AMERICAN JOURNAL OF EPIDEMIOLOGY
Vol. 114, No. 1
Printed in U.S.A.

A NEW ESTIMATOR FOR INFECTION RATES USING POOLS OF VARIABLE SIZE

CHAP T. LE[1]

Le, C. T. (Mayo Box 197, University of Minnesota, Minneapolis, MN 55455). A new estimator for infection rates using pools of variable size. *Am J Epidemiol* 1981;114:132-6.

A new estimator is proposed for the infection rate in a population of organisms when variably sized sample pools are analyzed. This new estimator has a closed form that can be easily evaluated and updated.

biometry; epidemiologic methods; infection rates

In estimating the infection rates in populations of organisms, it is impossible to assay each organism individually. Instead, the organisms are randomly divided into a number of pools and each pool is tested as a unit. The usefulness of group testing has been extensively studied, and the problem of estimating the infection rate P has been investigated previously by Thompson (1), Chiang and Reeves (2), and Bhattacharyya et al. (3). Recently, Walter et al. (4) considered the general case of pools of variable size and gave the maximum likelihood estimator $\hat{P}$ of P, but a solution to the estimating equation will involve iteration.

In this paper, we adopt the sampling scheme of Walter et al. and propose a new estimator $\tilde{P}$ for P. The method is somewhat similar to the Mantel-Haenszel procedure for combining 2×2 tables and the estimator $\tilde{P}$ has a closed form that can be easily evaluated. A Monte Carlo study demonstrated that the two estimators $\hat{P}$ and $\tilde{P}$ are about equally efficient and accurate.

METHOD AND RESULTS

In the general case of variable pool size, the maximum likelihood estimator $\hat{P}$, as presented by Walter et al. (4), is determined iteratively from the following equation

$$N = \sum_m \frac{mx_m}{1-(1-\hat{P})^m}, \qquad (1)$$

where

m = pool size;
x_m = number of positive pools of size m; and
N = total number of organisms used in all the pools.

In order to circumvent the iteration involved in equation 1, observe that only positive pools are involved in equation 1 and that in practice the sizes of positive pools are typically not highly variable. Hence, the above estimating equation can be written approximately as

$$N \approx \sum_m \frac{mx_m}{1-(1-\hat{P})^{\bar{m}_p}}, \qquad (2)$$

where $\bar{m}_p$ is the average size of positive pools. The new estimator $\tilde{P}$ defined in equation 2, an approximation to the maximum likelihood estimator $\hat{P}$, can now be obtained non-iteratively. Solving equation 2 for $\tilde{P}$, we obtain

Received for publication July 28, 1980, and in final form January 8, 1981.

Abbreviations: RA = relative accuracy; RE = relative efficiency.

[1] Biometry Division, School of Public Health, University of Minnesota, Minneapolis, MN 55455.

Address for reprints: Mayo Box 197, University of Minnesota, Minneapolis, MN 55455.

$$\tilde{P} = 1 - \left(1 - \frac{\sum mx_m}{N}\right)^{1/\bar{m}_p} = 1 - \left(1 - \frac{N_p}{N}\right)^{1/\bar{m}_p}, \quad (3)$$

where N_p is the total number of organisms in all positive pools.

Example

Walter et al. (4) gave the values of $\hat{P}$ for four data sets for studies conducted to determine if certain mosquitoes could transovarially transmit yellow fever virus. For example, for virus strain A with larval development interval of 11–15 days, the data

$$(m, n_m, x_m) = (80,3,1), (100,12,3), (103,1,0), (111,2,1), (115,1,0), (116,1,0), (123,1,0), (150,1,1), (152,1,0),$$

where n_m is the number of pools of size m. In our notation

$$N = 2421$$
$$N_p = 80 + (3)(100) + 111 + 150 = 641$$
$$\bar{m}_p = 641/6 = 106.83.$$

Hence,

$$\tilde{P} = 1 - \left(1 - \frac{641}{2421}\right)^{1/106.83} = 2.8748 \times 10^{-3},$$

while the corresponding maximum likelihood solution $\hat{P}$, found by iteration in Walter et al., is 2.8757×10^{-3}, a difference of only 0.03 per cent. Similar results for other data sets are shown in table 1.

The asymptotic variance of $\tilde{P}$ can be found by the method of error propagation, that is

$$\mathrm{Var}(\tilde{P}) = \sum \left(\frac{\partial \tilde{P}}{\partial x_m}\right)^2 \mathrm{Var}(x_m), \quad (4)$$

where $\partial \tilde{P}/\partial x_m$ is the partial derivative of $\tilde{P}$ with respect to x_m and

$$\mathrm{Var}(x_m) = n_m (1 - P)^m \left[1 - (1 - P)^m\right]. \quad (5)$$

For example, with the same data previously used and employing $\tilde{P}$ for P, the square root of equation 5 yields an estimated standard error for $\tilde{P}$ of 1.1893×10^{-3} compared with 1.1778×10^{-3} obtained by Walter et al. (4). The derivation of $\partial \tilde{P}/\partial x_m$ is given in Appendix 1.

