BOUNDARIES OF MINIMUM SIZE IN BINOMIAL SAMPLING

By R. L. Plackett

University of Liverpool

1. Introduction. Much attention has recently been concentrated on the problems arising when sampling a binomial population, since this is thought to form a suitable model for certain industrial and biological procedures. A general discussion of such procedures as applied in industry has been given by Barnard [2] and various particular cases have received detailed treatment by Burman [3] Stockman and Armitage [6], and Anscombe [1]. Unbiased estimation of the population parameter (the "fraction defective") has been investigated by Girshick, Mosteller and Savage [4] and Wolfowitz [7]. A paper by Haldane [5] is also relevant.

For such sampling procedures it is necessary to find the probabilities of accepting or rejecting material with a particular fraction defective; to calculate the average sample size; and to form an estimate of the fraction defective when sampling terminates. All three characteristics may be expressed in terms of quantities $N(x, y)$, defined in section 3, so that once these are known, the fundamental properties of the scheme are known.

Here we present a method for determining the $N(x, y)$; investigate the conditions under which it is valid; relate the method to the estimation problem; and exemplify its application. The schemes to which the method can successfully be applied are of a special type (to which the title refers) and include all inspection procedures with a finite upper limit to the sample size likely to be used in practice. Other schemes, when dissected in a manner similar to that used by Stockman and Armitage, can doubtless be formulated as an aggregate of the special types.

2. Nomenclature. Our nomenclature differs in some respects from that of Girshick, Mosteller and Savage, although the same collection of terms is employed. References to their paper should therefore be followed by a comparison of the terminology.

Taking a sample of one from a binomial population consists in observing either of two events, whose probabilities are p and $1 - p$ ($p \neq 0$ or 1). The results of successive samples of one can be represented by the path of a particle in a two-dimensional lattice of points with non-negative integer co-ordinates. This particle starts at the origin 0 and at any point (x, y) travels to $(x + 1, y)$ if the event whose probability is p has occurred, otherwise to $(x, y + 1)$. Sampling terminates when the particle reaches a *boundary point*, and the set of such points is denoted by B. Any point which can be reached during sampling, including the boundary points, is *accessible*, and any path from the origin to a point B which can be traversed during sampling is *admissible*; all other points are *inaccessible* and all other paths *inadmissible*. The *index* of a point is the sum of its coordinates.

It will probably help to note in particular that whereas Girshick, Mosteller and Savage used p to correspond to events causing the y co-ordinate to increase, we use it for x.

3. Determination of $N(x, y)$. The set B determines the sampling scheme and we are concerned with schemes in which all points of index greater than n, the finite maximum index of points in B, are inaccessible. This condition guarantees that if $N(x, y)$ denotes the number of admissible paths from the origin to a point (x, y) of B

$$\sum_B N(x, y)p^x(1 - p)^y \equiv 1,$$

the summation being over all boundary points. Consequently, to determine $N(x, y)$ equate coefficients of p in this identity, the coefficient of p^0 in the left hand side being 1 and all others zero. When all the $N(x, y)$ are known, the probability of reaching any subset of B can be calculated and the characteristics of the scheme found.

Sometimes it will be convenient to use

$$\sum_B N(x, y)q^y(1 - q)^x \equiv 1,$$

where $q = 1 - p$, but the resulting set of equations cannot be independent of the first set since if

$$\sum_{i=0}^{n} a_i p^i \equiv \sum_{j=0}^{n} b_j(1 - p)^j,$$

then

$$a_i = \sum_{j=i}^{n} (-1)^i \binom{j}{i} b_j.$$

The polynomial in either p or q is of degree n; the application of this method alone is therefore limited to boundaries containing at most $(n + 1)$ points, otherwise the number of unknowns exceeds the number of equations for them.

