Techniques

Group Testing for Estimating Infection Rates and Probabilities of Disease Transmission

William H. Swallow

Department of Statistics, North Carolina State University, Raleigh 27695-8203.

The author wishes to recognize Louise R. Romanow, who sparked his interest in this topic and contributed through discussion and critical reading of the manuscript. He also appreciates the encouragement of G. G. Kennedy and J. W. Moyer, and the helpful suggestions of the editor and referees.

This research was supported by USDA-CSRS Competitive Grant 81-CRCR-1-0765. The work was completed while the author was a guest of the International Agriculture Center and Department of Mathematics, Agricultural University, Wageningen, the Netherlands.

Accepted for publication 1 February 1985.

ABSTRACT

Swallow, W. H. 1985. Group testing for estimating infection rates and probabilities of disease transmission. Phytopathology 75: 882-889.

Group-testing or multiple-vector-transfer designs are shown to be usually more efficient than testing individuals for estimating p, which is an infection rate or a probability of pathogen transmission by a single vector. The bias, variance, and mean-squared-error properties of these designs are explored, as some understanding of them is essential to choosing experimental designs that are efficient, convenient, and safe. For the case in which N, the number of tests (or test plants), is limiting, a method is illustrated for selecting k, the number of individuals per test (group size, vectors per test plant), to obtain a near-optimal experimental design. For the case in which $N \times k$ (the total number of individuals [vectors]) is limiting, alternative choices of N and k are compared. Making an appropriate choice for a particular experiment requires considering relative costs and convenience. It is important that treatment differences be judged by comparing estimates of ps, and not by comparing observed fractions of positive tests, since the latter are functions of the ks that were used as well as of the treatments; this applies even when the same value of k is used throughout.

Additional key words: aphid vectors, insect vectors, maximum likelihood estimation, multiple-transfer designs, virus transmission.

An experimental design seen often in studies of insect-vectored plant diseases involves moving one or more vectors (aphids, leafhoppers, etc.) from an infected source plant to each of N noninfected test plants and observing the fraction of the N test plants developing symptoms of disease. No matter how many vectors may be transferred to each test plant, the intent always is, or should be, to estimate

$$p = \text{the probability of disease transmission by a } \textit{single} \text{ vector.} \quad (1)$$

In experiments designed to compare treatments (e.g., virus sources, recipient test plants, or vectors), comparison of treatments can be made by comparing estimates of their transmission probabilities (ps). Of course, one would like to estimate these ps as precisely, yet cheaply, as possible.

The most straightforward way to estimate p is by single-vector transfer, that is, by transferring one vector to each of the N test plants. The estimate of p is then simply the fraction of test plants that develop symptoms. Indeed, provided N can be increased at negligible cost, single-vector-transfer designs are always optimal

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.

(16), but it is seldom, if ever, true that very large numbers of test plants could be used at negligible cost. In practice, the cost of running the experiment is usually highly correlated with N, and in many cases N may even be constrained by the availability of screened cages, glasshouse space, suitable test plants, or some other factor.

An alternative to a single-vector-transfer design is to use a multiple-vector-transfer design or group testing, and to move more than one vector to each test plant. Group testing is not a new idea. It has a long history of application, particularly by British research workers (13,18), and its statistical aspects have been discussed by a number of authors (4,7–9,11,16). That notwithstanding, many workers have chosen not to use it, giving the weak justification that it is risky. Properly used, group testing poses minimal risk and, I would argue, few experimenters can afford *not* to consider using it. The following is therefore intended to make three important points. First, in application, some multiple-vector-transfer or group-testing design will almost always be preferable to a single-vector-transfer design. Second, the user must take care in the selection of a multiple-vector-transfer design, as the penalties for a poor choice can be great. And third, in most cases it is easy to choose a multiple-vector-transfer design which, though not optimal, is both safe and a considerable improvement over a single-vector-transfer design.

ESTIMATION OF INFECTION RATES AND PROBABILITIES OF DISEASE TRANSMISSION BY USING GROUP-TESTING DESIGNS

Using Thompson's (16) notation largely, the basic concept and formulae of group testing can be summarized as follows. The probability that any particular vector fails to transmit disease when moved from an infected source to a test plant is $(1-p)$. If k vectors are moved to the same test plant and they act independently, the probability that none of the k vectors transmits disease is $(1-p)^k$, the product of k probabilities, each being $(1-p)$. Thus, when k vectors are placed on each test plant,

$(1-p)^k$ = the probability a particular test plant is not infected
= the expected fraction of noninfected test plants.

If we define

H = the observed fraction of healthy or noninfected test plants,

then H is an estimator of $(1-p)^k$, that is,

$$\widehat{(1-p)^k} = H,$$

from which one obtains the usual estimator of p,

$$\hat{p} = 1 - H^{1/k}. \tag{2}$$

This estimator is the maximum likelihood (ML) estimator of p; that is, of all values that p could assume, $\hat{p}$ is the one that maximizes the probability or likelihood of the observed data, H. Put another way, the observed fraction of healthy or, equivalently, infected plants was more likely to have been observed when $p = \hat{p}$ than when p equalled any other value.

The importance of using $\hat{p}$, rather than H itself, is that H depends on k as well as on the treatment being tested. Results of tests (of the same or of different treatments) which used different ks can be compared directly through their $\hat{p}$s, but not through their Hs (the raw data). Even when the same $k>1$ has been used for two or more treatments, direct comparison of Hs or $(1-H)$s, fractions of healthy or of infected plants, respectively, is misleading. For example, suppose we wish to compare virus-resistance in two cultivars (same virus source, same vector) for which the true ps are $p_1 = 0.05$ and $p_2 = 0.10$. From $p_2/p_1 = 2$, we would say that the second cultivar is twice as susceptible as the first. Yet, if we use $k = 5$ vectors per test plant, the expected fractions of healthy plants in the two groups are $E(H_1) = 0.77$ and $E(H_2) = 0.59$ from the fact that the expected value of H is $E(H) = (1-p)^k$, giving a ratio of expected fractions of infected plants equal to $(1-0.59)/(1-0.77) = 1.78$. If we instead use $k = 10$ throughout, analogous calculations give $E(H_1) = 0.60$ and $E(H_2) = 0.35$, and a ratio of expected fractions of infected plants equal to 1.63. Thus, if we base our conclusion about relative susceptibility to infection on the ratio of fractions infected, that conclusion will be different with $k = 5$ than with $k = 10$. In fact, any conclusion about relative susceptibility which is based on comparing Hs or $(1-H)$s will in part be determined by the value of k used, and that value of k is often chosen somewhat arbitrarily. Comparison of treatments through their ps avoids this problem. Furthermore, as is shown below, in many experiments it is more appropriate to use different values of k for different treatments, according to their ps.

