Probabilities of Clumps in a Binary Sequence (and How to Evaluate Them Without Knowing a Lot)

DAVID M. BLOOM
Brooklyn College of CUNY
Brooklyn, NY 11210

1. Introduction When I was growing up in the 1940s and early '50s, my father, though a non-mathematician, encouraged my already strong interest in mathematics by bringing home books for me: problem and puzzle books, Hogben's *Mathematics for the Million*, Kasner and Newman's *Mathematics and the Imagination*, and others. In December 1985, when my son Eric and I visited him to celebrate his 80th birthday, we found that Dad hadn't changed his ways. He had picked up a copy of one of Martin Gardner's books (namely **[1]**), thinking that Eric and/or I might find in it items of interest. (Martin Gardner needs no introduction to most readers. He wrote regularly about mathematics for *Scientific American* magazine for many years and has written many books—some of them published by the MAA—about mathematical puzzles, curiosities, etc.)

What happened next was this: Eric (then 10) took a look at **[1]**, found in it the assertion (p. 124) that in an ordinary shuffled deck of 52 cards

> "there will almost always be a clump of six or seven [consecutive] cards of the same color," (0)

took out a deck of cards and did the experiment, obtained no such "clump," and came to me for an explanation. *Question*: Did Eric witness an extremely unlikely occurrence, or was **[1]** wrong? That is,

> In the case $(m, k, t) = (26, 26, 6)$, what is the probability that, in a random string of m red and k black objects, some t consecutive objects have the same color? (1)

Essentially the same problem, in a different guise, came to my attention more recently. In December 1992, I had to give a class test and a final exam in a required course for non-majors. To make it harder for a student to copy, I wrote two versions ("odd" and "even") of each exam. At the class test, I gave out the two versions alternately according to where the students had chosen to sit. Afterward, upon marking "o" or "e" (16 odds, 15 evens) next to each name on my alphabetically arranged roster, I was surprised to find that no three consecutive names had had the same version of the test. Even more surprising, the same thing happened at the final exam: Out of 15 "odds" and 17 "evens," no three alphabetically consecutive names had the same version. *Question*: Did I witness an extremely unlikely pair of occurrences, or was my surprise unwarranted? That is,

> In the cases $(m, k, t) = (16, 15, 3)$ and $(15, 17, 3)$, what is the probability that, in a random string of m "odd" and k "even" objects, no t consecutive objects have the same parity? (2)

Clearly, (1) and (2) are different cases of the same problem. What follows is a discussion of some elementary ways to solve it. In particular, we exhibit two (equivalent) recurrences for the probabilities, both of which can be proved by straightforward counting arguments. I found one of these recurrences purely by trial and error (how I did so is described in §3). Subsequently, David M. Jackson of the University of Waterloo showed me a simpler one, which we exhibit (with proof) in §5. In §2, a non-recurrence method is discussed briefly.

2. A false start, and an answer to question (1) What is the probability of a t-clump (some t or more consecutive cards of the same color) in the shuffled deck? When first trying to answer this question, I had a silly mental lapse. I reasoned as if the colors of successive cards were independent (as of course they are not!); i.e., as if the problem were to find the probability $P_t(n)$ that in n consecutive coin-tosses some t-clump occurs. The latter problem is easier than (1). Indeed, there are just two mutually exclusive ways that a t-clump can appear among n tosses: Either (i) a t-clump occurs among the first $n-1$ tosses, or (ii) the last t tosses form a clump, its type (heads or tails) is opposite to that of the $(n-t)$-th toss (unless $n-t=0$), and no t-clump occurs among the first $n-t$ tosses. Thus,

$$P_t(n) = P_t(n-1) + 2^{-t}(1 - P_t(n-t))$$

when $n > t$ (and we have $P_t(n) = 0$ when $n < t$, $P_t(t) = 2 \cdot 2^{-t}$). This recurrence for $P_t(n)$ was easily incorporated into a computer program and produced the value $P_6(52) = .5595\ldots$, which certainly would call into question Gardner's "...almost always... ." So I told Eric, back then in 1985. Three or four years later, I realized that I'd solved the wrong problem!

(Before addressing the *right* problem, note the intuitive likelihood that the correct probability of a 6-clump in the shuffled deck is even smaller than the value of $P_6(52)$ obtained above. If the first coin-toss is heads, the second has a 50% chance of being heads also; but if the first card is red, the probability that the next card is also red is only 25/51.)

