THE PROBABILITY DISTRIBUTION OF NEAR-MATCHES OF BALLS IN AN ORDERED SEQUENCE OF CELLS

Steven J. Schwager

Biometrics Unit, Cornell University, Ithaca, NY

BU-1028-M May, 1989

Abstract

A set of n balls are distributed independently at random into r ordered cells, numbered 1 to r. Fix a nonnegative integer threshold m; a near-match is said to occur when two balls are assigned to cells whose distance is m or less. The probability distribution of the number of near-matches is derived. This model arises in connection with experiments in genetic mapping of ripening-related cDNA clones in tomatoes. Implications of the results for genetic data are discussed, using a particular experiment for which m=5, n=38, and r=1200. For m=0, this situation reduces to a classical occupancy problem, finding the distribution of the number of empty cells. For m>0, this problem is a discrete analogue to a problem addressed for continuous variables by the theory of spacings. Computational formulas for probabilities developed here do not suffer from the extreme numerical instability shown to be inherent in the classical formulas for both the empty cells problem and the continuous-variable spacings problem. However, computational complexity increases with the number of near-matches.

1. Introduction

The well-known class of occupancy problems is based on the distribution of n balls into r cells. In the situation of interest here, which is known as Maxwell-Boltzmann statistics, both the balls and the urns are distinguishable. The number of possible outcomes, or arrangements of the balls in the urns, in this formulation, is r^n. The balls may be distributed independently and at random, so that each ball has probability 1/r of going into any of the r cells. In this case, all of the possible outcomes are equally likely. Probabilities of events involving the numbers of cells containing given numbers of balls are often of interest. For example, we may wish to obtain the probability that the number of empty cells is j for various values of j, or the probability that at least one cell contains two or more balls. When r=365, the latter is the solution to the simplest version of the birthday problem, which is to find the probability of at least one match among the birthdays of n people. For a detailed treatment of occupancy problems, see Johnson and Kotz (1977).

In the occupancy problem, there is no geometric structure among the cells. In the usual formulation of the birthday problem, for instance, birthdays occurring on May 5 and May 7 have the same relation as birthdays on May 5 and November 7, namely, they are not a match; the closeness of the cells is not relevant. However, there are many applied situations in which the cells do have a geometric structure and the distance between the occupied cells matters. In the birthday problem, we may wish to find the probability of at least one near-match among the birthdays of n people, where a near-match is defined to mean birthdays occurring within three days of one another. This is the problem addressed in this paper. Its general formulation is now given.

A set of n distinguishable balls are distributed independently and at random into r cells numbered from 1 to r, forming an ordered sequence. Fix a nonnegative integer threshold m; a near-match is said to occur when two balls are assigned to cells whose distance is m or less. We seek the probability distribution, or equivalently the probability density function (p.d.f.), of the number of near-matches. For m=0, this situation reduces to the classical occupancy

problem.

For m>0, the situation is a discrete analogue to a problem that the theory of spacings addresses for continuous variables. If n points are independently and uniformly distributed on the unit interval (0,1), the probability distributions of various functions of the n+1 resulting subintervals will be of interest. A slight modification is to replace the unit interval by the unit circle, which the n points will divide into n subintervals. The problem for these continuous situations that corresponds to the discrete model's p.d.f. of the number of near-matches is the p.d.f. of the number of subintervals smaller than a given size. This problem has been studied by several authors, whose work is described by Solomon (1978). The version involving the unit circle can be restated as the following problem of coverage on the circle: n arcs, each of length m/r, are thrown at random onto a circle whose circumference is 1; find the probability that there are exactly n-i gaps left uncovered between these arcs, where i=0,1,...,n-1. There are i near-matches whenever there are n-i gaps, e.g., n gaps correspond to no near-matches, n-1 gaps to 1 near-match,..., and 1 gap to n-1 near-matches.