Although H is an unbiased estimator of $(1-p)^k$, bias is introduced in the taking of the kth root, so $\hat{p}$ is a biased estimator of p (except when $k = 1$). The bias, the difference between the expected or average value of the estimator in repeated application, $E(\hat{p})$, and the true value of p, can be written as

$$\text{Bias}(\hat{p}) = E(\hat{p}) - p, \tag{3}$$

in which

$$E(\hat{p}) = 1 - \sum_{i=0}^{N} \left(\frac{i}{N}\right)^{1/k} \binom{N}{i} [(1-p)^k]^i [1-(1-p)^k]^{N-i}. \tag{4}$$

The variance of $\hat{p}$ can then be expressed as

$$\text{Variance}(\hat{p}) = E[\hat{p} - E(\hat{p})]^2 \tag{5}$$

$$= \sum_{i=0}^{N} \left(\frac{i}{N}\right)^{2/k} \binom{N}{i} [(1-p)^k]^i [1-(1-p)^k]^{N-i} - [1-E(\hat{p})]^2. \tag{6}$$

Although $\hat{p}$ itself depends on the fraction of healthy test plants, H, and not on N per se, increasing N reduces the variance of $\hat{p}$. The calculations of equations 4 and 6 are cumbersome and should be done by computer for all but tiny examples; those equations are given here only to indicate how values reported below may be obtained.

For a biased estimator, mean squared error (MSE) is a more appropriate measure of goodness than is variance. The MSE of $\hat{p}$ is

$$\begin{aligned} \text{MSE}(\hat{p}) &= E[\hat{p} - p]^2 \\ &= \text{Variance}(\hat{p}) + [\text{Bias}(\hat{p})]^2. \end{aligned} \tag{7}$$

$\text{MSE}(\hat{p})$ is the average squared deviation of $\hat{p}$ from the true p, whereas $\text{Variance}(\hat{p})$ is the average squared deviation from the (biased) expected value of $\hat{p}$. For an unbiased estimator, the expected value of the estimator equals the true value of the parameter being estimated, so the variance and MSE are the same.

As equation 7 indicates, MSE incorporates measures of both the accuracy (bias) and precision (variance) of the estimator. The MSE will be small only when the estimator is both accurate (small bias) and precise (small variance).

The bias, variance, and MSE of $\hat{p}$ are complicated functions of p, k, and N, as is evident from equations 4 and 6. However, it is important to have some understanding of the nature of their interrelationships. Figs. 1–3 illustrate the interrelationships for $N = 25$, taken as a typical example.

Bias. Fig. 1 illustrates the bias when $N = 25$ and $k = 1, 2, 3, 4, 5, 7, 10, 15$, and 25, by plotting the mean or expected value of the estimator, $E(\hat{p})$, versus the true p. When $k = 1$, $\hat{p}$ is unbiased [i.e., $E(\hat{p}) = p$] and Fig. 1 shows a straight diagonal line. When $k > 1$, the estimator has positive bias (2,16), that is, $\hat{p}$ overestimates p.

As Fig. 1 shows, the bias is strongly dependent on the value of k. When $k = 2$, the bias is never very great, and is negligible as long as p is less than about 0.5. When $k = 25$, the bias is negligible when $p < 0.06$, but may be huge for larger p. Intermediate values of k

have intermediate curves. Two important points emerge. First, choosing k too large must be avoided. For example, if $p = 0.3$, we would want $k \leqslant 5$ to keep Bias($\hat{p}$) small. Second, even when p is quite large (up to about 0.5 for $N = 25$) there is some $k > 1$ for which the bias is negligible. Since p is likely to be small in practice, it should be easy to choose $k > 1$ for which Bias($\hat{p}$) is negligible.

As one would expect, for larger N, the curves for these same k-values diverge less from the diagonal line and the region of negligible divergence extends to larger values of p; that is, the regions of negligible bias are larger. For smaller N, the regions of negligible bias are smaller, and more care must be exercised in choosing k.

There seems to be no generally satisfactory way to correct for the bias in $\hat{p}$ (7). One should simply choose a value of k for which the bias is tolerable.

Variance. Fig. 2 shows the variance of the estimator plotted against the true value of p for various values of k. The symmetric solid curve is for $k = 1$. Our principal interest is in seeing where curves for $k > 1$ lie below that for $k = 1$. The peaks of curves for $k > 2$ have been cropped, therefore, to allow "blowing up" the more interesting lower portion of the plot, which is what is shown as Fig. 2.

Fig. 2 offers clear evidence that choosing k simply to minimize the variance of $\hat{p}$, a common optimality criterion for unbiased estimators, would be disastrous. For example, when $p = 0.3$, the variance of $\hat{p}$ would be minimized by taking $k = 25$ (among the values of k shown in Fig. 2), since that curve is the lowest at $p = 0.3$. But the bias would be huge, as is seen from Fig. 1 [Bias($\hat{p}$) $= E(\hat{p}) - p = 1.0 - 0.3 = 0.7$, approximately]. The reason the variance is small for large k with even moderate p is that, when k is sufficiently large (how large depends on p), all test plants almost always become infected, giving small variance. Of course, values of k that yield badly biased $\hat{p}$s are undesirable.

Mean squared error. Fig. 3 amalgamates the information in Figs. 1 and 2, plotting MSE($\hat{p}$) versus the true value of p. As noted below equation 7, mean squared error incorporates both bias and variance, and will be inflated by either large bias (inaccuracy of the estimator) or large variance (poor precision).

Fig. 3 prompts several comments. First, when p is small, as is likely in practice, taking $k > 1$ rather than $k = 1$ can greatly reduce the MSE, often manyfold. Second, the smaller the value of p, the larger the optimal k. Third, even where in the figure $k = 15$ or $k = 25$ is optimal ($p \leqslant 0.08$), most of the reduction in MSE can be realized with a smaller k, say, $k = 5$. And fourth, although group testing is often discussed as being appropriate when p is very small, and indeed its benefits are greatest then, it is useful for larger p than is often supposed. In the case shown here ($N = 25$), group testing ($k > 1$) is preferable to single-vector transfer ($k = 1$) for all p between 0 and 0.58.

DESIGN CONSIDERATIONS

When N is fixed. The classic context for group testing is when N is considered fixed and the experimental design question is "What is the optimal k?" This situation arises either when N is set by, say, the number of test plants one can accommodate, but k can be whatever value the experimenter wishes, or when experimental costs (time, cost of laboratory analyses, etc.) are principally determined by N, not k, so N is more or less fixed by the budget.

Table 1 gives the following for a broad range of combinations of p and N: k^*, the value of k that is optimal in the sense that it minimizes MSE($\hat{p}$); the bias and MSE of $\hat{p}$ when $k = k^*$ [Bias($\hat{p}$;k^*) and MSE($\hat{p}$;k^*)]; and, for comparison, the MSE of $\hat{p}$ when $k = 1$ [MSE($\hat{p}$;1)]. For example, when $p = 0.05$ and $N = 25$, Table 1 indicates that the optimal number of vectors per test plant is $k^* = 18$ for which Bias($\hat{p}$) = 0.0016 and MSE($\hat{p}$) = 0.000200; if $k = 1$ had been used instead, $\hat{p}$ would have had MSE($\hat{p}$) = 0.001900, which is 9.5 times the minimum value realized with k^*. For $k = 1$, Bias($\hat{p}$;1) = 0 always. The column labeled $N = 200$ can be used for any $N > 200$. Table 1 is based on calculations using equations 4, 6, and 7 for $k = 1$ to 25 by 1, and 25 to 50 by 5. Recorded values of k^* greater than 25 are correct to within 5, except that $k^* = 50$ is the largest value considered. This degree of uncertainty about the exact value of $k^* > 25$ is of no practical concern, as will become clear. Portions of Table 1 are available elsewhere (8,9,11).

