Coupon Collector's Test for Random Digits

1. Introduction. Increasing use of random numbers, especially in Monte Carlo procedures and in large computing installations, has served to focus attention on the various tests for randomness. KENDALL and BABINGTON-SMITH[1] list four tests for so-called local randomness. While not giving the coupon collector's test (to be described below) a place in their now classical list of four tests, they did use a modified coupon collector's test in some of their investigations.

In an ordered set of digits, say, one may count the length of a sequence, beginning at a specified position, necessary to give or include the complete set of all ten digits. Or one may count the length required to give a set of k, $k < 10$, different digits. The distribution of these observed lengths for different initial positions can then be compared with a theoretically computed distribution. Such a test will be called the coupon collector's test from an analogy with certain sales promotion schemes.

The theoretical distribution may be computed from formulas given by H. von SCHELLING,[2] which formulas hold for the case where the individual category probabilities might be unequal. For the random digital case with category probabilities equal to 1/10, von SCHELLING's formulas simplify readily and may be conveniently related to the "differences of zero." These latter quantities are tabulated by FISHER and YATES[3] up to sequences of length 26. But in using the coupon collector's test for a complete set of all 10 digits it has been found that the mean of the length distribution is slightly greater than 29, and a table of probabilities associated with the sequence lengths 10 to 26 inclusive would hardly give a realistic picture of the entire distribution.

The present author has therefore extended this tabulation, and exact probabilities are given for sequence lengths $10 \leq n \leq 35$ and approximate probabilities for sequence lengths $36 \leq n \leq 75$. The probabilities were computed from the relation

$$p_n = \frac{1}{10^{n-1}} \sum_{j=0}^{q} (-1)^j \binom{q}{j} (q - j)^{n-1},$$

and are listed in Table 3.

If one were interested in sequence lengths necessary to obtain five different digits, the mean of this distribution is approximately 6.46. The range 5 to 26 inclusive available from FISHER and YATES[3] might be sufficient here.

2. Sequence Lengths for Decimal Expansions of π and e. It would be a simple matter to program a large digital computing machine so that it would tabulate the distribution of the sequence lengths needed for complete sets for a given ordered digital collection. However, the author did not have such a digital computing machine available, and he made a tabulation by hand for the decimal expansion of π. The 2035 decimal approximation to π given by GEORGE W. REITWIESNER[4] was used as the raw material for this count. Beginning with the initial position 3 in $\pi \simeq 3.14159\cdots$, it was recorded that a sequence length of 33 positions was needed to get all the ten digits. Beginning anew with the thirty-fourth position digit (which is a 2), it was recorded that a sequence of 18 positions was needed to get a complete set of all the 10 digits. Continuing this procedure, 67 sequences of complete sets were obtained, plus an incomplete sequence (at the end of the decimal expansion) of length 15. It was considered advisable to make the sequences non-overlapping as described above since there is considerable dependence among the set of sequence lengths if every position in the decimal expansion of π is regarded as a new starting point.

The sequence lengths for π are also included in Table 3. This tabulation for π was checked by Mr. WAYNE JONES of the Department of Defense, Washington, D. C.

Mr. JONES also made a tabulation based on the decimal expansion of e. REITWIESNER[4] gave a 2010 decimal approximation to e. An additional 490 places was given by METROPOLIS, REITWIESNER and von NEUMANN.[5] Mr. JONES found 82 complete sequences using 2486 digits in the expansion of e. This tabulation is also given in Table 3. The author desires to thank Mr. JONES for this count.

3. Statistical Tests. The mean and the standard deviation of the theoretical distribution may be computed from results given by von SCHELLING[2] or FELLER.[6] These theoretical values and the corresponding observed values for π and e are given below.

TABLE 1

	Theoretical	Observed π	Observed e
Mean	29.29	30.16	30.32
Standard deviation	11.21	11.83	10.64

To use a chi-square test, it is desirable that the expected values all exceed 10 in size. Since the sample size for π is small (67) some grouping of the sequence lengths is necessary to meet this desired minimum. The following results were obtained for a convenient grouping.