One practical application in which the number of near-matches arises is experiments in genetic mapping. The discussion here will focus on the mapping of ripening-related cDNA clones in tomatoes. This mapping results in the detection of one or more loci on the tomato genome for each clone. The genome may be viewed as a sequence of r ordered cells. It is of interest to know whether the loci occur independently at random locations on the genome or cluster on the genome. The value chosen for the threshold m could be based on the smallest resolvable distance of recombination, about 5 centimorgans (cm); or it could be 0, corrsponding to the occurrence of cosegregating pairs. Data from a genetic mapping experiment involving n=38 loci will be analyzed later in this paper.

Computational formulas for m=0 have long been known. These give the probability of having r-n+i empty cells, which corresponds to having i exact matches, for i=0,1,...,n-1. From Hoel, Port, and Stone (1971, p.45, equation (16)) or Feller (1968, p. 60, equation (11.7)), the probability that exactly k cells are empty is

$$(1) \qquad P[\text{exactly k cells empty}] = P[\text{k+n-r matches}] = \binom{r}{k}\sum_{j=0}^{r-k}(-1)^j\binom{r-k}{j}\left(1-\frac{j+k}{r}\right)^n.$$

Formulas for the arc coverage problem on the unit circle, a continuous analogue of the m>0 case, are also known. These give the probability of having i gaps in coverage among arcs on the unit circle. From Solomon (1978, p. 80, equation (14.4)), after defining g to be the greatest integer that is less than r/m,

$$(2) \qquad P[\text{exactly n-i gaps}] = P[\text{i near-matches}] = \binom{n}{i}\sum_{j=0}^{g-n+i}\binom{i}{j}(-1)^j\left[1-\frac{(j+n-i)m}{r}\right]^{n-1}.$$

Both formulas (1) and (2) suffer from extreme numerical instability. They involve sums whose terms grow increasingly large as we move through the summation. Many of these terms are several orders of magnitude greater than 1. In theory, the terms have alternating signs, so the positive and negative terms cancel almost exactly, leaving a probability as the overall series sum. In practice, however, problems arise when formulas (1) and (2) are used with parameter values suitable for gene mapping problems, e.g., n=20 to 50, r=1200 or 1500, and m=0 or 5. The presence of round-off errors, underflows, overflows, and similar numerical difficulties results in computed "probabilities" that can be much greater in magnitude than 1, and can also be negative in sign. Thus, these formulas are computationally unsatisfactory for obtaining the probabilities of given numbers of near-matches or matches.

Thus a more effective procedure for obtaining these probabilities is needed. Such a procedure is developed in Section 2. It is applied in Section 3 to an example using data from a gene mapping experiment.

2. Near-Match Probabilities

Assume that n balls are distributed independently at random among r cells. Because the r^n possible outcomes are equally likely under these conditions, classical equal likelihood

probability methods can be used to find event probabilities. We will derive formulas for counting the number of ways a given number of near-matches can occur. The analysis here applies to the case of m=0 as well as to m>0, so it applies to events involving matches as well as those involving near-matches.

Let the locations of the n balls be given by the vector $X = (x_1,x_2,...,x_n)$, and let $Y = (y_1,...,y_n)$ denote the vector of order statistics of X, arranged in increasing order, so $y_1 \leq y_2 \leq ... \leq y_n$. Define $d_i = y_{i+1} - y_i$ for i=1,2,...,n-1, so each $d_i \geq 0$, and let D denote the vector $(d_1,...,d_{n-1})$. Then each vector X can be transformed to Y and then to the pair (y_1,D); each Y corresponds to a single (y_1,D), and vice versa. Note that some Y's are obtained from more X's than others. To see this, observe that if all elements of Y are distinct, then n! different X's correspond to that Y, and thus to the pair (y_1,D). However, if two elements of Y, y_i and y_{i+1} for some i, are equal, then only n!/2! different X's correspond to that Y; and if all n elements of Y are equal, then only n!/n! = 1 vector X corresponds to that Y.

