STATISTICS IN MEDICINE, VOL. 11, 1591–1597 (1992)

ESTIMATING PREVALENCE BY GROUP TESTING USING GENERALIZED LINEAR MODELS

C. P. FARRINGTON

PHLS Communicable Disease Surveillance Centre, 61 Colindale Avenue, London NW9 5EQ, U.K.

SUMMARY

A method is described for estimating prevalence by group testing using generalized linear models. This provides a simple way of analysing such data using widely available software. Existing methodology to correct for overdispersion using quasi-likelihoods is applied to the group testing model. The methods are illustrated by an estimation of salmonella contamination in eggs, and of yellow fever virus infection in a mosquito population.

1. INTRODUCTION

The problem of estimating prevalence by group testing has received considerable attention in the epidemiological and statistical literature.[1–7] Group testing involves testing pools of experimental units rather than the individual units themselves, a negative result obtained on a pool then signifying that all the units making up the pool are negative. A judicious choice of pool size can lead to substantial reductions in the amount of testing required.

When a single pool size is used the maximum likelihood estimate of prevalence may be expressed in closed form.[4] In practice however the pool size is rarely unique. Existing methodology for estimating prevalence when multiple pool sizes are used is rather cumbersome, requiring iterative solutions to maximum likelihood equations using the Newton–Raphson algorithm.[5,7] This has led at least one author to seek simpler approximate solutions.[6]

The purpose of this paper is to suggest how the problem may be cast in the framework of generalized linear models,[8] thus enabling estimates to be obtained using widely available software such as the statistical package GLIM.[9] In addition to providing easily computed prevalence estimates and confidence intervals, the generalized linear model approach can be used to regress prevalence on covariates, and for model checking. In particular, the model may be adjusted to correct for overdispersion.[10,11] Existing methods to correct for extra-binomial variation may easily be applied to handle pooled data.

The methods are illustrated using published data on the infection rate of yellow fever virus in a mosquito population and are applied to estimate the prevalence of salmonella contamination in eggs. This second application will be described in the next section to motivate the methodology.

2. SALMONELLA CONTAMINATION IN EGGS

Prior to their arrival on the U.K. market, batches of eggs from overseas producers are routinely tested for salmonella contamination. From each batch (typically a lorryload containing a very

0277–6715/92/121591–07$08.50

Received July 1991
Revised April 1992

Table I. Salmonella contamination in eggs from 894 batches

Number of batches	Number of pools per batch	Pool size	Number (proportion) of positive pools
872	—	—	0
7	—	—	$\geqslant 1$
6	10	6	1 (0·10)
2	6	10	1 (0·17)
3	10	6	2 (0·20)
1	6	10	2 (0·33)
1	10	6	3 (0·30)
1	6	10	4 (0·67)
1	6	10	5 (0·83)

(—) Data missing (see text)

large number of eggs) a sample of 60 eggs is collected and divided into groups of equal size. The eggs in each group are pooled and the pools are tested for salmonella contamination. Data from 894 batches are given in Table I. The recommended pool size was 6, although some samples were tested in pools of 10. For samples with no positive pools the pool sizes were not recorded. For seven samples, it was only recorded that one or more pools were positive.

The problem is to estimate the prevalence of salmonella contamination in eggs. The prevalence depends on flock infection levels and hence is likely to vary between batches. The data may thus be expected to exhibit extra-binomial variability, as is indeed suggested by the variation in the proportion of positive pools in the different batches.

The next sections will show how systematic effects and overdispersion can be modelled using generalized linear models. Throughout the paper, the term 'unit' refers to the basic experimental units (in this example the units are eggs).

3. A GENERALIZED LINEAR MODEL FORMULATION

Suppose that for $i = 1, \ldots, k$ a random sample of n_i units is selected from a population i with covariate characteristics $\mathbf{x}_i$ and is tested in r_i pools of size m_i, so that $n_i = m_i r_i$. Typically, covariate characteristics will determine a classification of the units into categories within which the pools are formed.

Let θ_i be the prevalence in population i. Suppose that s_i out of the r_i pools from population i are found positive. Provided that there is no dilution effect,[12] the probability π_i that a pool of size m_i from population i is positive is given by:

$$\pi_i = 1 - (1 - \theta_i)^{m_i}. \tag{1}$$

Thus s_i may be regarded as binomial with index r_i and probability π_i. The implicit assumption that all pools from a given population are of the same size may be relaxed by allowing i to range over combinations of population and pool size. Similarly, different samples may be drawn from the same population.

