Improved Estimation of Pathogen Transmission Rates by Group Testing

Peter M. Burrows

Experimental Statistics Unit, Clemson University, Clemson, SC 29634-0367.
Technical Contribution 2466 from the South Carolina Agricultural Experiment Station, Clemson University.
Accepted for publication 19 May 1986.

ABSTRACT

Burrows, P. M. 1987. Improved estimation of pathogen transmission rates by group testing. Phytopathology 77:363-365.

Estimation of infection rates or probabilities of disease transmission is improved by adopting an alternative to maximum likelihood estimation with superior bias and mean square error properties. This improves the efficiency of group testing and extends the range of conditions where group testing is more efficient than individual testing. A simple formulation of optimal group size is presented for situations where the number of test plants is fixed by resource limitations.

Additional key words: multiple transfer designs, pathogen transmission, vectors.

Swallow (3) recently discussed the merits of group tests (multiple vector transfers) when estimating individual pathogen transmission rates. A single group test consists of transferring k vectors to each of N noninfected test plants, observing the subsequent count R of healthy plants, and calculating an estimate of transmission rate. There may be several, even many, such tests in any one experiment conducted to investigate variation in transmission rates associated with different sources of pathogen acquisition and with different test plant and vector genotypes. If p denotes the transmission rate per individual vector in a single group test, its maximum likelihood estimate is given by

$$\hat{p} = 1 - [R/N]^{1/k}$$

when it can be assumed that the vectors behave independently (even though transferred in groups) and that the N test plants are equally susceptible and respond independently. Swallow's discussion is based on the bias and mean square error properties of $\hat{p}$ in relation to design combinations (N, k) and the unknown p.

An alternative estimate, $\tilde{p}$, almost as simple to calculate as $\hat{p}$, is developed in the next section where it is shown that bias and mean square error properties of

$$\tilde{p} = 1 - [(2kR + k - 1)/(2kN + k - 1)]^{1/k}$$

are uniformly superior to those of $\hat{p}$ except for $k = 1$ when both estimates are identical to the minimum variance unbiased estimate. This does not modify the conclusion that group tests ($k > 1$) are usually preferable to individual testing ($k = 1$), but use of $\tilde{p}$ rather than $\hat{p}$ does improve the efficiency of group testing and changes the optimal choices of k in relation to p when N is fixed.

COMPARISON OF $\hat{p}$ AND $\tilde{p}$

Let $\theta = (1-p)^k$, the expected proportion of healthy plants under the assumptions stated in the previous section; then R follows a binomial distribution with rate parameter θ and sample size N. Accurate and efficient estimation of p requires an estimator of $\theta^{1/k}$ with favorable bias and mean square error properties. One approach to estimation of nonlinear functions of θ, given R and N, is found in the work of Haldane (2) and Anscombe (1) leading to the familiar estimate $\log[(R + 0.5)/(N - R + 0.5)]$ for logit $(\theta) = \log[\theta/(1 - \theta)]$. Application of that approach to the present problem begins with $[(R+a)/(N+b)]^{1/k}$ instead of the maximum likelihood estimate $(R/N)^{1/k}$, and then a and b are chosen so as to eliminate the dominant term of the bias when expanded as a power series in $(N\theta)^{-1}$. The result is

$$\tilde{p} = 1 - [(R + a)/(N + b)]^{1/k},\ a = b = (k - 1)/2k.$$

Bias $(\tilde{p})$, defined as the expectation of $(\tilde{p} - p)$, and mean square error of $\tilde{p}$, denoted by MSE$(\tilde{p})$ and defined as Variance $(\tilde{p})$ + [Bias $(\tilde{p})]^2$, can be calculated using the method described by Swallow (3) for calculation of Bias$(\hat{p})$ and MSE$(\hat{p})$. There are 154 combinations of N (= 10, 15, 20, 25, 30, 40, 50, 60, 80, 100, 200) and p (= 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.08, 0.10, 0.15, 0.20, 0.25, 0.30, 0.40, 0.50) in Table 1 of Swallow (3), which contains values $k = k^*$ yielding minimum mean square error of $\hat{p}$ provided that $k^* \leq 50$. Values of Bias$(\tilde{p})$, Bias$(\hat{p})$, Variance$(\tilde{p})$ and Variance$(\hat{p})$ have been calculated for each $k = 1,2,3,\ldots, 50$ in all of these 154 combinations of N and p. The following comparisons of bias and mean square error properties of $\tilde{p}$ and $\hat{p}$ are based on these calculations and some approximate theoretical results.