A near-match occurs for the i^{th} pair of locations, y_i and y_{i+1}, if $d_i \leq m$. Let A_k denote the event that exactly k of these near-matches occur. By convention, this specifies that the greatest possible number of near-matches is n-1 (rather than n(n-1)/2). For example, if 5 balls go into the same cell, the result will be counted as 4 matches, not as 10. This is a conservative approach, as it limits the influence of observing more than two balls in close proximity. If desired, it could be modified to increase the influence of such an occurrence.

It will be useful to divide the d_i's into two categories and to introduce additional notation. For a given X, let k be the number of near-matches that have occurred. Those k d_i's that are less than or equal to m, resulting in near-matches, will be denoted by $e_1,e_2,...,e_k$, where the e_i's are arranged in nondecreasing order, so $0 \leq e_1 \leq e_2 \leq ... \leq e_k \leq m$. Let E denote the vector $(e_1,...,e_k)$; it is implicit in this notation that the length of E represents the number of near-matches. Let $t = e_1 + ... + e_k$. Let $n_0, n_1,..., n_m$ denote the number of 0's, 1's,..., m's in E. Those n-1-k d_i's that are more than m, which do not give

near-matches, will be denoted by $f_1,f_2,\ldots,f_{n-1-k}$. Let $s = f_1 + \ldots + f_{n-1-k}$.

We now develop a counting argument to allow us to compute $\#(A_k)$, the number of vectors X that result in the occurrence of event A_k. This number is the sum of the numbers #(E), where E ranges over all possible vectors of length k with $0 \le e_1 \le \ldots \le e_k \le m$. For a given E, a range of values of s may occur. This range spans from a minimum of (n-1-k)(m+1), when each of the f_i's is just large enough to avoid a near-match, to a maximum of r-1-t, when the observed locations in X include 1 and r.

For given E and s, four aspects of the vector X must be considered:

1. The number of possible position sequences within D occupied by E. Of the n-1 positions in D, n_0 must contain 0's, n_1 must contain 1's,..., and n_m must contain m's, so the number of possible sequences within D occupied by E is given by the multinomial coefficient #(D,E) given by

$$\binom{n-1}{n_0,n_1,\ldots,n_m,n-1-k}.$$

If m=0, this reduces to 1.

2. The number of distinct sequences $f_1,\ldots,f_{n-1-k}$ whose sum equals s. Each of the f_i's must be at least m+1, so we need to know how many ways the excess, s - (n-1-k)(m+1), can be distributed among the n-1-k elements f_i. By an elementary combinatorial argument, the number of ways this can be done, #(F|s), is equal to

$$\binom{(n-2-k) + (s - (n-1-k)(m+1))}{n-2-k}.$$

3. The number of distinct values of y_1 compatible with the value of s. Since y_1 can range from 1 to r-s-t, this number is obviously r-s-t.

4. The number of orderings of X compatible with Y, or equivalently, with (y_1,D). If all of the e_i's are greater than 0, this number is n!. If some of the e_i's are 0, though, the corresponding elements of Y are equal, so the number of orderings must be reduced. Let

$a_1, a_2, \ldots, a_m$ be the number of times in D that a sequence of exactly 1,2,...,m consecutive 0's occurs. The number of orderings of X compatible with Y is then

$$\#(X,D) = n! \; / \; (2!)^{a_1}(3!)^{a_2}\ldots(\{m+1\}!)^{a_m}.$$

Combining these aspects of X and summing over all E and s compatible with A_k gives

$$\#(A_k) = \sum_E \sum_s \#(D,E) \times \#(F|s) \times (r\text{-}s\text{-}t) \times \#(X,D) \tag{3}$$

where summation over E ranges over all E with $0 \le e_1 \le e_2 \le \ldots \le e_k \le m$, and summation over s ranges from $s = (n\text{-}1\text{-}k)(m+1)$ to $s = r\text{-}1\text{-}t$. This computation is feasible for many combinations of r, n, m, and k of practical importance in genetic mapping and other applications. The calculation becomes more laborious as the value of k increases, because summing over E involves a k-fold sum. The probability of exactly k near-matches is then given by

