequences for research, the random number generators used to produce simulated data must be reliable.

The purpose of this paper is to examine the distributional properties of random binomial variables produced by each of two random number generators. The first generator examined is the RANBIN function of SAS(R) Institute Inc. The second generator is a derived procedure that uses a uniform random variable to sample directly from an expected binomial distribution. This method uses the UNIFORM and the PROBBNML functions of SAS software.

METHODS

To test the RANBIN function, 10,000 random variates were generated using the CALL RANBIN subroutine. Sample sizes (N) of 100, 500, 750 and 1,000 were used with parameter values (π) of 0.001, 0.01, and 0.1. A chi-square (χ^2) goodness of fit statistic was computed for the observed frequency distribution to test the hypothesis that the distribution is binomial.

An independent method was derived to produce binomially distributed random variates by using the UNIFORM function to sample from a binomial distribution. This binomial distribution was produced using the PROBBNML function.

As an illustration, consider the expected binomial distribution for a sample size of 100 and π value of 0.2. The cumulative probability of this distribution can be obtained by using the PROBBNML function with parameters 100 and 0.2; that is

$$P\{X <= x \mid N, \pi\} = \sum_{X=0}^{x} \left[\begin{array}{c} 100 \\ X \end{array} \right] 0.2^X * 0.8^{100-X}$$

If one million samples, each of size 100, were independently drawn from this binomial distribution, then the expected frequency of a random variable, X, with a value less than or equal to x, would be $P[X <= x] * 1{,}000{,}000$. To select a random variate from this binomial distribution of one million values, a uniformly distributed random variate, U, multiplied by 1,000,000 was used to locate the corresponding value of X in the cumulative binomial distribution. For example, suppose the UNIFORM function returned the value of 0.345982. Multiplying this value by 1,000,000 yields 345,982. Therefore, the value of X which satisfied

$$\{ P[X = x-1] * 1{,}000{,}000 < 345{,}982 <= P[X = x] * 1{,}000{,}000 \}$$

would be selected as one of the random variates.

Ten thousand random variates were selected in this manner from binomial distributions with parameters N of 100, 500, 750 and 1000 and π of 0.001, 0.01, and 0.1. A χ^2 goodness of fit statistic was obtained for each of the observed distributions to test the hypothesis that the distribution is binomial.

RESULTS

Our results indicate that the RANBIN function does not perform well when the expected value of the binomial distribution is small. RANBIN produces an overabundance of values at the low end of the distribution and a corresponding underproduction of values at the high end of the distribution. RANBIN consistently produced significant χ^2 goodness of fit values when sample sizes were less than 1,000 and π values were less than 0.1. RANBIN began to produce variates that were distributed as expected when sample size reached 1,000 with a π value of 0.1; yet, more than half of the simulations performed for this distribution were significantly different from the expected distribution.

In contrast, the derived method produced random variates that were close to the expected distribution for all combinations of the parameters of sample size and π value. This derived method does not appear to be influenced by small expected values of the binomial distribution unlike the RANBIN method.

Our results displayed consistent patterns for all combinations of parameter values. Therefore, only representative combinations of sample size and π values examined are presented here. The results for sample size of 1,000 were chosen for presentation (Tables 1-3) since they have the largest number of degrees of freedom for the distributions.

DISCUSSION

Many simulation studies may be designed to investigate a model that has a binomial response variable with a low probability of occurence. In these situations, even with large sample sizes, the expected value of the distribution will be small. The use of RANBIN to generate simulated values in this situation would produce a distribution that had an excess of small values. The interpretation of the results from a simulation study that used RANBIN would be biased in favor of a positive response of the binomial variable of interest. Thus, a simulation study may show, for example, that a new drug results in a number of positive responses for survival that is inflated due to the biased output from RANBIN.

As with all random number generators, caution must be exercised when producing a random set of

variates. Calculation of a goodness of fit χ^2 for the simulated distribution is suggested when using any random number generator. Our conclusion is that RANBIN should be used with caution when π values less than 0.2 or greater than 0.8 are of interest.