4. Properties of the boundary.

THEOREM 1. *If n is the maximum index of points in B and if any point of greater index is inaccessible, then B contains at least $n + 1$ points.*

There must be at least two boundary points of index n for any such point (a_n, b_n) must be approached from $(a_n - 1, b_n)$ or $(a_n, b_n - 1)$; in which case either $(a_n - 1, b_n + 1)$ or $(a_n + 1, b_n - 1)$ is a boundary point. Let P be any one of these points. At least one admissible path exists from 0 to P; suppose one such path to consist of the points (a_0, b_0), (a_1, b_1), $\cdots$, (a_n, b_n) where $a_k + b_k = k$ $(k = 0, 1, 2, \cdots, n)$. It is clear that one or more boundary points exist on the line $x = a_k$, having $y > b_k$, for otherwise the particle could travel indefinitely along this line; similarly one or more exist on $y = b_k$ with $x > a_k$; and if there is just one on each they cannot be identical unless $k = n$ since (a_k, b_k) is not then a boundary point. Initially (a_0, b_0) contributes two boundary points; since then either $a_{k+1} = a_k$ and $b_{k+1} \neq b_k$ or $a_{k+1} \neq a_k$ and $b_{k+1} = b_k$ it follows that each succeeding point up to and including (a_{n-1}, b_{n-1}) contributes at least one more; the point (a_n, b_n) is counted as soon as x reaches a_n or y reaches b_n, whichever occurs first. Consequently there are at least $n + 1$ boundary points.

Reversely, if the boundary contains $n + 1$ points whose maximum index is m, such that any point of greater index is inaccessible, then $m \leq n$. For suppose $m > n$ and apply the preceding result.

An important class of boundaries therefore comprises those with the minimum number of points necessary to attain a given maximum index; they may conveniently be termed boundaries of minimum size and for them alone the method of equating coefficients yields the number of equations equal to the number of unknowns, the first being otherwise less than the second.

If there are exactly $n + 1$ boundary points then $(a_1, b_1), (a_2, b_2), \cdots, (a_{n-1}, b_{n-1})$ must each contribute to just one; since $a_{k+1} = a_k$ or $a_k + 1$ there is one point of B on each of the lines $x = 0, x = 1, \cdots, x = a_n$ and this set of points $(0, d_0)(1, d_1), \cdots, (a_n, b_n)$ can be denoted by U, the upper part of the boundary. Clearly $d_{k+1} \geq d_k - 1$ for otherwise more than one boundary point is required on the line $x = k + 1$. Similarly, there must be a second group of points of B $(c_0, 0), (c_1, 1), \cdots, (a_n, b_n)$ with $c_{k+1} \geq c_k - 1$ forming the lower boundary L; and all $(n + 1)$ points have now been enumerated, the point P belonging to both U and L. The characteristic of such sets B is that the sequences U and L both have monotonically non-decreasing index; the special case of sequences with monotonically increasing index provides the rejection and acceptance boundaries of non-rectifying industrial inspection procedures. (The difference between rectifying and non-rectifying procedures is clearly stated in the introduction to Anscombe [1]).

THEOREM 2. *For boundaries of minimum size any two accessible points not in B of the same index m cannot be separated on the line $x + y = m$ by boundary or inaccessible points* In the terminology of Girshick, Mosteller and Savage the accessible points not in B form a *simple* region.

Let $Q(x_1, y_1)$ and $R(x_2, y_2)$ be any two such accessible points of index m and suppose $x_1 < x_2$. There are two possibilities: (a_m, b_m) does or does not lie between Q and R.

(i) (a_m, b_m) lies between Q and R, i.e. $x_1 < a_m < x_2$. In this case there must be points of B at $Q'(x_1, Y_1)$ with $Y_1 > y_1$ and at $R'(X_2, y_2)$ with $X_2 > x_2$. The boundary from Q' to P and from R' to P has non-decreasing index; hence all points of U on the lines $x = x_1, x = x_1 + 1, \cdots, x = a_m - 1$ have index at least $x_1 + Y_1 > m$; similarly all points of L on the lines $y = y_2, y = y_2 + 1, \cdots, y = b_m - 1$ have index at least $X_2 + y_2 > m$. By definition of the boundary there are no additional points of B on either group of lines between the path $0P$ and the line $x + y = n$, so the proof of the theorem is completed.