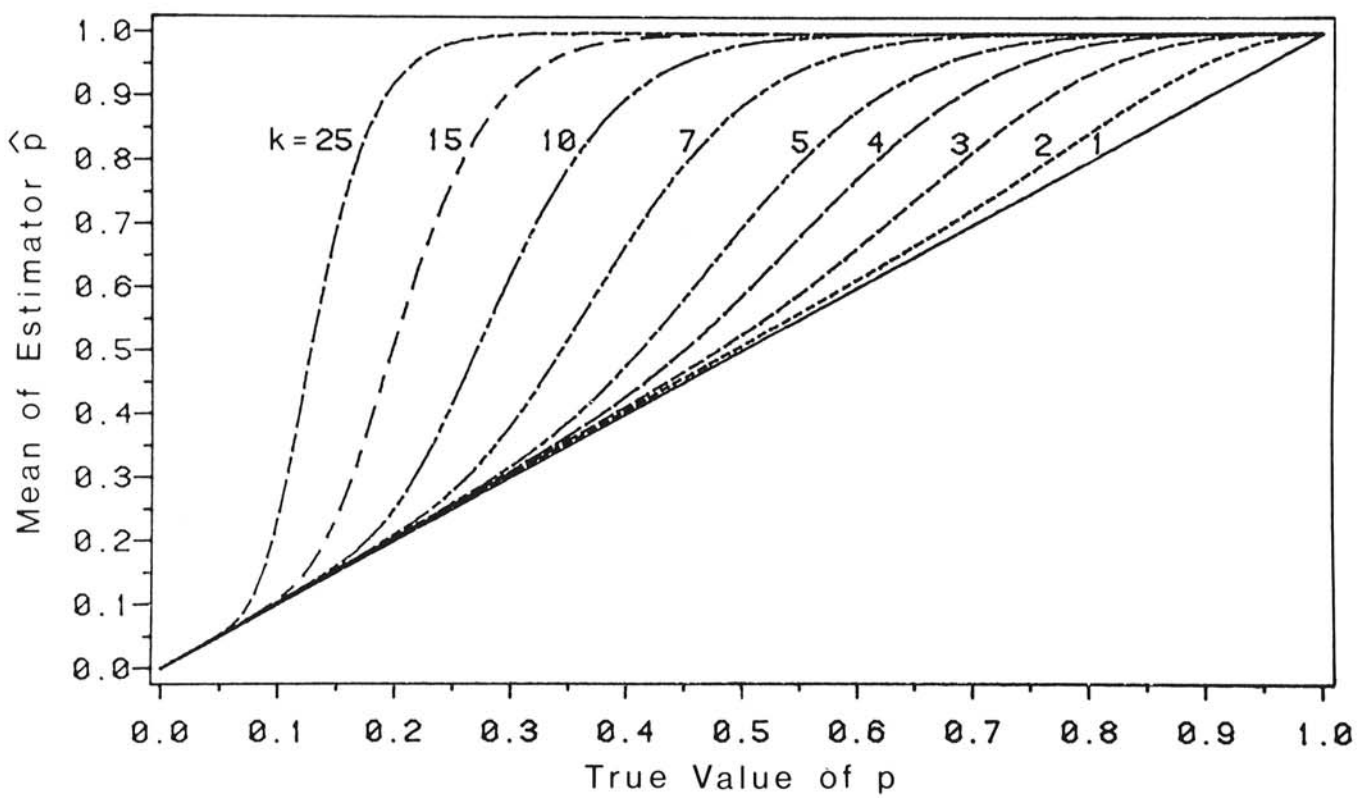

Fig. 1. Expected value (mean) of the maximum likelihood estimator ($\hat{p}$) of the infection rate or probability (p) of disease transmission by a single vector versus the true value of p for tests employing $k = 1$ to 25 vectors per test plant with $N = 25$ test plants.

Comparison of MSE($\hat{p}$;k*) with MSE($\hat{p}$;1) shows the gains attainable with group testing. When p is small, MSE($\hat{p}$;1) is often 10–20 times the minimum MSE, MSE($\hat{p}$;k*). Increasing N reduces both MSE($\hat{p}$;k*) and MSE($\hat{p}$;1), but increases the relative advantage of group testing. For example, when $p = 0.05$ and $N = 25$, MSE($\hat{p}$;1)/MSE($\hat{p}$;k*) = 0.001900/0.000200 = 9.5, but when $N = 50$ the ratio is 0.000950/0.000084 = 11.3, and when $N = 100$ the ratio is 11.9. Of course, increasing N changes k*, too. As Table 1 also shows, the bias in $\hat{p}$ when the optimal k is used is never large, and decreases with increasing N (i.e., across a row of the table).

To reduce MSE($\hat{p}$) with a single-vector-transfer design (k = 1), one must double N to halve MSE($\hat{p}$). This follows from the fact that, for the special case when $k = 1$, equation 6 simplifies

TABLE 1. Optimal number of vectors (k*) per plant to use in estimating infection rate or probability (p) of disease transmission by a single vector, when the number of test plants (N) is fixed, and the bias [Bias($\hat{p}$;k*)] and mean squared error (MSE) of the estimator of p for k* [MSE($\hat{p}$;k*)] or one vector [MSE ($\hat{p}$;1)] per test plant