A Monte Carlo study was undertaken to compare the two estimators using ten different pool sizes at six levels of variability in pool size. The pool size of type i is $m_i = 10 + (i - 1)M$; $i = 1, 2, \ldots, 10$. The two estimators are compared in terms of the relative efficiency (RE) and relative accuracy (RA) defined by

$$\mathrm{RE}_P(\hat{P} \text{ to } \tilde{P}) = \frac{\hat{\mathrm{Var}}(\tilde{P})}{\hat{\mathrm{Var}}(\hat{P})} \quad (6)$$

$$\mathrm{RA}_P(\hat{P} \text{ to } \tilde{P}) = \frac{(\text{Average } \tilde{P} - P)^2 + \hat{\mathrm{Var}}(\tilde{P})}{(\text{Average } \hat{P} - P)^2 + \mathrm{Var}(\hat{P})} \quad (7)$$

Results for various levels of infection rate P are shown in table 2, where each entry is the average of 100 independent runs. For example,

$$\mathrm{RE}_P(\hat{P} \text{ to } \tilde{P} \mid M = 5, P = 4 \times 10^{-3}) = 1.02$$
$$\mathrm{RA}_P(\hat{P} \text{ to } \tilde{P} \mid M = 5, P = 4 \times 10^{-3}) = 1.01.$$

TABLE 1

Estimates of transovarial infection rates of yellow fever virus (data and $\hat{P}$ from Walter et al. (4))

Larval development time (days)	Strain A		Strain H	
	$\hat{P} \times 10^3$	$\tilde{P} \times 10^3$	$\hat{P} \times 10^3$	$\tilde{P} \times 10^3$
6–10	0.5338	0.5338	1.4704	1.4698
11–15	2.8757	2.8748	5.8743	5.8556

TABLE 2

Results from the Monte Carlo study comparing $\tilde{P}$ and $\hat{P}$ (in each cell, the estimates are in order $\tilde{P} \times 10^3$, $\hat{P} \times 10^3$, RE, RA)

M	$P = 2 \times 10^{-3}$	4×10^{-3}	6×10^{-3}	8×10^{-3}
5	1.8640	4.9827	6.4656	8.7770
	1.8639	4.9815	6.4627	8.7669
	1.00	1.02	1.01	0.99
	1.00	1.01	1.01	0.99
10	2 5176	4.0648	6.5519	9.1432
	2 5169	4.0618	6.5389	0.0987
	1.02	1.02	0.99	0.96
	1 01	1.01	0.99	0.95
15	2.2398	4.2845	6.5406	8.8044
	2 2388	4.2762	6.5079	8.7209
	1.01	1.00	0.97	0.93
	1.01	0.99	0.95	0.90
20	2.2363	4.6175	6.6839	8.6882
	2.2343	4.5991	6.6243	8.5583
	1.01	0.99	0.95	0.92
	1.01	0.97	0.91	0.86
25	2.0312	4.3240	6.4759	8.0754
	2.0289	4.2974	6.3964	7.9197
	1.00	0.96	0.94	0.90
	1.00	0.94	0.89	0.90
30	1.9985	4.3156	5.8952	8.2155
	1.9951	4.2789	5.8115	7.9945
	1.00	0.94	0.94	0.89
	1.00	0.91	0.97	0.86

DISCUSSION

We have proposed an estimator $\tilde{P}$ of the infection rate P using pools of variable size. The new estimator has a closed form that can be easily evaluated, as shown in the above numerical example. The estimator $\tilde{P}$ can also be obtained by a method which is somewhat similar to the Mantel-Haenszel procedure for combining 2 × 2 tables. If only one pool size m is used throughout, with x_m positive pools out of n_m, the maximum likelihood estimator $\hat{P}_m$ is given by equation 10 of Bhattacharyya et al. (3)

$$\hat{P}_m = 1 - \left(1 - \frac{x_m}{n_m}\right)^{1/m}. \qquad (8)$$

Pools of various sizes are then combined to yield

$$\tilde{P} = 1 - \left(1 - \frac{\sum m x_m}{\sum m n_m}\right)^{\sum x_m / \sum m x_m} = 1 - \left(1 - \frac{N_p}{N}\right)^{1/\bar{m}_p}.$$

Of course, the computation of the asymptotic variance of $\tilde{P}$ using equations 4 and 5 is also complicated. But our simulation results demonstrated that this new estimator and the maximum likelihood estimator $\hat{P}$ are about equally efficient and accurate. In all 24 cases, the differences of the two estimates are no more than 2 per cent and when the difference is large—for either a large P value or large variability within pool sizes—the new estimate tends to be a little more accurate.