Let CLL denote the complementary log–log function,

$$\mathrm{CLL}(x) = \log[-\log(1 - x)]. \tag{2}$$

It is easily seen that:

$$\mathrm{CLL}(\pi_i) = \log(m_i) + \mathrm{CLL}(\theta_i) \tag{3}$$

so that if

$$\mathrm{CLL}(\theta_i) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} \tag{4}$$

then a generalized linear model for binomial data, with complementary log-log link function and the vector of fixed constants $\log(m_i)$ (these are offsets in GLIM terminology) provides a suitable representation of the data, the parameters $\boldsymbol{\beta}$ denoting changes in prevalence on the complementary log-log scale.

If the θ_i are very small then the distribution of positive units is approximately Poisson. Thus:

$$\pi_i = 1 - \exp(-\theta_i m_i), \quad \mathrm{CLL}(\pi_i) = \log(m_i) + \log(\theta_i) \tag{5}$$

so that provided $\log(\theta_i) = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$, the contrasts may be interpreted more straightforwardly as logarithmic relative prevalences.

Finally, note that the kernel log-likelihood is:

$$l(\boldsymbol{\theta}; \mathbf{s}) = \sum_{i=0}^{k} s_i \ln(1 - (1-\theta_i)^{m_i}) + m_i(r_i - s_i)\ln(1-\theta_i). \tag{6}$$

Thus the contribution of those samples for which $s_i = 0$ is the same as if all units from these samples had been individually tested, that is as if $m_i = 1$. This provides a statistical foundation for the method of group testing. The point estimates of the parameters and profile likelihood confidence intervals do not depend on the pool sizes used for the negative samples. The Fisher information for $\boldsymbol{\beta}$ is given by

$$I(\boldsymbol{\beta})_{j,k} = \sum_{i=0}^{k} n_i m_i (1-\theta_i)^{m_i}(1-(1-\theta_i)^{m_i})^{-1}\ln^2(1-\theta_i)x_{ij}x_{ik} \tag{7}$$

and the parameter variances therefore depend on m_i, although this dependence is rather weak when $m_i\theta_i \ll 1$.

4. CORRECTING FOR OVERDISPERSION

After correcting for fixed covariate effects, there may still remain variability over and above that consistent with the binomial model, as revealed by a high Pearson chi-square goodness of fit statistic or deviance. In some circumstances this may be due to overdispersion arising from random variation in the θ_i.

Overdispersion is sometimes corrected by rescaling the model.[8] This method has been shown to retain high efficiency for modest overdispersion provided that the maximum likelihood estimates of the θ_i have bias of order n^{-1}.[13] However, this condition does not apply in the present context. Thus for instance for a unique sample and pool size m the bias of the maximum likelihood estimate of θ is given by

$$E(\hat{\theta} - \theta) = (2r)^{-1}(m-1)m^{-2}(1-(1-\theta)^m)(1-\theta)^{1-m} + o(r^{-1}) \tag{8}$$

where r is the number of pools tested. Thus methods more involved than rescaling the model may be required.

A standard way of correcting for overdispersion is to allow the parameters θ_i to vary according to a beta distribution with mean μ_i and variance $\phi\mu_i(1-\mu_i)$. This approach requires rather cumbersome computation and may in some circumstances be unduly restrictive. To retain generality while remaining within the generalized linear model formulation, a natural approach is to use quasi likelihoods.[10,11]

Assume therefore that, conditionally on θ_i, s_i is binomially distributed (r_i, π_i), with π_i defined as in expression (1), and that θ_i is distributed with mean μ_i and variance $\phi\mu_i(1-\mu_i)$. Unconditionally, the following relations hold approximately provided that $m\theta_i \ll 1$:

$$E(s_i) = r_i\tau_i, \qquad v(s_i) = r_i\alpha_i\tau_i(1-\tau_i) \tag{9}$$

where

$$\tau_i = 1-(1-\mu_i)^{m_i}, \qquad \alpha_i = 1+\phi m_i{}^2(r_i-1)\mu_i(1-\mu_i)^{-1}(1-\tau_i)\tau_i{}^{-1} \tag{10}$$

the value $\phi = 0$ corresponding to zero random variability in the θ_i. When $m_i = 1$, these expressions reduce to those of Williams.[10] The model defined in expressions (9) and (10) can be fitted by the method of Bennett[11] using the variance inflation factor α_i in (10). References 10 and 11 provide GLIM listings which may readily be adapted for this purpose. Note that the model defined in (9) and (10) can be fitted whatever the values of the $m_i\theta_i$. When these are much less than 1 the θ_i may be regarded as distributed with mean μ_i and variance $\phi\mu_i(1-\mu_i)$.

This procedure yields quasi-likelihood estimates and standard errors, together with an estimate of the parameter ϕ. The parameter estimates from this method have been shown consistent, and the asymptotic standard errors are unaffected by the fact that ϕ has been estimated from the data.[14] If the m_i and r_i are constant and no covariates are involved then the variance inflation factors α_i are constant and the quasi likelihood approach is equivalent to rescaling the model.