(ii) If $x_1 \geq a_m$ or $x_2 \leq a_m$ the proof is precisely analogous to that given in (i).

5. Justification of the method. THEOREM 3. *For boundaries of minimum size the equations for $N(x, y)$ are soluble and of rank $n + 1$.*

To prove this we give a general method of solution for the system of equations, using powers of p and q alternately: as already remarked, this is equivalent to using the equations from the coefficients of powers of p only. In the first place, note that the coefficient of q^u is a linear combination of numbers $N(x,y)$ with $x + y \geq u$ and $y \leq u$; and the coefficient of p^t has $x + y \geq t$ and $x \leq t$.

$$\text{Let } s = \text{Min}(d_0, d_1, d_2, \cdots, b_n) - 1.$$

Then from the coefficients of $q^0, q^1, \cdots, q^s$ can successively be determined $N(c_0, 0)\, N(c_1, 1), \cdots, N(c_s, s)$, the matrix of the equations being triangular with ones in the main diagonal. The points in U at $(r_1, s + 1), (r_2, s + 1), \cdots$ now appear in the coefficients of $q^{s+1}, q^{s+2}, \cdots$ and complicate the solution.

$$\text{Let } r = \text{Max}(r_1, r_2, \cdots).$$

If either (r, d_r) or (c_s, s) is the point P then all the remaining $N(x, y)$ can successively be determined from the coefficients of powers of p when the values of $N(c_0, 0), N(c_1, 1), \cdots, N(c_s, s)$ are substituted in the equations. Otherwise the path OP for $y \geq s + 1$ must have $x \geq r + 1$ so that all points of L on $y \geq s + 1$ have $x \geq r + 2$ i.e. any point of L on $x = 0, x = 1, \cdots, x = r$ has $y \leq s$; for such points the number of admissible paths is now known. Therefore from the coefficients of $p^0, p^1, \cdots, p^r$ can successively be determined $N(0, d_0)$, $N(1, d_1), \cdots, N(r, d_r)$, the matrix of these unknowns being again triangular; in particular $N(r_1, s + 1), N(r_2, s + 1), \cdots$ can now be found.

Let $s_1 = \text{Min}\ (d_{r+1}, d_{r+2}, \cdots, b_n) - 1$, so that $s_1 > s$. The coefficients of $q^{s+1}, q^{s+2}, \cdots, q^{s_1}$ give successively $N(c_{s+1}, s + 1)\, N(c_{s+2}, s + 2), \cdots, N(c_{s1}, s_1)$; for the points in U at $(r_{11}, s_1 + 1), (r_{12}, s_1 + 1) \cdots$. Let

$$r_1 = \text{Max}\ (r_{11}, r_{12}, \cdots).$$

Since there is only one point of U on each line $x = \text{constant}$, $r_1 > r$. As before, if either (r_1, d_{r_1}) or (c_{s_1}, s_1) is P the remaining points of U are soon determined. Otherwise the process continues and there result an increasing sequence of points of L and a similar sequence for U; the process terminates when (a_n, b_n) has been reached in both, when all $N(x, y)$ will have been found.

It is clear that for particular cases alternative methods of solution will prove more convenient.

6. Connection with estimation. Suppose that the point (t, u) is accessible and let $N^*(x, y)$ be the number of admissible paths from (t, u) to (x, y) where (x, y) is in B. Then Girshick, Mosteller and Savage have shown that $N^*(x, y)/N(x, y)$ is an unbiased estimate of $p^t(1 - p)^u$; and a necessary and sufficient condition for it to be the unique unbiased estimate is that the accessible points not in B form a simple finite region. Hence from theorem 2 such estimates are unique for schemes with boundaries of minimum size. An alternative proof is given by considering that if two unbiased estimates of any function of p exist and $f(x, y)$ is the difference between them at (x, y)

$$\sum_B f(x, y)N(x, y)p^x(1 - p)^y \equiv 0,$$

where $f(x, y)$ is not everywhere zero. The equations formed by equating coefficients have rank $(n + 1)$ as shown by Theorem 3, so that the only solution is $f(x, y)N(x, y) = 0$. Since each $N(x, y)$ is certainly positive it follows at once that $f(x, y) = 0$ and there can only be one unbiased estimate.