TABLE 2

	π		e	
Sequence lengths, n	Observed	Expected	Observed	Expected
10–19	13	11.604	12	14.202
20–23	13	11.720	11	14.344
24–27	9	11.491	14	14.064
28–32	5	11.480	15	14.050
33–39	13	10.195	17	12.477
40 and over	14	10.510	13	12.863
Totals	67	67.000	82	82.000
Chi-squared test values	6.436		2.826	

Neither of these chi-square test values is unusually out of line. It has been previously reported[5,7] that (using a sample of 2000 digits for e) excessive flatness in the single frequencies was noted, and an indication was obtained that the single digits in e are "non-random."[5] Apparently, this phenomenon did not reflect itself

TABLE 3

Table of Probabilities and Empirical Distributions

n	p_n	Observed for π	Observed for e
10	.0003 6288	0	0
11	.0016 3296	0	0
12	.0041 9126 4	0	0
13	.0080 9315 2	0	0
14	.0130 4560 8576	0	1
15	.0186 3435 9744	1	2
16	.0243 5958 6451 2	1	1
17	.0297 8461 8864	6	2
18	.0345 7819 0373 1264	3	2
19	.0385 2892 7611 5744	2	4
20	.0415 3577 5577 4998 4	2	3
21	.0435 8654 2461 1780 8	2	2
22	.0447 3311 6259 6932 2752	3	3
23	.0450 6836 4358 6388 8896	6	3
24	.0447 0706 5704 2485 9072	2	3
25	.0437 7151 9771 4451 888	1	1
26	.0423 8153 3617 2618 0440 544	3	5
27	.0406 4806 4094 7299 5986 8	3	5
28	.0386 6968 0608 0430 0677 8256	2	3
29	.0365 3106 7596 0890 3842 8456	2	1
30	.0343 0291 1584 0099 4076 1298 5728	0	4
31	.0320 4266 1164 2497 4751 5573 3056	1	2
32	.0297 9578 2029 8315 1051 7926 5414 4	0	5
33	.0275 9724 1577 4565 5030 9198 2083 2	1	2
34	.0254 7304 5949 3494 9321 3424 0393 4016	2	2
35	.0234 4171 8456 6112 5667 0619 1553 5264	2	3
36	.0215 1565 5696 6141 0012	3	3
37	.0197 0233 0293 7275 2140	2	5
38	.0180 0533 0690 9430 5978	3	1
39	.0164 2524 1844 5333 7918	0	1
40	.0149 6037 8429 7183 4300	1	3
41	.0136 0738 6073 5944 1433	1	1
42	.0123 6172 7525 9630 8189	1	1
43	.0112 1807 0507 6953 1223	6	1
44	.0101 7059 2895 9431 7444	1	0
45	.0092 1321 9356 5092 1003	1	1
46	.0083 3980 1802 1014 6739	0	0
47	.0075 4425 4318 8464 5255	0	1
48	.0068 2065 1566 0407 8968	0	1
49	.0061 6329 8170 0629 7386	0	0
50	.0055 6677 5325 1020 3197	0	0

TABLE 3—*Continued*

n	p_n	Observed for π	Observed for e
51	.0050 2596 9683 5475 6580	0	1
52	.0045 3608 8658 5862 1077	0	0
53	.0040 9266 5455 6666 7056	0	1
54	.0036 9155 6480 2793 7180	0	0
55	.0033 2893 3218 7175 0148	0	0
56	.0030 0127 0238 7102 7949	0	0
57	.0027 0533 0592 0946 3793	0	0
58	.0024 3814 9607 8519 0648	0	0
59	.0021 9701 7828 5704 8789	1	0
60	.0019 7946 3656 1777 4941	1	0
61	.0017 8323 6124 7283 4517	0	0
62	.0016 0628 8101 7164 6560	0	0
63	.0014 4676 0128 6430 1251	0	0
64	.0013 0296 5041 3524 9640	0	0
65	.0011 7337 3456 8177 1422	0	0
66	.0010 5660 0172 2241 9129	0	1
67	.0009 5139 1491 6632 0338	0	0
68	.0008 5661 3473 2828 0290	0	0
69	.0007 7124 1073 6049 1625	0	0
70	.0006 9434 8154 4810 4916	0	0
71	.0006 2509 8310 7043 5014	0	0
72	.0005 6273 6471 7289 2795	0	1
73	.0005 0658 1228 5628 1531	0	0
74	.0004 5601 7836 1436 4579	0	0
75	.0004 1049 1841 9142 4169	0	0
76 and over	.0036 9745 8744 5702 0432	1 (77)	0
	Total	67	82

in materially changing the characteristics of the sequence length distribution for the coupon collector's test. Some question arises as to whether the single frequency test and the coupon collector's test are independent, and also which test has the greater power.