Bias$(\tilde{p})$ is found to be uniformly less than Bias$(\hat{p})$ for all $k = 2,3,4,\ldots,50$. When attention is restricted to $k = k^* > 1$ (153 combinations), Bias$(\tilde{p})$ attains a maximum of 5.2% of Bias$(\hat{p})$ when $N = 10$ and $p = 0.4$, exceeds 3% of Bias$(\hat{p})$ only when $N = 10 (p > 0.1)$, $N = 15 (p > 0.25)$ and $N = 20 (p > 0.4)$, and is less than 1% of Bias$(\hat{p})$ for 72 of the combinations. Provided that $N\theta > 1$, an approximate comparison of Bias$(\tilde{p})$ and Bias$(\hat{p})$ is available from the following series expansions for expected values of $\tilde{p}$ and $\hat{p}$:

$$E(\hat{p}) = p + b(1-2b)(1-\theta)(1-p)\left(\frac{1}{(N\theta)} + \frac{(B_2 + B'_2)}{(N\theta)^2} + \text{terms of order } (N\theta)^{-3}\right)$$

$$E(\tilde{p}) = p + b(1-2b)(1-\theta)(1-p)\left(\frac{B_2}{(N\theta)^2} + \text{terms of order } (N\theta)^{-3}\right)$$

where the coefficients B_2 and B'_2 are given by

$$B_2 = (1 + \theta)(1 - b)/6 \text{ and } B'_2 = (1 - \theta)b^2 + b.$$

Observe that there is no term with denominator $N\theta$ in the expression for Bias$(\tilde{p}) = E(\tilde{p}) - p$; estimate $\tilde{p}$ was constructed to eliminate this term, which accounts for the dramatic reduction in bias when compared with $\hat{p}$ in the previous paragraph.

For the same 154 combinations, Variance$(\tilde{p})$ is found to be uniformly less than Variance$(\hat{p})$ at all $k = 2,3,4,\ldots,50$; thus

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. § 1734 solely to indicate this fact.

reduction of bias has *not* sacrificed precision. Because both bias and variance of $\tilde{p}$ are superior to these properties of $\hat{p}$, MSE($\tilde{p}$) is uniformly less than MSE($\hat{p}$) for all $k = 2,3,4,\ldots,50$. When attention is restricted to $k = k^* > 1$, MSE($\tilde{p}$) is less than 80% of MSE($\hat{p}$) for eight combinations (occurring in columns $N = 10$ and $N = 15$), is 80–89.8% of MSE($\hat{p}$) for 47 combinations (all with $N < 50$), is 90–94.6% of MSE($\hat{p}$) for 55 combinations, and greater than 95% of MSE($\hat{p}$) for the remaining 43 combinations.

Expressions enabling approximate comparisons of MSE($\tilde{p}$) and MSE($\hat{p}$) provided that $N\theta > 1$, are given next:

$$\text{MSE}(\hat{p}) = \frac{(1-\theta)(1-p)^2}{k^2}\left(\frac{1}{(N\theta)} + \frac{(M_2 + M'_2)}{(N\theta)^2} + \text{terms of order } (N\theta)^{-3}\right)$$

$$\text{MSE}(\tilde{p}) = \frac{(1-\theta)(1-p)^2}{k^2}\left(\frac{1}{(N\theta)} + \frac{M_2}{(N\theta)^2} + \text{terms of order } (N\theta)^{-3}\right)$$

where the coefficients M_2 and M'_2 are given by

$$M_2 = 2(1-\theta)b^2 \text{ and } M'_2 = 5(1-\theta)b^2 + 2\theta b.$$

Unfortunately, these approximate comparisons are not very accurate for those N, p, and k combinations yielding large differences between MSE($\hat{p}$) and MSE($\tilde{p}$) because these series expansions are slow to converge unless $N\theta >> 1$.