$$P(A_k) = \#(A_k) \; / \; r^n \tag{4}$$

3. Application to Genetic Mapping

We now examine the practical application mentioned earlier, the genetic mapping of ripening-related cDNA clones in tomatoes. This mapping results in the detection of one or more loci on the tomato genome for each clone. We will treat the genome as a sequence of r ordered cells. The value of r has been estimated at 1200 cm (centimorgans), but more recent data indicate a genome size of r = 1500 cm, so we will use both values in our analysis. Two values were chosen for the threshold m: m = 5, based on the smallest resolvable distance of recombination; and m = 0, corresponding to the occurrence of cosegregating pairs. A genetic mapping experiment generated data involving n = 38 loci, the number of loci homologous to the ripening clones. The loci may occur independently, at random locations on the genome, or their locations on the genome may exhibit clustering. We would like to determine which

of these hypotheses is supported by the observed locations of the 38 loci. These observed locations included three cosegregating pairs, so $k = 3$ when we use the threshold $m = 0$. The locations included ten pairs of tightly linked (within 5 cm) loci, so $k = 10$ when we use the threshold $m = 5$.

Under the null hypothesis, the n balls (loci) are distributed independently at random among the r cells. Under the alternative hypothesis, that there is clustering among the locations of the n balls, the number of near-matches will be stochastically greater than it is under the null hypothesis. We therefore reject the null hypothesis if the observed k lies in the upper 5% tail of the distribution of k under the null hypothesis. We now obtain the probability of observing k near-matches under the null hypothesis, for the specified m and r. For k=0,1,...,4, this was obtained from a computer implementation of formula (4) above. For $k \geq 5$, the computer algorithm is still under development, so a simulation of the process was performed. A set X of 38 independent random integers was generated, each uniformly distributed between 1 and r, and the number k of near-matches with threshold m was determined for X; probabilities in the table here are based on 100,000 repetitions of the simulation. As a check on the accuracy of the simulation, it agreed with the exact values found from (4) to within .0015 or less for all values of $k \leq 4$, m, and r.

Probability of exactly k near-matches under the null hypothesis

k	m=0 r=1200	m=0 r=1500	m=5 r=1200	m=5 r=1500
0	.5532	.6233	.0008	.0039
1	.3343	.2995	.0074	.0253
2	.0939	.0669	.0299	.0772
3	.0163	.0092	.0760	.1476
4	.0020	.0009	.1360	.1989
5	.0003	.0002	.1829	.2000
6	.0000	.0000	.1890	.1591
7	.0000	.0000	.1617	.1006
8	.0000	.0000	.1085	.0524
9	.0000	.0000	.0616	.0221
10	.0000	.0000	.0299	.0074
11	.0000	.0000	.0114	.0022
12	.0000	.0000	.0039	.0006
13	.0000	.0000	.0013	.0002
14	.0000	.0000	.0004	.0000

The observed value k = 3 has a p-value of .0186 for r = 1200, .0103 for r = 1500. The observed value k = 10 has a p-value of .0469 for r = 1200, .0104 for r = 1500. From these results, we would reject the null hypothesis and conclude that significant evidence of clustering is present. (In the actual analysis of these data, it was suspected that 1 of the 3 matches (using m = 0) and 4 of the 10 near-matches (using m = 5) occurred within suspected areas of reduced recombination. This required an adjustment in the observed values to k = 2 (for m = 0) and k = 6 (for m = 5), and resulted in acceptance of the null hypothesis, reversing the earlier conclusion.)

References

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications.* Volume I, 3rd Edition, Wiley, New York.

Hoel, P.G., S.C. Port, and C.J. Stone. (1971). *Introduction to Probability Theory.* Houghton Mifflin Company, Boston.

Johnson, N.L. and S. Kotz. (1977). *Urn Models and Their Application.* Wiley, New York.

Solomon, H. (1978). *Geometric Probability.* SIAM, Philadelphia.