OK, what next? It occurred to me to try the well-known *principle of inclusion-exclusion*, which states that if $A_1, \ldots, A_n$ are events and P denotes probability, then

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = B_1 - B_2 + B_3 - \cdots + (-1)^{n-1} B_n \tag{3}$$

where

$$B_1 = \sum_i P(A_i); \quad B_2 = \sum_{i<j} P(A_i \cap A_j); \quad B_3 = \sum_{i<j<k} P(A_i \cap A_j \cap A_k); \quad \text{etc.}$$

If G is the probability that a 6-clump occurs in a randomly shuffled 52-card deck, then G will equal the quantity (3) if we let n be a sufficiently large integer and then define A_i to be the event that some 6-clump *begins* with the i-th card in the deck; i.e., that cards $i, i+1, \ldots, i+5$ have the same color *and* this color is opposite to that of card $i-1$ if $i > 1$. Thus, e.g., $P(A_i) = 0$ if $i > 47$, $P(A_i \cap A_j) = 0$ if $i < j < i+6$, etc.; and $B_k \neq 0$ only for $1 \leq k \leq 8$. I won't make you wade through the calculations. Suffice it to say that two days' work with a hand calculator (a computer wasn't needed!) produced the bounds

$$.4640 < G < .4644, \tag{4}$$

indicating that a 6-clump won't even appear in the shuffled deck *half the time*—a result wholly incompatible with statement (0). Not sure that I myself hadn't erred, I tried a "random" simulation by computer. Among 2000 simulated "shuffled decks," only about 45% had 6-clumps, a figure roughly 1.3 standard deviations from the value (4) (but in the same ballpark). **[1]** was wrong, after all.

3. A better method The foregoing method, though viable for the particular parameters $(m, k, t) = (26, 26, 6)$ (using the notation of (1)), is far too cumbersome for the general case, which calls for a general recurrence. Yet I had already tried, and failed, to find such a recurrence, using a table obtained by brute-force enumeration for $t = 3$, the smallest nontrivial value of the clump-length t. To fix notation, let $C_t(m,k)$ denote the number of strings of m indistinguishable objects of Type A and k indistinguishable objects of Type B (say 1's and 0's) in which *no* t-clump (run of length t) occurs. The following table (5) gives values of $C_3(m, k)$ for small m, k.

VALUES OF $C_3(m, k)$

$m \backslash k$	0	1	2	3	4	5	6	7	8	9	10	11	12
0	1	1	1	0	0	0	0	0	0	0	0	0	0
1	1	2	3	2	1	0	0	0	0	0	0	0	0
2	1	3	6	7	6	3	1	0	0	0	0	0	0
3	0	2	7	14	18	16	10	4	1	0	0	0	0
4	0	1	6	18	34	45	43	30	15	5	1	0	0
5	0	0	3	16	45	84	113	114	87	50	21	6	1
6	0	0	1	10	43	113	208	285	300	246	157	77	
7	0	0	0	4	30	114	285	518	720	786	683		
8	0	0	0	1	15	87	300	720					
9	0	0	0	0	5	50	246	786					
10	0	0	0	0	1	21	157	683					

(5)

Because the roles of Type A and Type B are interchangeable, the matrix (5) is symmetric. *Challenge*: Can you find a recurrence that generates it? (If you'd like, stop and do some trial-and-error before reading further. It may take a while.) My own attempt had left me stumped.

But the events of December 1992 (my "odd" and "even" tests) brought the problem back to my attention, and I took another look at Table (5). Let's now examine it together. The rows for $m = 0, 1, 2$ are familiar: They are the sequences of coefficients in the expansion of $(1 + x + x^2)^n$ (the trinomial coefficients) for $n = 1, 2, 3$. In particular, each entry in rows $m = 1$ and $m = 2$ of Table (5) (say in column k) is the sum of the three entries from the preceding row in columns $k, k-1, k-2$. However, for $m \geq 3$ this pattern fails; in fact, the rows no longer have left-right symmetry. So let's ask: *By how much* does the pattern fail? Let $D_3(m, k)$ be the answer to that question; that is,

$$D_3(m, k) = \sum_{j=0}^{2} C_3(m-1, k-j) - C_3(m, k). \tag{6}$$

Let's tabulate D_3:

VALUES OF $D_3(m, k)$

$m \backslash k$	0	1	2	3	4	5	6	7	8	9	10
0	−1	−1	−1	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0
3	1	2	3	2	1	0	0	0	0	0	0
4	0	1	3	5	5	3	1	0	0	0	0
5	0	1	4	9	13	13	9	4	1	0	0
6	0	0	2	9	21	32	34	26	14	5	1
7	0	0	1	7	24	52	79	88	73	45	20