OPTIMAL GROUP SIZES WHEN N IS FIXED

All comparisons of MSE($\hat{p}$) and MSE($\tilde{p}$) in the previous section were made at the same group size k, and, in particular, at choices $k = k^*$ minimizing MSE($\hat{p}$), provided that $k^* \leq 50$. It is more pertinent to compare MSE($\tilde{p}$), at choice $k = k^*$, with MSE($\tilde{p}$) minimized at choice $k = k^0$ for the specified combinations of N and p. With a slight modification to the definition of k^0, this is done in Table 1, which contains (for each of the 154 combinations) the values $k^0(k^*)$ together with the mean square error efficiency of group testing with group size k^* (using $\hat{p}$ as the estimate) relative to group testing with group size k^0 (using $\tilde{p}$ as the estimate):

$$\text{mean square error efficiency (\%)} = \frac{100[\text{MSE}(\tilde{p}),\ k = k^0]}{[\text{MSE}(\hat{p}),\ k = k^*]}$$

The criterion of minimizing MSE($\tilde{p}$) with respect to k is not sufficient to construct Table 1 for two reasons. First, the calculations were truncated to $k \leq 50$; thus table entries $k^0 = 50$ or $k^* = 50$ indicate that group sizes minimizing MSE($\tilde{p}$) or MSE($\hat{p}$) are not less than 50, and values $k^0 = 50$ or $k^* = 50$ have been used instead. In such cases, the listed mean square error efficiency is biased in favor of $\hat{p}$ because the optimal k^0 is greater than the optimal k^*. Second, there are combinations of N and p, indicated by the symbol ':' in Table 1, where MSE($\tilde{p}$) experiences two local minima over the range $k = 1,2,3,\ldots,50$ and the value k^0 selected

TABLE 1. Optimum group sizes k^0 (k^*) and the corresponding modified mean square error efficiencies (%) for selected combinations of individual transmission rate p and sample size N