	N										
p	10	15	20	25	30	40	50	60	80	100	200
0.01k*	35	50	50	50	50	50	50	50	50	50	50
Bias($\hat{p}$;k*)	0.0007	0.0005	0.0003	0.0003	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.00003
MSE($\hat{p}$;k*)	0.000046	0.000021	0.000014	0.000011	0.000009	0.000007	0.000005	0.000004	0.000003	0.000003	0.000001
MSE($\hat{p}$;1)	0.000990	0.000660	0.000495	0.000396	0.000330	0.000248	0.000198	0.000165	0.000124	0.000099	0.000050
0.02k*	19	30	35	40	45	50	50	50	50	50	50
Bias($\hat{p}$;k*)	0.0013	0.0010	0.0008	0.0006	0.0006	0.0004	0.0004	0.0003	0.0002	0.0002	0.0001
MSE($\hat{p}$;k*)	0.000162	0.000078	0.000048	0.000035	0.000027	0.000019	0.000015	0.000012	0.000009	0.000007	0.000003
MSE($\hat{p}$;1)	0.001960	0.001307	0.000980	0.000784	0.000653	0.000490	0.000392	0.000327	0.000245	0.000196	0.000098
0.03k*	14	20	25	30	30	35	40	45	45	45	50
Bias($\hat{p}$;k*)	0.0020	0.0014	0.0012	0.0010	0.0008	0.0007	0.0006	0.0005	0.0004	0.0003	0.0002
MSE($\hat{p}$;k*)	0.000337	0.000164	0.000104	0.000076	0.000059	0.000041	0.000031	0.000025	0.000018	0.000015	0.000007
MSE($\hat{p}$;1)	0.002910	0.001940	0.001455	0.001164	0.000970	0.000728	0.000582	0.000485	0.000364	0.000291	0.000146
0.04k*	11	16	19	22	25	30	30	35	35	35	35
Bias($\hat{p}$;k*)	0.0026	0.0019	0.0015	0.0013	0.0012	0.0010	0.0008	0.0007	0.0006	0.0004	0.0002
MSE($\hat{p}$;k*)	0.000565	0.000281	0.000180	0.000131	0.000102	0.000072	0.000055	0.000045	0.000032	0.000026	0.000012
MSE($\hat{p}$;1)	0.003840	0.002560	0.001920	0.001536	0.001280	0.000960	0.000768	0.000640	0.000480	0.000384	0.000192
0.05k*	9	13	16	18	20	23	25	25	25	30	30
Bias($\hat{p}$;k*)	0.0032	0.0024	0.0020	0.0016	0.0015	0.0012	0.0010	0.0008	0.0006	0.0006	0.0003
MSE($\hat{p}$;k*)	0.000842	0.000424	0.000274	0.000200	0.000157	0.000110	0.000084	0.000069	0.000050	0.000040	0.000019
MSE($\hat{p}$;1)	0.004750	0.003167	0.002375	0.001900	0.001583	0.001188	0.000950	0.000792	0.000594	0.000475	0.000238
0.06k*	8	11	13	15	17	19	21	22	23	23	24
Bias($\hat{p}$;k*)	0.0038	0.0028	0.0023	0.0019	0.0018	0.0014	0.0012	0.0010	0.0008	0.0006	0.0003
MSE($\hat{p}$;k*)	0.001158	0.000592	0.000385	0.000282	0.000222	0.000156	0.000120	0.000098	0.000071	0.000056	0.000027
MSE($\hat{p}$;1)	0.005640	0.003760	0.002820	0.002256	0.001880	0.001410	0.001128	0.000940	0.000705	0.000564	0.000282
0.08k*	6	9	10	12	13	15	16	16	17	17	18
Bias($\hat{p}$;k*)	0.0048	0.0038	0.0030	0.0027	0.0023	0.0019	0.0016	0.0013	0.0010	0.0008	0.0004
MSE($\hat{p}$;k*)	0.001922	0.001002	0.000656	0.000484	0.000382	0.000269	0.000208	0.000170	0.000124	0.000097	0.000047
MSE($\hat{p}$;1)	0.007360	0.004907	0.003680	0.002944	0.002453	0.001840	0.001472	0.001227	0.000920	0.000736	0.000368
0.10k*	5	7	8	9	10	12	12	13	13	14	14
Bias($\hat{p}$;k*)	0.0057	0.0045	0.0036	0.0031	0.0027	0.0023	0.0018	0.0016	0.0012	0.0010	0.0005
MSE($\hat{p}$;k*)	0.002807	0.001482	0.000987	0.000732	0.000579	0.000409	0.000317	0.000258	0.000189	0.000149	0.000072
MSE($\hat{p}$;k*)	0.009000	0.006000	0.004500	0.003600	0.003000	0.002250	0.001800	0.001500	0.001125	0.000900	0.000450
0.15k*	4	5	6	6	7	8	8	8	9	9	9
Bias($\hat{p}$;k*)	0.0086	0.0064	0.0054	0.0042	0.0040	0.0033	0.0026	0.0022	0.0018	0.0014	0.0007
MSE($\hat{p}$;k*)	0.005409	0.002976	0.002014	0.001516	0.001202	0.000858	0.000665	0.000544	0.000398	0.000314	0.000152
MSE($\hat{p}$;1)	0.012750	0.008500	0.006375	0.005100	0.004250	0.003188	0.002550	0.002125	0.001594	0.001275	0.000638
0.20k*	3	4	4	5	5	6	6	6	6	7	7
Bias($\hat{p}$;k*)	0.0099	0.0083	0.0059	0.0058	0.0047	0.0042	0.0033	0.0027	0.0020	0.0019	0.0009
MSE($\hat{p}$;k*)	0.008356	0.004764	0.003284	0.002459	0.001975	0.001416	0.001100	0.000901	0.000662	0.000523	0.000253
MSE($\hat{p}$;1)	0.016000	0.010667	0.008000	0.006400	0.005333	0.004000	0.003200	0.002667	0.002000	0.001600	0.000800
0.25k*	3	3	3	4	4	4	5	5	5	5	5
Bias($\hat{p}$;k*)	0.0146	0.0085	0.0062	0.0067	0.0055	0.0040	0.0041	0.0034	0.0025	0.0020	0.0010
MSE($\hat{p}$;k*)	0.012089	0.006652	0.004735	0.003514	0.002831	0.002056	0.001597	0.001306	0.000959	0.000757	0.000370
MSE($\hat{p}$;1)	0.018750	0.012500	0.009375	0.007500	0.006250	0.004688	0.003750	0.003125	0.002344	0.001875	0.000938
0.30k*	2	3	3	3	3	4	4	4	4	4	4
Bias($\hat{p}$;k*)	0.0104	0.0118	0.0083	0.0064	0.0053	0.0056	0.0044	0.0036	0.0027	0.0021	0.0011
MSE($\hat{p}$;k*)	0.014516	0.008861	0.006000	0.004616	0.003769	0.002716	0.002114	0.001733	0.001276	0.001009	0.000494
MSE($\hat{p}$;1)	0.021000	0.014000	0.010500	0.008400	0.007000	0.005250	0.004200	0.003500	0.002625	0.002100	0.001050
0.40k*	2	2	2	2	2	3	3	3	3	3	3
Bias($\hat{p}$;k*)	0.0169	0.0100	0.0072	0.0056	0.0047	0.0065	0.0051	0.0042	0.0031	0.0025	0.0012
MSE($\hat{p}$;k*)	0.020264	0.011981	0.008632	0.006780	0.005589	0.004032	0.003143	0.002580	0.001902	0.001506	0.000739
MSE($\hat{p}$;1)	0.024000	0.016000	0.012000	0.009600	0.00800	0.006000	0.004800	0.004000	0.003000	0.002400	0.001200
0.50k*	1	2	2	2	2	2	2	2	2	2	2
Bias($\hat{p}$;k*)	0.0000	0.0157	0.0108	0.0082	0.0067	0.0049	0.0039	0.0032	0.0024	0.0019	0.0009
MSE($\hat{p}$;k*)	0.025000	0.015722	0.01759	0.008247	0.006722	0.004932	0.003901	0.003228	0.002400	0.001911	0.000946
MSE($\hat{p}$;1)	0.025000	0.016667	0.012500	0.010000	0.008333	0.006250	0.005000	0.004167	0.003125	0.002500	0.001250

tremendously to

$$\text{Variance}(\hat{p}) = p\,(1-p)/N \qquad (8)$$

which is $\text{MSE}(\hat{p})$ also, since $\text{Bias}(\hat{p}) = 0$. Equation 8 is helpful. For example, it can be used to show that to obtain an equally good $\hat{p}$ (equal MSE) as realized with the optimal multiple-vector-transfer design with $N = 25$ and $p = 0.05$ [$\text{MSE}(\hat{p}) = 0.000200$], a single-vector-transfer design would require $N = 238$ [using equation 8 and solving $0.000200 = 0.05(1-0.05)/N$, in which $p = 0.05$]. The