This time, the rows have left-right symmetry all the way through $m = 5$, with nonsymmetry beginning at $m = 6$, whereas in (5) the nonsymmetry began at $m = 3$. This suggests comparing row 6 of the D_3 matrix with row 3 of the C_3 matrix:

k	0	1	2	3	4	5	6	7	8	9	10
$C_3(3,k)$	0	2	7	14	18	16	10	4	1	0	0
$D_3(6,k)$	0	0	2	9	21	32	34	26	14	5	1

Aha! Do you see the pattern? If not, let's make it easier by shifting the D_3 row $1\frac{1}{2}$ spaces to the left:

$$\begin{array}{ll} (C_3 \text{ row}) & 0\ 2\ 7\ \ 14\ \ 18\ \ 16\ \ 10\ \ 4\ 1\ 0 \\ (D_3 \text{ row}) & 2\ 9\ 21\ \ 32\ \ 34\ \ 26\ \ 14\ 5\ 1 \end{array}$$

and now the scheme is as evident as in Pascal's triangle: $D_3(6,k) = C_3(3,k-1) + C_3(3,k-2)$. A check of other such pairs of rows (row m of D_3 versus row $m-3$ of C_3) reveals a similar pattern:

$$D_3(m,k) = C_3(m-3,k-1) + C_3(m-3,k-2) \tag{7}$$

with *two exceptions*: when $0 \le k \le 2$ and $m = 0$ or 3, the left side of (7) minus the right side equals -1 or $+1$, respectively. Thus, the correct recurrence for C_3 (in view of (6)) appears to be

$$\begin{aligned} C_3(m,k) &= \sum_{i=0}^{2} C_3(m-1,k-i) - \sum_{i=1}^{2} C_3(m-3,k-i) + e(m,k) \\ e(m,k) &= \begin{Bmatrix} 1, & \text{if } m = 0 \text{ and } 0 \le k \le 2 \\ -1, & \text{if } m = 3 \text{ and } 0 \le k \le 2 \\ 0, & \text{in all other cases} \end{Bmatrix}. \end{aligned} \tag{8}$$

Generalizing (8) to arbitrary values of t in place of $t = 3$, a natural guess was that, for all positive integers t and all integers m, k,

$$\begin{aligned} C_t(m,k) &= \sum_{i=0}^{t-1} C_t(m-1,k-i) - \sum_{i=1}^{t-1} C_t(m-t,k-i) + e_t(m,k) \\ e_t(m,k) &= \begin{Bmatrix} 1, & \text{if } m = 0 \text{ and } 0 \le k < t \\ -1, & \text{if } m = t \text{ and } 0 \le k < t \\ 0, & \text{in all other cases} \end{Bmatrix}, \end{aligned} \tag{9}$$

and numerical data (brute-force enumeration again) seemed to confirm it. At this point, I found proof easier than discovery and soon had an elementary combinatorial proof of (9). The proof is available to the reader on request, but will not be given here; instead, we shall exhibit in §5 a simpler such proof of the equivalent Jackson recurrence. At any rate, next on my agenda was to use (9) to obtain numerical results, including answers to the specific questions (1) and (2) posed in §1.

4. Numerical results For a random sequence of m objects of one type and k of another, the probability that a t-clump occurs is clearly

$$P_t(m,k) = 1 - C_t(m,k) \Big/ \binom{m+k}{k}.$$

Using (9), a program to compute $C_t(m, k)$, and hence $P_t(m, k)$, was written, tested, and run for various values of the parameters. One result was

$$P_6(26,26) = .46424\ldots,$$

agreeing with (4) (at which the programmer felt great relief). Two other results were

$$P_5(26,26) = .77307\ldots; \qquad P_4(26,26) = .97396\ldots,$$

so that the phrase "almost always" in assertion (0) still seems exaggerated even with $t = 5$ in place of Gardner's $t = 6$. (For $t = 4$, the phrase seems more appropriate.)

As for my "odd" and "even" tests in December '92: The probability of no 3-clump on the class test was

$$C_3(16,15)\Big/\binom{31}{15} = .0042342\ldots, \tag{10}$$

and the probability of no 3-clump at the final exam was

$$C_3(15,17)\Big/\binom{32}{15} = .0028779\ldots. \tag{11}$$

If the two distributions were independent,* we would conclude that the probability of a 3-clump occurring on *neither* exam was the product of the numbers (10) and (11), namely

$$.00001218\ldots,$$

less than 1 out of 82,000. (And yet it happened. A nonmathematical friend to whom I reported the event—and the odds against its occurrence—reacted thus: "So I *could* win the lottery!")