Problems may arise if many of the binomial denominators r_i are equal to 1, in which case convergence may be very slow and the parameter estimates may be biased. In the extreme case when all the r_i are equal to 1, ϕ cannot be estimated by this method. This despite the fact that the data may nevertheless indicate the presence of heterogeneity, since for example a pool of 100 units testing negative suggests a lower prevalence than a pool of 5 testing positive.

When many of the r_i are small, the following procedure may be tried. For all batches with negative pools, that is with $s_i = 0$, replace r_i by $n_i = m_i r_i$ and m_i by 1. The justification for this is that any inferences drawn from the data should be the same as if all units in negative pools had been individually tested and found negative. As discussed in Section 3, the model before correction for overdispersion will not greatly be affected when the $m_i\theta_i \ll 1$, while the bias in the dispersion parameter estimate ϕ may be reduced.

5. YELLOW FEVER VIRUS INFECTION

Walter *et al.*[5] estimated transovarial infection rates in the progeny of an insect population infected by yellow fever virus. The data published in Reference 5 are stratified by virus type (A and H) and by larval developmental interval (6–10 days, and 11–15 days). Using GLIM and fitting the full four-parameter model with factors for virus (level 1: A, level 2: H) and interval (level 1: early; level 2: late developmental interval) with first levels set to zero, gives estimates which, suitably transformed, are virtually identical to those quoted by Walter *et al.*[5] (see Table II 'unadjusted').

The Pearson chi-square statistic for the model was 78·64 on 59 degrees of freedom, suggesting a mild degree of overdispersion. However the great majority of data points (54 of 63) correspond to unique pools, all but 11 of which tested negative. Direct estimation of the dispersion parameter ϕ may consequently yield biased estimates. For the negative pools r_i was therefore replaced by $n_i = m_i r_i$ and m_i by 1. The parameter estimates and standard errors are not substantially affected by these changes (Table II 'adjusted'). The quasi likelihood method described in the previous section produces the estimates corrected for overdispersion shown in Table III; the table also gives the estimates obtained if the iteration is carried out on the unadjusted data.

Table II. Parameterr estimates and standard errors for yellow fever virus infection

	Adjusted data	Unadjusted data
Grand mean	− 7·535 (0·696)	− 7·535 (0·695)
Virus	0·998 (0·757)	0·998 (0·758)
Interval	1·685 (0·803)	1·685 (0·807)
Interaction	− 0·283 (0·901)	− 0·282 (0·914)
Chi square (59 d.f.)	74·84	78·64

Data source: Walter *et al.*[5]

Table III. Parameter estimates and standard errors for yellow fever virus infection after correction for overdispersion

	Adjusted data	Unadjusted data
Grand mean	− 7·270 (0·717)	− 7·437 (0·887)
Virus	0·937 (0·779)	1·395 (0·951)
Interval	1·575 (0·826)	1·548 (1·025)
Interaction	− 0·071 (0·929)	− 0·880 (1·145)
Dispersion ϕ	0·0000611	0·00166

Table IV. Parameter estimates for yellow fever virus infection after model simplification

	Adjusted data corrected for overdispersion*	Unadjusted data not corrected for overdispersion
Grand mean	− 7·227 (0·451)	− 7·375 (0·442)
Virus	0·887 (0·421)	0·806 (0·420)
Interval	1·519 (0·376)	1·465 (0·376)

* dispersion $\phi = 0{\cdot}0000611$ (see Table III)

Finally, the model is simplified by removing the non-significant interaction term, retaining the previously estimated value of ϕ. The estimates for the original model and for the final model (using adjusted data, corrected for overdispersion) are shown in Table IV. The differences between the models are slight, but sufficient to nudge the virus effect to statistical significance at the 5 per cent level. The covariate effects may be summarized as relative risks: the relative risk of infection by virus H compared with virus A is 2·43 (95 per cent confidence limits 1·05, 5·64), and the relative risk of infection for late developing larvae compared with early developing larvae is 4·57 (95 per cent confidence limits 2·15, 9·70).

6. SALMONELLA CONTAMINATION IN EGGS

We now return to the example described in Section 2 (see Table I). No covariate information was collected on the batches, and hence the systematic part of the model is simply the grand mean. It

Table V. Mean prevalence of salmonella contamination standard error and dispersion parameter for different pool sizes for the negative samples

	Pool size in negative samples			
	1	6	10	60
Without correction for heterogeneity				
Prevalence ($\times 10^4$)	6·56	6·56	6·56	6·56
SE($\times 10^4$)	1·11	1·11	1·09	1·12
With correction for heterogeneity				
Dispersion ϕ	0·0180	0·0196	0·0210	3·65*
Prevalence ($\times 10^4$)	8·26	8·02	7·80	1·36*
SE($\times 10^4$)	1·75	1·73	1·71	0·50*

* These estimates are biased (see text)

will be assumed that the recommended pool size of 6 was used for the negative batches; the sensitivity of the results to this assumption will be investigated. For the 7 positive batches with missing pool size, the likelihood is $1 - (1 - \theta_i)^{m_i r_i}$, identical to that obtained with $r_i = s_i = 1$. For these batches it will therefore be assumed that a single pool of 60 was tested and found positive.