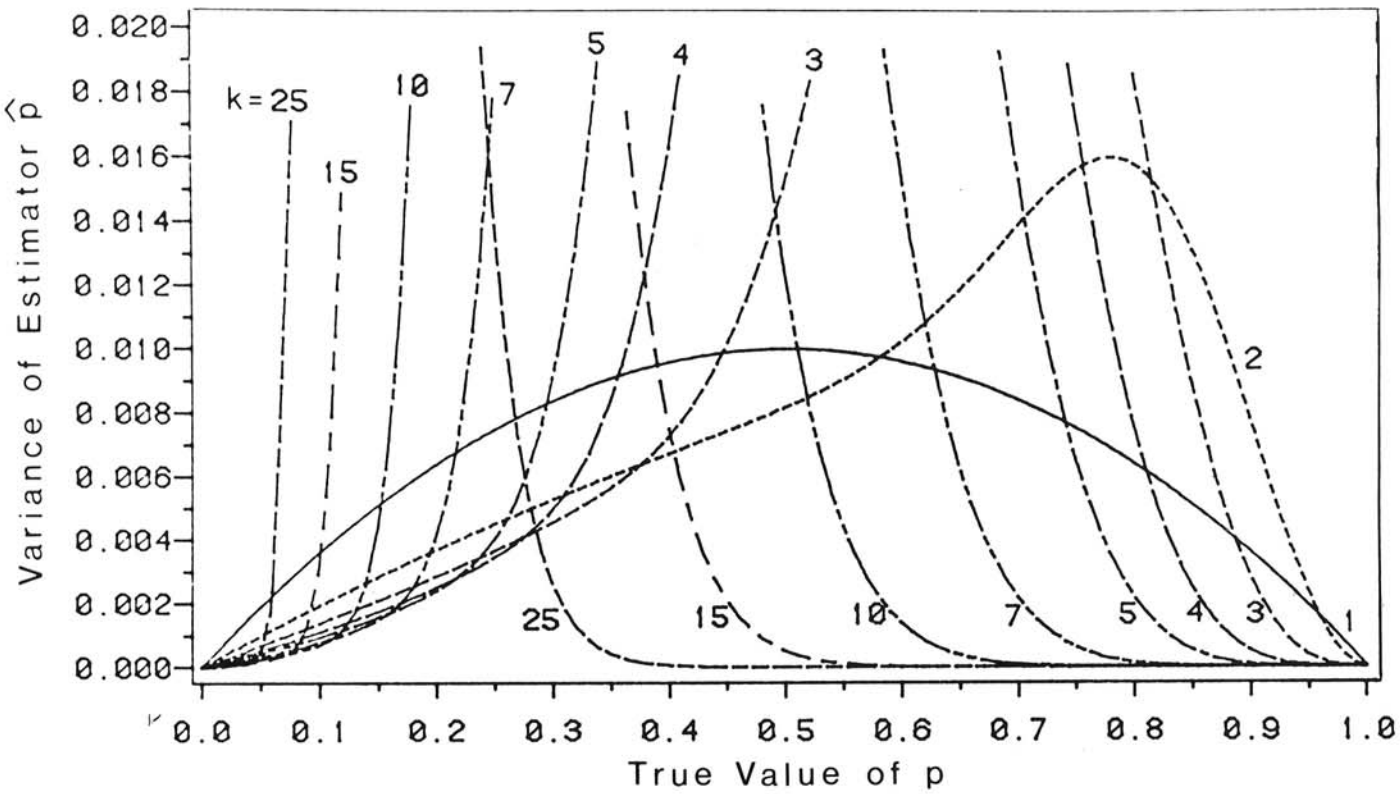

Fig. 2. Variance of the maximum likelihood estimator ($\hat{p}$) of the infection rate or probability (p) of disease transmission by a single vector versus the true value of p for tests employing $k = 1$ to 25 vectors per test plant with $N = 25$ test plants.

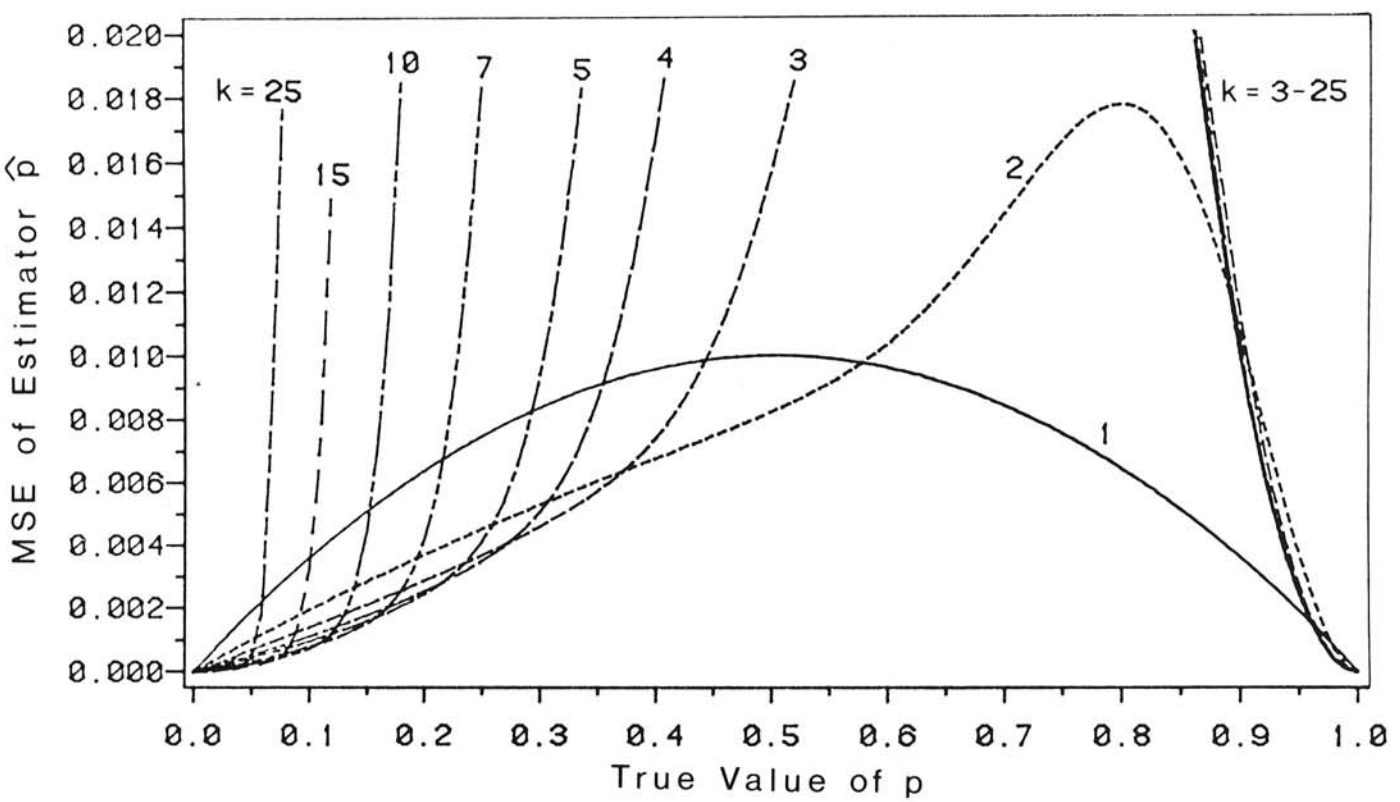

Fig. 3. Mean squared error of the maximum likelihood estimator ($\hat{p}$) of the infection rate or probability (p) of disease transmission by a single vector versus the true value of p for tests employing $k = 1$ to 25 vectors per test plant with $N = 25$ test plants.