Postscript. It is easy to find the expected *number* $E = E(m, k, t)$ of noncontiguous t-clumps in a sequence of m 1's and k 0's. Indeed, the probability that the i-th term of the sequence *begins* such a run is $([m]_t + [k]_t)/[m+k]_t$ if $i = 1$ and is $(k[m]_t + m[k]_t)/[m+k]_t(m+k-t)$ if $1 \le i-1 \le m+k-t$, where $[x]_n$ denotes the product $x(x-1)\ldots(x-(n-1))$. (Compare with the discussion following display (3).) Summing over all i, since expectation is additive, we get

$$E(m,k,t) = \{(k+1)[m]_t + (m+1)[k]_t\}/[m+k]_t \qquad (1 \le t \le m+k).$$

In particular, $E(26,26,6) = .610\ldots$. Since this number clearly must exceed the probability of at least *one* 6-clump in the shuffled deck, we don't even need the actual probability to see that statement (0) is much too strong. Similarly, with respect to our question (2), $E(16,15,3) + E(15,17,3) = 7.55\ldots$, so that the extraordinariness of "no clumps" seems evident even before we have found the recurrence (8) or (9).

For n (fair) coin tosses, the formula for the expected number of t-clumps (obtained similarly) is even simpler:

$$E(n,t) = (n+2-t)/2^t. \tag{12}$$

*Approximate independence, at least, seems likely to me; I had instructed "friends" to sit apart, and the two exams were held in different rooms with different seat layouts.

It has been shown (e.g., in [4]) that if L_n is the length of the *longest* run in a sequence of n tosses, then $E(L_n) \sim \log_2 n$ as $n \to \infty$. (12) makes the latter intuitively plausible. For example, if $n = 2^{1000}$ ($\log_2 n = 1000$), then $E(n, 997)$ is extremely close to 8 and $E(n, 1003)$ is extremely close to $1/8$, making it seem very likely that $997 \le L_n \le 1002$. Viewed in this light, it is not surprising that the variance of L_n is nearly constant when n is large, a fact that Schilling's award-winning article [7] calls "remarkable" (as indeed it seems when first encountered). [7] and [4] give more precise expressions for $E(L_n)$ and $\mathrm{Var}(L_n)$.

5. A more efficient recurrence The number of terms on the right side of (9) increases with t. In [6], Jackson gave a partial proof, using the theory of combinatorial generating functions (as developed, e.g., in [5]) of the following alternate recurrence for $C_t(m, k)$, in which the right side has only six terms no matter how large t is:

$$C_t(m,k) = \begin{cases} C_t(m-1,k) + C_t(m,k-1) - C_t(m-t,k-1) - C_t(m-1,k-t) \\ \qquad + C_t(m-t,k-t) + e_t^*(m,k) \end{cases}$$

$$e_t^*(m,k) = \left\{ \begin{array}{rl} 1, & \text{if } (m,k) = (0,0) \text{ or } (t,t) \\ -1, & \text{if } (m,k) = (0,t) \text{ or } (t,0) \\ 0, & \text{otherwise} \end{array} \right\}. \tag{13}$$

A referee of this article has pointed out that (13) can in fact be derived without generating functions, as follows:

For fixed t, we call a sequence of 1's and 0's *good* if it contains no t-clump. A good sequence of m ones and k zeros will be denoted by $S(m, k)$; an $S(m, k)$ beginning with the digit i ($= 0$ or 1) will be denoted by $S_i(m, k)$; and $x^{(t)}$ (where $x = 0$ or 1) will denote the sequence $(x, x, \ldots, x)$ (t terms). Also, let $[A, B]$ denote the sequence consisting of the sequence A followed by the sequence B.

By inspection, (13) holds if $(m, k) = (0, 0)$ or $(0, t)$ or $(t, 0)$, so we assume (m, k) is not one of those three pairs. Since $(m, k) \neq (0, 0)$, every $S(m, k)$ has the form

$$[1, S(m-1,k)] \quad \text{or} \quad [0, S(m,k-1)]. \tag{14}$$

Conversely, since $(m, k) \neq (0, t)$ or $(t, 0)$, a sequence (14) is an $S(m, k)$ if and only if it is not of the form

$$\left[1^{(t)}, 0, S(m-t,k-1)\right] \quad \text{or} \quad \left[0^{(t)}, 1, S(m-1,k-t)\right]. \tag{15}$$