Using the model of Section 3 the prevalence was estimated to be 6·56 per ten thousand eggs, with approximate 95 per cent confidence limits of 4·71 and 9.13 per ten thousand. However, the Pearson Chi-square of 2046 on 893 degrees of freedom suggests that the binomial model is inappropriate. Some adjustment for overdispersion is clearly necessary. Under the quasi-likelihood formulation the batch prevalence is regarded as a random variable. Applying the method of Section 4 gives $\phi = 0{\cdot}0196$, the point estimate of prevalence increasing to 8·02 per ten thousand eggs, with approximate 95 per cent confidence limits of 5·25 and 12·24 per ten thousand. The standard deviation of the distribution of batch prevalences is $[\phi\mu(1 - \mu)]^{1/2} = 39{\cdot}6 \times 10^{-4}$ namely 39·6 per ten thousand eggs.

To test sensitivity to the assumption that the negative samples were tested in pools of size 6, the estimates were recalculated assuming that pool sizes of 1, 10 and 60 were used. The results are presented in Table V.

As expected, the prevalence estimate is unaffected, and the standard errors vary little. In addition, for pool sizes of 1, 6 or 10, the correction for overdispersion produces broadly similar results. However, if it is assumed that each negative sample was tested in a single pool of size 60, the correction for overdispersion breaks down, for the reasons outlined in a previous section. Specifically, the variance inflation factor α_i is 1 for the 872 negative batches and > 1 for all but 7 of the 22 positive batches. Thus greater weight is given to the negative batches, biasing the prevalence estimate towards zero and inflating the dispersion parameter ϕ.

In conclusion, the estimated prevalence of salmonella contamination is about 8 per ten thousand eggs, with 95 per cent confidence limits of 5 and 12 per ten thousand. The batch to batch variation in prevalence may be represented approximately by a distribution with mean 8×10^{-4} and standard deviation 40×10^{-4}.

ACKNOWLEDGEMENTS

I thank Dr. De Louvois of the Public Health Laboratory Service who collated the salmonella data and made it available for this analysis, and two referees for very valuable comments.

REFERENCES

1. Thompson, K. H. 'Estimation of the proportion of vectors in a natural population of insects', *Biometrics*, **18**, 568–578 (1962).
2. Chiang, C. L. and Reeves, W. C. 'Statistical estimation of virus infection rates in mosquito vector populations', *American Journal of Hygiene*, **75**, 377–391 (1962).
3. Sobel, M. and Elashoff, R. M. 'Group testing with a new goal, estimation', *Biometrika*, **62**, 181–193 (1975).
4. Bhattacharyya, G. K., Karandinos, M. G. and DeFoliart, G. R. 'Point estimates and confidence intervals for infection rates using pooled organisms in epidemiologic studies', *American Journal of Epidemiology*, **109**, 124–131 (1979).
5. Walter, S. D., Hildreth, S. W. and Beaty, B. J. 'Estimation of infection rates in populations of organisms using pools of variable size', *American Journal of Epidemiology*, **112**, 124–128 (1980).
6. Le, C. T. 'A new estimator for infection rates using pools of variable size', *American Journal of Epidemiology*, **114**, 132–136 (1981).
7. Chen, C. L. and Swallow, W. H. 'Using group testing to estimate a proportion, and to test the binomial model', *Biometrics*, **46**, 1035–1046 (1990).
8. McCullagh, P. and Nelder, J. A. *Generalised Linear Models*, 2nd edn., Chapman and Hall, London, 1991.
9. Payne, C. D. *The GLIM System Manual; Release 3.77*, The Numerical Algorithms Group Ltd., Oxford, 1985.
10. Williams, D. A. 'Extra-binomial variation in logistic linear models', *Applied Statistics*, **31**, 144–148 (1982).
11. Bennett, S. 'An extension of Williams' method for overdispersion models', *GLIM newsletter*, **17**, 12–18 (1989).
12. Hwang, F. K. 'Group testing with a dilution effect', *Biometrika*, **63**, 671–673 (1976).
13. Cox, D. R. 'Some remarks on overdispersion', *Biometrika*, **70**, 269–274 (1983).
14. Moore, D. F. 'Asymptotic properties of moment estimators for overdispersed counts and proportions', *Biometrika*, **73**, 583–588 (1986).