optimal multiple-vector-transfer design with $k^* = 18$ would have used $18 \times 25 = 450$ vectors, not 238, but we are assuming at the moment that N, not k, determines the cost or is otherwise difficult to increase. By this method one can determine, for any of the multiple-vector-transfer designs of Table 1, the N needed to obtain as good a $\hat{p}$ using $k = 1$. These calculations can be quite startling, especially for a small p.

The expected fraction of infected plants is $E(1 - H) = 1 - E(H) = 1 - (1 - p)^k$. Even for the optimal designs given in Table 1, this ranges from 0.30 to 0.79, and depends on p and k. For example, when $p = 0.04$ and $N = 25$, it equals $1 - (1 - 0.04)^{22} = 0.59$, in which $k^* = 22$ from Table 1. For the designs in Table 1, the expected fraction of infected plants is highest when both p and N are small, and decreases as either p or N increases. Of course, in any application the observed fraction of infected plants will vary around the expected fraction.

Although Figs. 1–3 and Table 1 illustrate many of the key points in the theory of group testing, the discussion based on them has glossed over one point that is critical in application, namely, that p is unknown. In practice, one must choose an appropriate k for an experiment without knowing p. In effect, one wants to use Table 1 to choose the optimal k, but doesn't know which row of the table to enter. A very workable solution is to enter Table 1 with N and p_e, a value which is believed to be an upper bound for p (16). For example, if one feels reasonably confident that the unknown p is in the range 0.01 to 0.10, then one can enter Table 1 with $p_e = 0.10$ to obtain a value of k for the proposed experiment. Often one already has enough preliminary data or other information to do this. If not, a small trial can be run to get a rough estimate of p; using some small $k > 1$ in this preliminary trial is generally advisable, both for increased efficiency (versus using $k = 1$) and because any bias introduced thereby, being positive, can only lead to overestimation of p, making the choice of p_e more conservative (7).

The reason for taking $p_e \geq p$ is illustrated in Table 2 for a hypothetical example in which $N = 25$ and the true (unknown in practice) $p = 0.10$. As Table 2 indicates, if, in planning the experiment to estimate p, one serendipitously enters Table 1 with $p_e = 0.10$ (the true p), one is told to use $k = 9$, the optimal value since p_e equals the true p, for which $MSE(\hat{p}) = 0.0007$. If one enters with $p_e = 0.20$, one is told to use $k = 5$ for which $MSE(\hat{p}) = 0.0010$, and so forth. Although p_e determines the value of k to be used, the associated bias and MSE values shown in Table 2 are calculated using the true $p = 0.10$, not p_e, with that k; they could not be calculated in practice, since the true p would be unknown.

One sees from Table 2 that when p_e exceeds the true p, a smaller than optimal k is used, but most of the potential gains of group testing are still realized. For example, when $p_e = 0.20$, twice the true p, $MSE(\hat{p}) = 0.0010$, which differs little from the minimum $MSE(\hat{p}) = 0.0007$, and is still less than 30% of the value $MSE(\hat{p}) = 0.0036$ for $k = 1$. Using p_e too large leads to a smaller bias than that found with the optimal k. On the other hand, using a p_e much smaller than the true p leads to using k too large, with $\hat{p}$ then suffering from (perhaps greatly) increased bias and MSE. Although, ideally, one would like p_e as close to the true p as possible, if in doubt, take p_e too large for a conservative choice of k.

Thus, in application, one cannot determine the optimal k, since to do so one must know p, the probability to be estimated. However, by using Table 1 as suggested, with a sensibly chosen p_e, it is easy to select a multiple-vector-transfer design ($k > 1$) which is more efficient than a single-vector-transfer design. The smaller p is, the larger the gain in efficiency is likely to be. The only risk is in using p_e too small and, thereby, k too large. Knowing that, the user can keep the risk small by a conservative choice of p_e or, equivalently, k.

A brief comment should be made about available asymptotic results, and a warning given to potential users. These results appear in a number of references (8,9,14,16), usually with appropriate cautions, but their simplicity makes indiscriminate use overly tempting. Roughly speaking, asymptotic results are results that are correct with infinite sample sizes (there may be other conditions which must be met, too), and may be approximately correct and thus useful with smaller sample sizes. The asymptotic result that is relevant here is this: when N is sufficiently large and k is large relative to p (the latter condition being satisfied in most applications of interest), then the optimal k is approximately $k = 1.5936/p$. But how large an N is "sufficiently large" for this asymptotic result to be reasonably accurate? The answer is "Larger than the values of N commonly used, unfortunately!" Values of $k = 1.5936/p$ are larger than the values of k^* given under $N = 200$ in Table 1. For $N < 25$, and perhaps larger, these values of k are clearly too big, especially when p is small. They often greatly exceed the k^* values in Table 1. Using the asymptotically optimal k when one has only moderate sample sizes (N) always leads to adopting too large a value of k and, as has already been shown, this is exactly what we wish to avoid. The user is better advised to use results based on exact calculations (using equations 4 and 6), as in Table 1.

When $N \times k$ is fixed. A less common situation, but one which also occurs, is that in which it is more appropriate to think of the number of vectors ($N \times k$), rather than the number of test plants (N), as fixed. Perhaps the number of available vectors is limited, or labor constraints (time or cost) determine that only a certain number of vectors can be transferred. The problem is then to choose both N and k such that the product, $N \times k$, equals the predetermined value.

As mentioned earlier, if costs are unrelated to N (or are ignored), then $k = 1$ is always optimal. That is, it will always be best to use N test plants with $k = 1$ vector per test plant, and the reason can be seen intuitively as follows. When $k = 1$, we determine for *each* vector whether it has transferred the disease. When $k > 1$, an infected plant indicates that *one or more* of the k vectors has transmitted the disease, but there is uncertainty as to whether exactly one or more than one vector transmitted the disease. This uncertainty is the price paid by the multiple-vector-transfer design—the single-vector-transfer design provides more precise information. Extending this argument, the smaller the value of p, the smaller is the relative likelihood that more than one of k, rather than exactly one of k, caused a plant to become infected. In other words, the smaller p is, the more certain it becomes that an infected plant was infected by only one of the k vectors and, therefore, the less bias there is in $\hat{p}$.