The chi-square test values in Table 2 were calculated by assuming that the sequence lengths for complete sets of digits are independent draws from a known (infinite) multinomial probability distribution. (Null hypothesis.) The alternatives would include unspecified sorts of dependency and other underlving probabilities different from those given in Table 3.

ROBERT E. GREENWOOD

The University of Texas
Austin, Texas

[1] M. G. KENDALL & B. BABINGTON SMITH, "Randomness and random sampling numbers," Royal Stat. Soc., *Jn.*, v. 101, 1938, p. 147–166.

[2] H. von SCHELLING, "Auf der Spur des Zufalls," *Deutsches Statisches Zentralblatt*, v. 26, 1934, p. 137–146. Also, "Coupon collecting for unequal probabilities," *Amer. Math. Mon.*, v. 61, 1954, p. 306–311.

[3] R. A. FISHER & F. YATES, *Statistical Tables for Biological Agriculture and Medical Research.* 3rd edition, London, 1948, Table XXII.
[4] GEORGE W. REITWIESNER, "An ENIAC determination of π and e to more than 2000 decimal places," *MTAC*, v. 4, 1950, p. 11–15.
[5] N. C. METROPOLIS, G. REITWIESNER, & J. VON NEUMANN, "Statistical treatment of values of first 2000 decimal digits of e and π calculated on the ENIAC," *MTAC*, v. 4, 1950, p. 109–111.
[6] W. FELLER, *Probability Theory and Its Applications.* Volume 1, New York, 1950, p. 175–181.
[7] F. GRUENBERGER, "Further statistics on the digits of e," *MTAC*, v. 6, 1952, p. 123–134.

A Method for the Evaluation of a System of Boolean Algebraic Equations

With the advent of large scale electronic devices whose logical design is described by a system of Boolean algebraic equations, a method to mechanize the evaluation of such a system and shorten this evaluation with respect to time will be increasingly useful. Such a method will be described in this paper.

The problem may be described as follows: Given a set of n variables, Q^k, $(k = 1, 2, \cdots, n)$ each of which may take on the value 1 (true) or 0 (false) at any time t; then the value of any Q^k at time $t + 1$ may be defined by the system of Boolean equations

$$R_t^k = f_k(Q_t^q) \tag{1}$$

$$S_t^k = g_k(Q_t^q) \tag{2}$$

$$Q^k{}_{t+1} = \phi(Q_t^k, R_t^k, S_t^k) \tag{3}$$

where $1 \leq q \leq n$. For example, the recirculation loop of a dynamic flip-flop may be defined simply by

$$Q^k{}_{t+1} = \phi(Q_t^k, R_t^k, S_t^k) = R_t^k.$$

In another system, a more complex definition

$$Q^k{}_{t+1} = \phi(Q_t^k, R_t^k, S_t^k) = Q_t^k \cdot \overline{R_t^k} \cdot \overline{S_t^k} + R_t^k \cdot \overline{S_t^k} + \overline{Q_t^k} \cdot R_t^k \cdot S_t^k$$

may be taken, where R_t^k and S_t^k are the two inputs to flip-flop Q^k.

We shall use the symbols for conjunction, disjunction, and negation

$Q^1 \cdot Q^2$	"Q^1 and Q^2"	conjunction
$Q^1 + Q^2$	"Q^1 or Q^2"	disjunction
$\overline{Q^1}$	"Not Q^1"	negation

which are defined by the truth tables[1]

Q^1	Q^2	$Q^1 \cdot Q^2$	$Q^1 + Q^2$	$\overline{Q^1}$
0	0	0	0	1
0	1	0	1	1
1	0	0	1	0
1	1	1	1	0.

A "term" is defined as one or more variables conjoined together, e.g., $Q^1 \cdot Q^2 \cdot \overline{Q^3}$; and an "equation" as M terms, T_m, $(m = 1, 2, \cdots, M)$ disjoined together, e.g., $Q^1 \cdot Q^2 \cdot \overline{Q^3} + Q^1 \cdot Q^4$. Now note that the value of a term is zero if any variable in