7. An illustration. As an application of the method we take the interesting rectifying sequential inspection scheme discussed by Anscombe. The boundary points are at $(H, 0)$, $(H + b, 1)$, $\cdots$ $(H + \mu b, \mu)$, where μ is the greatest integer less than $(N - H)/(b + 1)$, and thereafter on the line $x + y = N$. The equations for $N(x, y)$ take here their simplest form, namely equation (4) of Barnard's paper. From the coefficients of $q^0, q^1, \cdots, q^y, \cdots,$

$$1 = N(H, 0);$$

$$0 = N(H + b, 1) - HN(H, 0) \quad \text{whence} \quad N(H + b, 1) = H;$$

$$0 = N(H + 2b, 2) - \binom{H + b}{1}H + \binom{H}{2} \quad \text{whence} \quad N(H + 2b, 2)$$

$$= \frac{H(H + 2b + 1}{2!};$$

$$0 = N(H + 3b, 3) - \binom{H + 2b}{1}\frac{H(H + 2b + 1)}{2!} - \binom{H + b}{2}H + \binom{H}{3};$$

$$\text{whence } N(H + 3b, 3) = \frac{H(H + 3b + 2)(H + 3b + 1)}{3!}.$$

It now appears reasonable to guess the general term as

$$\frac{H}{y!}(H + yb + y - 1)(H + yb + y - 2) \cdots (H + yb + 1).$$

The proof is therefore complete if we show

$$\binom{H}{y} - \binom{H + b}{y - 1}H + \binom{H + 2b}{y - 2}\frac{H(H + 2b + 1)}{2!}$$

$$- \binom{H + 3b}{y - 3}\frac{H(H + 3b + 2)(H + 3b + 1)}{3!}$$

$$+ \cdots + (-1)^y \frac{H(H + yb + y - 1)(H + yb + y - 2) \cdots (H + yb + 1)}{y!} = 0.$$

Put $(b + 1) = \xi$, and the left hand side becomes

$$\frac{(H-1)!}{(H-y)!y!} - \frac{(H+\xi-1)!}{(H+\xi-y)!(y-1)!1!} + \frac{(H+2\xi-1)!}{(H+2\xi-y)!(y-2)!2!} - \cdots (-1)^y \frac{(H+y\xi-1)!}{(H+y\xi-y)!y!},$$

which is y times the coefficient of t^{H-y} in $(1+t)^{H+y\xi-1} \times [(1+t)^{-\xi} - t^{-\xi}]^y$. Rewriting the latter as $(1+t)^{H-1}[1 - (1+t^{-1})^{\xi}]^y$, it becomes clear that the highest power of t is t^{H-y-1}, whence the required result follows.

REFERENCES

[1] F. J. Anscombe, "Linear sequential rectifying inspection for controlling fraction defective," *Roy. Stat. Soc. Jour.* (*supplement*), Vol. 8 (1946), pp. 216–222.

[2] G. A. Barnard, "Sequential tests in industrial statistics," *Roy. Stat. Soc. Jour.* (*supplement*), Vol. 8 (1946), pp. 1–21.

[3] J. P. Burman, "Sequential sampling formulas for a binomial population," *Roy. Stat. Soc. Jour.* (*supplement*), Vol. 8 (1946), pp. 98–103.

[4] M. A. Girshick, F. Mosteller, and L. J. Savage, "Unbiased estimates for certain binomial sampling problems with applications," *Annals of Math. Stat.*, Vol. 17 (1946), pp. 13–23.

[5] J. B. S. Haldane, "On a method of estimating frequencies," *Biometrika*, Vol. 33 (1945), pp. 222–225.

[6] C. M. Stockman and P. Armitage, "Some properties of closed sequential schemes," *Roy. Stat. Soc. Jour.* (*supplement*), Vol. 8 (1946), pp. 104–112.

[7] J. Wolfowitz, "On sequential binomial estimation," *Annals of Math. Stat.*, Vol. 17 (1946), pp. 489–492.