p	$N = 10$	$N = 15$	$N = 20$	$N = 25$	$N = 30$	$N = 40$	$N = 50$	$N = 60$	$N = 80$	$N = 100$	$N = 200$
.01	50(33) 58.5	50(50) 84.0	50(50) 90.9	50(50) 92.8	50(50) 94.0	50(50) 95.5	50(50) 96.4	50(50) 97.0	50(50) 97.8	50(50) 98.2	50(50) 99.1
.02	50(19) 46.5	50(28) 63.1	50(35) 73.6	50(42) 81.1	50(47) 86.4	50(50) 91.9	50(50) 93.6	50(50) 94.7	50(50) 96.1	50(50) 96.9	50(50) 98.4
.03	50(14) 41.6	50(20) 61.9	44(25) 72.4	45(29) 78.3	46(32) 82.4	47(38) 87.3	49(41) 90.0	49(44) 91.6	50(46) 93.6	50(47) 94.8	50(50) 97.3
.04	50(11) 31.8	50(16) 56.7	33(19) 73.7	33(22) 79.3	34(25) 83.1	35(29) 87.7	36(31) 90.2	37(33) 91.8	37(34) 93.7	38(35) 94.8	38(37) 97.3
.05	50(9) 22.3	50(13) 43.7	26(16) 74.8 :	27(18) 80.1	27(20) 83.7	28(23) 88.0	29(25) 90.4	29(26) 91.9	30(27) 93.7	30(28) 94.9	31(29) 97.3
.06	50(8) 18.8	50(11) 32.3	22(13) 75.5 :	22(15) 80.7 :	23(17) 84.2 :	23(19) 88.2	24(21) 90.5	24(22) 92.0	25(23) 93.8	25(23) 94.9	25(24) 97.4
.08	38(6) 19.9	42(9) 30.5	16(10) 77.0 :	16(12) 81.8 :	17(13) 84.9 :	17(15) 88.6 :	18(16) 90.8 :	18(16) 92.2	18(17) 94.0	18(17) 95.1	19(18) 97.4
.10	30(5) 20.9	33(7) 31.6	13(8) 77.9 :	13(9) 82.3 :	13(10) 85.4 :	14(12) 89.0 :	14(12) 91.0 :	14(13) 92.4	15(13) 94.1	15(14) 95.2	15(14) 97.5
.15	19(4) 23.4	21(5) 34.1	8(6) 80.5 :	8(6) 84.1 :	9(7) 87.0 :	9(8) 89.8 :	9(8) 91.7 :	9(8) 92.8	9(9) 94.6	10(9) 95.7	10(9) 97.7
.20	14(3) 25.8	16(4) 36.1	6(4) 81.8 :	6(5) 85.9 :	6(5) 88.1 :	7(6) 90.8 :	7(6) 92.5 :	7(6) 93.5	7(6) 94.8	7(7) 95.5	7(7) 97.8
.25	11(3) 26.5	12(3) 38.5	5(3) 82.3 :	5(4) 87.2 :	5(4) 89.1 :	5(4) 90.7 :	5(5) 92.7 :	5(5) 94.1	5(5) 95.6	5(5) 96.5	5(5) 98.3
.30	9(2) 30.1	10(3) 39.3	4(3) 86.3 :	4(3) 88.3 :	4(3) 89.2 :	4(4) 91.7 :	4(4) 93.6 :	4(4) 94.8	4(4) 96.1	4(4) 96.9	4(4) 98.5
.40	6(2) 35.0	7(2) 45.7	3(2) 89.4 :	3(2) 89.8 :	3(2) 89.9 :	3(3) 92.3 :	3(3) 94.2	3(3) 95.2	3(3) 96.5	3(3) 97.2	3(3) 98.6
.50	5(1) 37.7	5(2) 47.8	2(2) 89.8	2(2) 92.9 :	2(2) 94.6 :	2(2) 96.1 :	2(2) 97.0	2(2) 97.5	2(2) 98.1	2(2) 98.5	2(2) 99.3

TABLE 2. Bias and mean square error properties of maximum likelihood estimate $\hat{p}$ and the alternative estimate $\tilde{p}$, for selected group sizes (k) when the individual transmission rate $p = 0.1$ and the sample size $N = 25$

k	Bias($\hat{p}$)	MSE($\hat{p}$)	Bias($\tilde{p}$)	MSE($\tilde{p}$)
7	0.002555	0.000797	0.000031	0.000735
8	0.002795	0.000755	0.000036	0.000686
9	0.003053	0.000732[a]	0.000042	0.000652
10	0.003340	0.000733	0.000050	0.000628
11	0.003684	0.000777	0.000060	0.000614
12	0.004137	0.000915	0.000073	0.000605
13	0.004801	0.001248	0.000088	0.000603[b]
14	0.005847	0.001947	0.000105	0.000605
15	0.007527	0.003266	0.000123	0.000611
20	0.038920	0.032273	0.000071	0.000654
21	0.052627	0.045217	−0.000021	0.000656[c]
22	0.069465	0.061118	−0.000162	0.000653
35	0.469203	0.429973	−0.008624	0.000374
36	0.500732	0.458432	−0.009731	0.000365
37	0.530902	0.485585	−0.010875	0.000362[d]
38	0.559597	0.511337	−0.012049	0.000365
39	0.586741	0.535630	−0.013247	0.000370

[a] Minimum MSE($\hat{p}$) at $k = 9$.
[b] Local minimum MSE($\tilde{p}$) at $k = 13$.
[c] Local maximum MSE($\tilde{p}$) at $k = 21$.
[d] Local minimum MSE($\tilde{p}$) at $k = 37$.