However, costs cannot be ignored altogether. Even when costs cannot be calculated precisely, some choices of N and k will appear more cost-efficient, or at least more convenient, than others. Table 3 gives $MSE(\hat{p})$s for combinations of several values of p with some possible choices of N and k for which $(N \times k) = 100, 500, 1{,}000$, or 2,000. This allows comparison of different ways of allocating $N \times k$ vectors to N test plants with k vectors each. The minimum $MSE(\hat{p})$ when $N \times k$ is fixed is always attained with $k = 1$, as discussed above. However, especially when p is small, other allocations yield $MSE(\hat{p})$s which differ little from the minimum. For example, when $(N \times k) = 500$ and $p = 0.01$, the estimate of p based on $k = 1$ and $N = 500$ is only slightly better than that based on $k = 25$ and $N = 20$, but uses 25 times as many plants. When $(N \times k) = 500$ and $p = 0.05$, $MSE(\hat{p})$ with $k = 20$ and $N = 25$ is about twice that for $k = 1$ and $N = 500$, but requires one twentieth as many test plants.

TABLE 2. Example showing the relationship between the value (p_e) used to enter Table 1, and the bias [Bias($\hat{p}$)] and mean squared error [MSE($\hat{p}$)] of the resulting estimator of infection rate or probability (p) of disease transmission by a single vector, when $N = 25$ test plants are used and the true $p = 0.10$

p_e	k from Table 1	Bias($\hat{p}$)	MSE($\hat{p}$)
0.02	40	0.6123	0.5584
0.04	22	0.0695	0.0611
0.06	15	0.0075	0.0033
0.08	12	0.0041	0.0009
0.10 (true p)	9	0.0031	0.0007
0.15	6	0.0023	0.0009
0.20	5	0.0021	0.0010
0.25	4	0.0018	0.0011
	$k = 1$	0	0.0036

I have considered two basic cases (N fixed and $N \times k$ fixed) both because each of these cases occurs in practice, and because each is a convenient vehicle for illustrating some of the principles of group testing. However, many experiments do not fall neatly into either case. Perhaps N is fixed, but costs also depend on k. Tables 1 and 3, and the principles they elucidate, can still be exploited in arriving at a sensible experimental design. The tables can be used directly, or the researcher can, starting with values given in Tables 1 and 3, compare an even wider variety of possible designs by making some additional simple calculations. In doing so, the following fact is useful: starting with a design in Tables 1 or 3 and holding k constant, increasing N will decrease MSE($\hat{p}$) at least proportionally. For example, doubling N will reduce MSE($\hat{p}$) by at least half. When $k > 1$ this is conservative, because MSE($\hat{p}$) decreases proportionally faster than N increases. [Note that going the other way (decreasing N) poses some risk. When $k > 1$, halving N more than doubles MSE($\hat{p}$), markedly so when N is small.] Thus, continuing the last example of the previous paragraph, it would be expected that doubling N would achieve with $k = 20$ and $N = 50$, approximately the same MSE($\hat{p}$) as with $k = 1$ and $N = 500$. Compared to using $k = 1$ and $N = 500$, this requires twice as many vectors, but only one tenth as many test plants. [In this case, the value MSE($\hat{p}$) = 0.000088 for $k = 20$ and $N = 50$ appears in Table 3 under ($N \times k$) = 1,000]. After comparing a variety of alternative designs, the researcher can choose an appropriate design given the ways cost and convenience depend on N and k for the particular experiment.

Biological considerations and statistical assumptions. I have focused thus far on statistical considerations in choosing k, without reference to biological considerations. In fact, the two are closely interrelated. The following brief comments are meant only to suggest the kinds of issues involved; the issues are important, but the specifics depend on the particular application.

A key assumption in the theory of multiple-vector-transfer designs is that each of the k vectors transferred to a particular test plant transmits the virus with probability independent of k, the number of vectors on that plant. For most aphid-vectored plant virus diseases, for example, the weight of evidence is that this independence assumption is valid (3,15,18). However, in some applications one may have to impose a limit on k if the notion of independently operating vectors is to be plausible (4,16). For example, it may be that the behavior of the vectors is altered in some important way when their density on the plant is too great, or the response of the plant itself may be affected. This is another reason to use a value of k which is perhaps smaller than the statistically optimal one, and a reason why values of $k > 50$ were not considered in constructing Table 1.

Another assumption, usually less important than that of independent action, is that each vector has the same probability of transmitting the disease. This requires appropriate standardization of exposure to the disease source and opportunity to infect the test plant. If, for example, certain leaves of the test plants are more susceptible than others, then this should be considered in the experimental design.

CONFIDENCE INTERVALS

For a single p. An approximate confidence interval for p can be constructed from the point estimate, $\hat{p}$, as

$$\hat{p} \pm z\,[\widehat{\text{Variance}(\hat{p})}]^{1/2}, \qquad (9)$$

in which, for a 95% confidence interval, z equals 1.96, the value of the standard normal random variable (Z) which is exceeded with probability $0.05/2 = 0.025$. $\widehat{\text{Variance}(\hat{p})}$, the estimated variance of $\hat{p}$, can be obained as

$$\widehat{\text{Variance}(\hat{p})} = [1 - (1 - \hat{p})^k]/[Nk^2(1 - \hat{p})^{k-2}]. \qquad (10)$$

Equation 10 follows from the asymptotic variance of $\hat{p}$ (16), but is satisfactory for the combinations of N, p, and k of Table 1, or for those N and p with smaller k. One can also estimate Variance($\hat{p}$) by using equations 4 and 6 with $\hat{p}$ in place of p on the right-hand sides of those equations, but the gain is not enough to justify the considerable extra effort. The confidence interval from equation 9 will be approximate in any case, because $\hat{p}$ is not normally distributed; the approximation improves with increasing N. Construction of exact confidence intervals is very tedious, but has been done for a few special cases (4).

TABLE 3. Values of the mean squared error [MSE($\hat{p}$)] of the estimator of infection rate or probability (p) of disease transmission by a single vector, for combinations of values of p with some alternative choices of the number of test plants (N) and number of vectors (k) per test plant for which ($N \times k$) = 100, 500, 1,000, or 2,000

			p				
$N \times k$	k	N	0.01	0.05	0.10	0.20	0.30
100	1	100	0.000099	0.000475	0.000900	0.001600	0.002100
	2	50	0.000101	0.000493	0.000962	0.001828	0.002601
	4	25	0.000104	0.000535	0.001120	0.002543	0.005080
	5	20	0.000106	0.000560	0.001224	0.003365	0.017278
	10	10	0.000116	0.000851	0.012960	0.206815	0.370598
	20	5	0.000345	0.098504	0.424224	0.604807	0.488246
500	1	500	0.000020	0.000095	0.000180	0.000320	0.000420
	5	100	0.000020	0.000107	0.000228	0.000542	0.001039
	10	50	0.000021	0.000126	0.000327	0.003535	0.117996
	20	25	0.000023	0.000207	0.032273	0.479741	0.480773
	25	20	0.000024	0.001609	0.182826	0.593906	0.488781
	50	10	0.000123	0.405333	0.769364	0.639912	0.490000
1,000	1	1,000	0.000010	0.000048	0.000090	0.000160	0.000210
	5	200	0.000010	0.000053	0.000113	0.000267	0.000501
	10	100	0.000010	0.000062	0.000157	0.000624	0.029329
	20	50	0.000011	0.000088	0.001623	0.358913	0.471420
	25	40	0.000012	0.000111	0.041524	0.550674	0.487520
	40	25	0.000013	0.029133	0.558438	0.637927	0.489993
	50	20	0.000014	0.182045	0.730651	0.639823	0.490000
2,000	1	2,000	0.000005	0.000024	0.000045	0.000080	0.000105
	10	200	0.000005	0.000031	0.000077	0.000285	0.002656
	20	100	0.000006	0.000042	0.000170	0.201033	0.453108
	25	80	0.000006	0.000050	0.002366	0.473295	0.484978
	40	50	0.000006	0.001031	0.384830	0.635836	0.489986
	50	40	0.000007	0.036797	0.658927	0.639644	0.490000

It should be noted that when none or all of the test plants become infected ($H = 1$ or 0, $\hat{p} = 0$ or 1), one cannot estimate Variance($\hat{p}$). Choices of N and k which are likely to yield such data should be avoided.