For the parameter combinations investigated, the method derived by the authors performed well for all sample sizes and parameter values. This method could be used to generate random variates from other probability functions, given access to a function that adequately calculates the probabilty of values in the desired cumulative distribution. The SAS program for our derived method and for the investigation of RANBIN is presented in Appendix A.

Reprint requests or inquiries may be directed to:
Susan Kenny
Dept. Biostatistics and Epidemiology CHB-309
University of Oklahoma, PO Box 26901
Oklahoma City, Oklahoma 73190

Table 1. Expected and observed distributions for N=1000 and p=0.001.

X	E[X]	DERIVED	χ^2	RANBIN	χ^2
0	3677	3633	0.53	2630	298.13
1	3681	3752	1.37	4032	33.47
2	1840	1803	0.74	2653	359.22
3	613	637	0.94	612	0.00
4	153	132	2.88	68	47.22
5	30	37	1.63	4	22.53
6	6	6	0.00	1	4.17
		χ^2_{GOF}	8.1		764.7
		p	0.32		0.00

Table 2. Expected and observed distributions for N=1000 p=0.01

X	E[X]	DERIVED	χ^2	RANBIN	χ^2
2	26	20	1.38	78	104.00
3	74	82	0.86	110	17.51
4	186	172	1.05	242	16.86
5	375	363	0.38	334	4.48
6	627	650	0.84	555	8.27
7	900	843	3.61	804	10.24
8	1128	1174	1.87	1014	11.52
9	1256	1241	0.18	1205	2.07
10	1257	1268	0.09	1289	0.82
11	1143	1159	0.22	1198	2.65
12	952	955	0.01	1036	7.41
13	731	742	0.16	815	9.65
14	520	523	0.02	533	0.32
15	345	327	0.94	370	1.81
16	215	217	0.02	220	0.12
17	126	124	0.03	123	0.07
18	69	76	0.71	45	8.35
19	36	39	0.25	24	4.00
20	18	16	0.22	4	10.89
21	8	5	1.13	1	6.13
22	7	4	1.29	0	7.00
		χ^2_{GOF}	15.3		234.2
		p	0.8		0.000

Table 3. Expected and observed distributions for N=1000 and p=0.1.

X	E[X]	DERIVED	χ^2	RANBIN	χ^2
72	12	9	0.75	23	10.08
73	6	5	0.17	4	0.67
74	8	10	0.50	8	0.00
75	11	11	0.00	11	0.00
76	15	18	0.60	23	4.27
77	20	17	0.45	20	0.00
78	26	30	0.62	24	0.15
79	34	33	0.03	48	5.76
80	43	44	0.02	42	0.02
81	55	73	5.89	52	0.16
82	68	66	0.06	77	1.19
83	84	94	1.19	74	1.19
84	102	104	0.04	90	1.41
85	122	121	0.01	135	1.38
86	144	132	1.00	144	0.00
87	168	187	2.14	153	1.33
88	194	200	0.18	154	8.25
89	221	207	0.89	227	0.16
90	248	285	5.52	212	5.22
91	276	300	2.09	244	3.71
92	303	306	0.03	287	0.84
93	329	305	1.75	325	0.05
94	352	370	0.92	318	3.28
95	373	399	1.81	373	0.00
96	391	376	0.57	403	0.37
97	405	367	3.56	390	0.56
98	415	393	1.17	442	1.76
99	420	370	5.95	428	0.15
100	420	447	1.73	423	0.02
101	416	394	1.16	402	0.47
102	407	396	0.29	418	0.29
103	395	383	0.36	417	1.22
104	378	360	0.86	395	0.76
105	359	367	0.17	389	2.50
106	336	307	2.50	324	0.42
107	312	323	0.39	310	0.01
108	287	287	0.00	295	0.22
109	261	281	1.53	248	0.65
110	235	252	1.22	239	0.07
111	209	191	1.55	220	0.58
112	184	193	0.44	200	1.39
113	161	165	0.09	195	7.18
114	139	137	0.03	142	0.06
115	119	119	0.00	135	2.15
116	101	102	0.01	98	0.09
117	85	106	5.18	90	0.29
118	71	75	0.22	64	0.69
119	58	45	2.91	60	0.07
120	47	58	2.57	64	6.15
121	38	32	0.94	30	1.68
122	31	28	0.29	25	1.16
123	24	22	0.17	22	0.16
124	19	22	0.47	14	1.31
125	15	16	0.07	13	0.27
126	11	11	0.00	13	0.36
127	9	12	1.00	6	1.00
128	7	9	0.57	7	0.00
129	18	28	5.55	11	2.72
		χ^2_{GOF}	70.3		86.0
		p	0.13		0.01