Finally, it is interesting to note that our combination method can also be used

TABLE 3

Estimates of transovarial infection rates of yellow fever virus using the Poisson model (Data from Walter et al. (4))

Larval development time (days)	Strain A $\tilde{P}^* \times 10^3$	Strain B $\tilde{P}^* \times 10^3$
6–10	0.5339	1.4709
11–15	2.8790	5.8728

to obtain an estimator of the infection rate using the Poisson model. If only one pool size, m, is used throughout, Bhattacharyya et al. (3) give the following estimator using the method of moments,

$$\hat{P}_m^* = -\frac{\ln[1 - x_m/n_m]}{m}$$

$$= -\ln[1 - x_m/n_m]^{1/m}$$

If now variable pool sizes are used, these pools can be combined by the new method to yield the new estimator

$$\tilde{P}^* = -\ln\left[1 - \frac{\sum m x_m}{\sum m n_m}\right]^{\sum r_m / \sum m r_m}$$

$$= \frac{-\sum x_m}{\sum m x_m} \ln\left[1 - \frac{\sum m x_m}{\sum m n_m}\right]$$

$$= -\frac{\ln\left[1 - \frac{N_p}{N}\right]}{\bar{m}_p}. \qquad (9)$$

For illustration, with the same data set used in the above example of section 2,

$$\tilde{P}^* = \frac{-\ln\left(1 - \frac{641}{2421}\right)}{106.83}$$

$$= 2.8790 \times 10^{-3},$$

with a standard error of 1.1862×10^{-3}. The derivation of $\widehat{\text{Var}}(\tilde{P}^*)$ is given in Appendix 2. Similar results for other data sets are shown in table 3. These values of $\tilde{P}^*$ are very close to those of $\hat{P}$ or $\tilde{P}$ given in table 1. In fact, since the true infection rate P is usually very small, the Poisson distribution could be used as a reasonably good approximation to the binomial distribution.

REFERENCES

1. Thompson KH. Estimation of the proportion of vectors in a natural population of insects. Biometrics 1962;18:568-78.
2. Chiang CL, Reeves WC Statistical estimation of virus infection rates in mosquito vector populations. Am J Hyg 1962;75:377-91.
3. Bhattacharyya GK, Karandinos MG, DeFoliart GR. Point estimates and confidence intervals for infection rates using pooled organisms in epidemiologic studies. Am J Epidemiol 1979; 109:124-31.
4. Walter SD, Hildreth SW, Beaty BJ. Estimation of infection rates in populations of organisms using pools of variable size. Am J Epidemiol 1980;112:124-8.

APPENDIX 1

From equation 3, we can write

$$\tilde{P} = 1 - \left(1 - \frac{\sum m_j x_j}{N}\right)^{\sum r_j / \sum m_j r_j}.$$

Then

$$\frac{\partial \tilde{P}}{\partial x_j} = -(1 - \tilde{P})\left[\frac{\partial A}{\partial x_j} \ln B + A \frac{\partial \ln B}{\partial x_j}\right],$$

where

$$A = \sum x_j / \sum m_j x_j$$

$$\frac{\partial A}{\partial x_j} = \sum_i (m_i - m_j) x_i / \left(\sum_i m_i x_i\right)^2$$

$$B = 1 - \sum m_j x_j / N$$

$$\frac{\partial \ln B}{\partial x_j} = -m_j / NB.$$

APPENDIX 2

Using equation 9,

$$\hat{P}^* = -\frac{\sum x_m}{\sum m x_m} \ln\left[1 - \frac{\sum m x_m}{N}\right]$$
$$= -A \ln B$$

with

$$A = \frac{\sum x_m}{\sum m x_m}$$
$$B = 1 - \sum m x_m / N.$$

The asymptotic variance can be found by the method of error propagation, that is,

$$\mathrm{Var}(\hat{P}^*) = \sum \left(\frac{\partial \hat{P}^*}{\partial x_m}\right)^2 \mathrm{Var}(x_m)$$

where

$$\mathrm{Var}(x_m) = n_m(1 - e^{-mP})\, e^{-mP},$$

and

$$\frac{\partial \hat{P}^*}{\partial x_m} = \left(\frac{\partial A}{\partial x_m} \ln B + A \frac{\partial \ln B}{\partial x_m}\right).$$