Next, if $(m, k) \neq (t, t)$ then a sequence (15) also has the form (14) if and only if it is not of the form

$$\left[1^{(t)}, 0^{(t)}, S_1(m-t,k-t)\right] \quad \text{or} \quad \left[0^{(t)}, 1^{(t)}, S_0(m-t,k-t)\right]; \tag{16}$$

the excluded sequences (16) all have the form (15); and their number is $C_t(m-t, k-t)$. If instead $(m, k) = (t, t)$, there are exactly two excluded sequences, namely $[1^{(t)}, 0^{(t)}]$ and $[0^{(t)}, 1^{(t)}]$, and $2 = C_t(0, 0) + 1$. In either case, the number of sequences (15) not of the form (14) is $C_t(m-t, k-t) + e_t^*(m, k)$, so that

$$\begin{aligned} \text{no. of } S(m,k)\text{'s} &= (\text{no. of sequences } (14)) - (\text{no. of sequences } (15)) \\ &\quad + C_t(m-t,k-t) + e_t^*(m,k), \end{aligned}$$

which is precisely (13).

6. Some exercises We asserted (§1) that the recurrences (9) and (13) are equivalent; a nice exercise for the student is to prove this assertion algebraically, without reference to combinatorics. Here are two more such exercises:

I. For fixed t, if $d_n = \Sigma_k C_t(n-k, k)$ (the n-th northeast-to-southwest diagonal sum in the matrix C_t), then

$$d_n = 2^n \ (0 \le n < t); \quad d_n = \sum_{j=1}^{t-1} d_{n-j} \ (n \ge t). \tag{17}$$

For example, when $t = 3$ then $d_n = d_{n-1} + d_{n-2}$ ($n \ge 3$), and in fact the d's are twice the Fibonacci numbers: $d_n = 2F_{n+1}$ when $n \ge 1$ (see table (5)). (17) can be proved either combinatorially or by induction.

II. For fixed t, if $r_m = \Sigma_k C_t(m, k)$ (the m-th row sum in C_t), then

$$r_m = t^{m+1} \ (0 \le m < t); \quad r_m = (t-1) \cdot \sum_{j=1}^{t-1} r_{m-j} \ (m \ge t). \tag{18}$$

This can be proved by induction using (9); I haven't found a combinatorial argument.

Either (17) or (18) can be used to check the matrix C_t, after constructing C_t from (9) or (13).

7. Related problems Space does not permit a comprehensive listing here of the literature on clump-related problems, but a few quite recent references (called to my attention by a knowledgeable referee) deserve brief mention. Godbole [**2**] obtains an explicit formula (as a sum) for the probability that, in the first n terms of a sequence of m 1's and k 0's, no run of t consecutive 1's occurs. (I know of no such explicit formula for "consecutive 1's *or* consecutive 0's," which was the problem addressed herein.) Sequences whose successive terms are independent (i.e., no parameters m, k) are easier to deal with, and several articles have done so in considerable generality. In particular, two papers in [**3**] treat "random n-letter words formed from an r-letter alphabet" ($r \ge 2$): Suman [**3**, 119–130] obtains three formulas (involving sums) for the probability that no t-clump occurs in such a "word", and Chryssaphinou, et al. [**3**, 231–241] study the waiting time until at least one of a given set of patterns occurs ("at least one t-clump" would be a special case). For readers wishing to pursue such matters further, the aforementioned articles also contain useful bibliographies; in addition, [**3**] contains recent articles on other clump-related topics.

Editor's Note. After this paper was accepted, it was pointed out that a recursion for the probability of clumps has been obtained by E. F. Schuster in [**3**], pp. 91–111. His recurrence is more complicated in that the terms of his recurrence must themselves be obtained from a different recurrence. Schuster presents a table of the probabilities that no t-clump occurs in a sequence of m 1's and n 0's, up to $m + n = 50$. Just short of what's needed for a deck of cards!

REFERENCES

1. Martin Gardner, *aha! Gotcha*, Freeman, New York, NY, 1982.
2. A. P. Godbole, On hypergeometric and related distributions of order k, *Commun. Statist.: Theory and Methods* 19 (1990), 1291–1301.
3. A. P. Godbole and S. Papastavridis (eds.), *Runs and Patterns in Probability: Selected Papers*, Kluwer Academic, Boston, MA, 1994.
4. L. Gordon, M. F. Schilling, and M. S. Waterman, An extreme value theory for long head runs, *Probability Theory and Related Fields* 72 (1986), 279–287.
5. I. P. Goulden and D. M. Jackson, *Combinatorial Enumeration*, Wiley-Interscience, New York, NY, 1983.
6. D. M. Jackson, personal communication, 3/12/93.
7. M. F. Schilling, The longest run of heads, *Coll. Math. J.* 21 (1990), 196–207.