does not correspond to the smaller of these minima because of unacceptable Bias($\tilde{p}$) accompanying the latter. The example in Table 2, for $N = 25$ and $p = 0.1$, should make this clear. For this combination, $k^* = 9$ yields minimum MSE($\hat{p}$) equal to 0.0007324; there is a local minimum value (0.0006026) of MSE($\tilde{p}$) at $k = 13$ accompanied by Bias($\tilde{p}$) = 0.000088, and another (0.0003616) at $k = 37$ accompanied by Bias($\tilde{p}$) = −0.010875 (which is 10.9% of p). In this case, $k^0 = 13$ is selected and the mean square error efficiency is 100(0.0006026)/(0.0007324) = 82.3%; so the entry in Table 1 reads as follows:

13(9)
82.3 :

In those cases where this modified minimum mean square error criterion has been applied, the comparison is biased in favor of $\hat{p}$ (k^*) again. For example, had $k^0 = 37$ been selected in Table 2, the mean square error efficiency would have been 100(0.0003616)/(0.0007324) or 49.4% instead of 82.3%.

CONCLUSIONS

On the basis of its superior bias and mean square error properties, estimate $\tilde{p}$ is preferable to estimate $\hat{p}$ when estimating pathogen transmission rates by group testing for all combinations of N, k, and p of practical interest.

When designing group tests with N fixed by resource limitations, Table 1 can be used to choose near optimal group sizes in the manner described by Swallow (3). There are differences between optimal group sizes k^0 and k^* of practical importance for all $N < 50$, with potential for very much improved testing efficiency when $N < 25$. When mean square error is adopted as the criterion of efficiency, group tests with group sizes k^* (using $\hat{p}$) are very inefficient (20 to 80%) relative to tests with group sizes k^0 (using $\tilde{p}$) for $N < 25$ and p in the working range of most practical interest (< 0.10).

These results reinforce the argument that group testing is usually more efficient than individual testing and extend the range of conditions where this is demonstrable. When $N = 10$ and $p = 0.5$, for example, the value $k^* = 1$ is given by Swallow (3); but even for this extreme combination, $k^0 = 5$ and the efficiency of individual testing is only 38% of optimal group testing.

For each p value in Table 1, k^0 is quite stable in relation to $N \geqslant 20$, increasing slightly with increasing N only for the lower p values. This suggests that it should be possible to formulate a simple approximation for k^0 in relation to p that is serviceable for all $N \geqslant 20$. Values of $\theta^0 = (1-p)^{k^0}$, for all combinations with $N \geqslant 20$ and $k^0 < 50$, are highly concentrated around an average value of 0.237 (range 0.20–0.27). That is to say, the expected proportion of test plants recorded as healthy is approximately 0.237 for optimal group testing when $N \geqslant 20$ and $p \geqslant 0.03$ (and possibly for $p < 0.03$ where $k^0 > 50$ has not been investigated). For very large N, greater than 200, the corresponding theoretical value is 0.20319 obtained as the solution to $\log(\theta) = 2(\theta - 1)$ derived by minimizing the first term of the expression for MSE($\tilde{p}$) given above. Since $\log(0.237) \cong -1.44$, the rule

$$k^0 \cong \text{nearest integer}[-1.44/\log(1-p)],\ 20 \leqslant N \leqslant 200,$$

is suggested. In Table 1, restricted to $N \geqslant 20$ and $k^0 < 50$, deviations of k^0 from this approximation are insignificant from a practical viewpoint: the largest discrepancy is $k^0 \cong 28$ instead of $k^0 = 31$ when $p = 0.05$ and $N = 200$.

LITERATURE CITED

1. Anscombe, F. J. 1956. On estimating binomial response relations. Biometrika 43:461-464.
2. Haldane, J. B. S. 1955. The estimation and significance of the logarithm of a ratio of frequencies. Ann. Hum. Genet. 20:309-311.
3. Swallow, W. H. 1985. Group testing for estimating infection rates and probabilities of disease transmission. Phytopathology 75:882-889.