Appendix A.
SAS code (vers. 6.03) used for the derived method and to test RANBIN.

```
DATA V500_1;
RETAIN SKYSQ SKYSQU DF TOT_C TOT_E
       TOT_CU A Z U  0  POINTS 10000;
SEED1=INT(UNIFORM(0)*1000000)*10+7;
N=500; PO=.1; POPSIZE=1000000;
ARRAY FREQ (501) _TEMPORARY_;
ARRAY CU (501) _TEMPORARY_;
ARRAY C (501) _TEMPORARY_;
ARRAY E (501) _TEMPORARY_;
DO I=1 TO N+1; FREQ(I)=0; CU(I)=0;
   C(I)=0; E(I)=0;
END;
DO I=0 TO N; J=I+1;
   XX=PROBBNML(PO,N,I);
   FREQ(J)=INT(POPSIZE*XX+.5);
END;
DO I=0 TO N;
   XX=PROBBNML(PO,N,I); PP=XX-A; A=XX;
   E(I+1)=INT(POINTS*PP+.5);
END;
DO H=1 TO POINTS;
   U=UNIFORM(0);
   II=INT(U*FREQ(501)+1.5);
   IF II<1 THEN II=1;
   IF II>=FREQ(501) THEN DO;
     I=N+1; GOTO SKIP; END;
   I=1;
   DO WHILE (FREQ(I)<=II); I=I+1; END;
   SKIP:  CU(I)+1;
   CALL RANBIN(SEED1,N,PO,Z); C(Z+1)+1;
END;
RETAIN EXP_LO CU_LO RB_LO 0 EXP_HI
       CU_HI RB_HI  0;
L=1;
DO UNTIL (EXP_LO>5 AND E(L)>5);
   EXP_LO=EXP_LO+E(L);
   CU_LO=CU_LO+CU(L);
   RB_LO=RB_LO+C(L);
   L+1;
END;
E(L-1)=EXP_LO; CU(L-1)=CU_LO;
C(L-1)=RB_LO;  LO=L-1;
H=N+1;
DO UNTIL (EXP_HI>5 AND E(H)>5);
   EXP_HI=EXP_HI+E(H);
   CU_HI=CU_HI+CU(H);
   RB_HI=RB_HI+C(H);
   H=H-1;
END;
E(H+1)=EXP_HI; CU(H+1)=CU_HI;
C(H+1)=RB_HI; HI=H+1;
DO J=LO TO HI; X=J-1; KYSQ=.; KYSQU=.;
   IF E(J)>0 THEN DO;
     KYSQ=((C(J)-E(J))**2)/E(J);
     SKYSQ+KYSQ; DF+1;
     TOT_C+C(J); TOT_E+E(J);
     KYSQU=((CU(J)-E(J))**2)/E(J);
     SKYSQU+KYSQU; TOT_CU+CU(J);
     SIGNIF=1-PROBCHI(SKYSQ,DF);
     SIGNIFU=1-PROBCHI(SKYSQU,DF);
     EXP=E(J); UN1=CU(J); RAN=C(J);
     OUTPUT;
   END;
END;
RUN;
```