For a difference between *p*s. Many experiments are run to compare two or more treatments, in which case there will be one p to estimate for each treatment. One need not use the same value of k or of N for all treatments in estimating these ps. Indeed, it may be inappropriate to do so, since good choices of k and N depend on p, and it may be known (or strongly suspected) in advance that different treatments have ps of different magnitudes. When all treatments are viewed as equally important or interesting, one may choose to strive for $\hat{p}$s with approximately equal MSEs. Or, if some treatments are of greater importance than others, available resources can be deliberately allocated unequally to provide better estimates for the more important treatments.

An approximate confidence interval for the difference between p_1 and p_2 can be calculated as

$$(\hat{p}_1 - \hat{p}_2) \pm z\,[\widehat{\text{Variance}(\hat{p}_1)} + \widehat{\text{Variance}(\hat{p}_2)}]^{1/2}. \qquad (11)$$

An approximate two-tailed test of hypothesis regarding equality of p_1 and p_2 follows. If the 95% confidence interval includes the value zero, then p_1 and p_2 are declared to be not significantly different at the 5% level of significance. Otherwise, they are declared to be different. This can be done for any or all pairs of ps. [Although a detailed discussion is beyond the scope of this paper, I note in passing that available data sometimes provide more than one $\hat{p}$ for each of some or all of the treatments. In that case, one may wish to use analysis of variance on the collection of $\hat{p}$s (with weighting or after appropriate transformation, if necessary) to test treatment differences.]

EXTENSIONS AND RELATED PROBLEMS

Brief mention should be made of several extensions of the group-testing model which deal with more complicated problems. First, results have been obtained for the case in which different values of k (pool size) have been used on subsets of the N test plants (pools) that will be used to estimate a single p (4,10,17). As this complicates the problem considerably, I recommend using the same k within any group of test plants, if possible. Second, Sobel and Elashoff (14) have considered models that are applicable when the same individual can be retested in more than one pool. Retesting is usually not possible in vector-transfer designs, but may be in other contexts, for example, where portions of the sample from an individual can be entered into different pools. Third, Bhattacharyya et al (2) have considered finite population models. In most experiments, including vector-transfer designs, the population is viewed as infinite (conceptual, not real).

Although the presentation above is in the context of vector-transfer designs, the discussion, tables, and figures have broader applicability. For example, in estimating the fraction of a human population that has been exposed to a rare viral disease, one may wish to test for prevalence of antigen in pooled blood samples, each sample containing blood of k individuals. Or, in determining the proportion of vectors carrying a virus, one may want to test pools of k vectors each (perhaps with ELISA), rather than test individual vectors, to control laboratory costs. The problem is the same—to choose the optimal k.

Other applications are closely related, but less obviously so (12,16). One example is estimation of bacterial densities by the standard "most probable number," whereby a solution containing bacteria is diluted, and unit volumes are plated, cultured, and classified as containing bacteria or not (5,6). A second is grid sampling to estimate the density of a plant species (or of infected plants) by recording its presence or absence in each of a number of randomly selected quadrats (squares of the grid) (1). Although solutions to these problems are customarily formulated by using models based on the Poisson distribution, and the group-testing model is based on the binomial distribution, when p is small relative to k, the Poisson and binomial are nearly indistinguishable. Thus, choosing the optimal k for group testing is analogous to choosing the optimal dilution factor for estimating bacterial density, or the optimal grid size for estimating plant density.

LITERATURE CITED

1. Bartlett, M. S. 1935. Mathematical appendix to Blackman, G. E. A study by statistical methods of the distribution of species in grassland associations. Ann. Bot. 49:749-777.
2. Bhattacharyya, G. K., Karandinos, M. G., and DeFoliart, G. R. 1979. Point estimates and confidence intervals for infection rates using pooled organisms in epidemiological studies. Am. J. Epidemiol. 109:124-131.
3. Bindra, O. S., and Sylvester, E. S. 1961. Effect of insect numbers on aphid transmission of potato leafroll virus. Hilgardia 31:279-325.
4. Chiang, C. L., and Reeves, W. C. 1962. Statistical estimation of virus infection rates in mosquito vector populations. Am. J. Hyg. 75:377-391.
5. Cochran, W. G. 1950. Estimation of bacterial densities by means of the "most probable number." Biometrics 6:105-116.
6. Finney, D. J. 1978. Statistical Method in Biological Assay. 3rd ed. (Pages 425-439) Griffin, London. 508 pp.
7. Gibbs, A. J., and Gower, J. C. 1960. The use of a multiple-transfer method in plant virus transmission studies—Some statistical points arising in the analysis of results. Ann. Appl. Biol. 48:75-83.
8. Griffiths, D. A. 1972. A further note on the probability of disease transmission. Biometrics 28:1133-1139.
9. Kerr, J. D. 1971. The probability of disease transmission. Biometrics 27:219-222.
10. Le, C. T. 1981. A new estimator for infection rates using pools of variable size. Am. J. Epidemiol. 114:132-135.
11. Loyer, M. W. 1983. Bad probability, good statistics, and group testing for binomial estimation. Am. Statistician 37:57-59.
12. Mantel, N. 1975. Group testing with the goal of estimation. Biometrics 31:994-995.
13. Smith, K. M., and Lea, D. E. 1946. The transmission of plant viruses by aphides. Parasitology 37:25-37.
14. Sobel, M., and Elashoff, R. M. 1975. Group testing with a new goal, estimation. Biometrika 62:181-193.
15. Sylvester, E. S. 1954. Aphid transmission of nonpersistent plant viruses with special reference to the *Brassica nigra* virus. Hilgardia 23:53-98.
16. Thompson, K. H. 1962. Estimation of the proportion of vectors in a natural population of insects. Biometrics 18:568-578.
17. Walter, S. D., Hildreth, S. W., and Beaty, B. J. 1980. Estimation of infection rates in populations of organisms using pools of variable size. Am. J. Epidemiol. 112:124-128.
18. Watson, M. A. 1936. Factors affecting the amount of infection obtained by aphis transmission of the virus Hy. III. Philos. Trans. R. Soc. London, Ser. B 226:457-489.