MAXIMUM-LIKELIHOOD ESTIMATION OF THE RELATIVE REMOVAL RATE FROM THE DISTRIBUTION OF THE TOTAL SIZE OF AN INTRA-HOUSEHOLD EPIDEMIC

BY NORMAN T. J. BAILEY

Nuffield Lodge, Regent's Park, London

(1) INTRODUCTION

In a previous paper (Bailey, 1953*a*) I discussed the distribution of the *total* size of a stochastic epidemic, involving both infection and removal, in a given group of homogeneously mixing susceptibles. The model employed was of the 'continuous infection' type, according to which infected individuals continue as sources of infection until removed from circulation by recovery, death or isolation. This may be contrasted with the chain-binomial type of model which entails short periods of high infectivity and approximately constant incubation periods (see, for example, Greenwood, 1931, 1949; Lidwell & Sommerville, 1951; Bailey, 1953*b*). The basic assumptions are that, with x susceptibles and y infectious persons in circulation, the chance of one new infection taking place in time dt is $\beta xy\,dt$, while the chance of a removal is $\gamma y\,dt$, where β and γ are the infection and removal rates, respectively. For a full discussion, the paper referred to (Bailey, 1953*a*) should be consulted. Particular attention was paid to the total size, i.e. when $t \to \infty$, of the epidemic occurring in small groups following the introduction of a single infectious case, the obvious application being to intra-household epidemics. It is important to note that from this ultimate distribution of epidemic size we cannot estimate β and γ separately, though we can estimate the *relative removal rate*, $\rho = \gamma/\beta$. If n is the number of susceptibles in addition to the first case, then for $n = 1$ we have explicit expressions for the maximum-likelihood estimate of ρ and its variance. For $n > 1$ we can resort to the well-known maximum-likelihood scoring procedure. Formulae were given for the probabilities P_w, $0 \leqslant w \leqslant n$, of an epidemic of size w in a group of n susceptibles, not counting the primary case, for the range of values $n = 2$, 3, 4 and 5. Expressions were also given in each case for the maximum-likelihood score for ρ. The actual procedure of estimation in any specific instance is liable to be tedious, especially for $n > 2$. As the score is always a linear function of the observations, it was suggested that it might be worth while tabulating the coefficients of the observational quantities over a suitable range of values of ρ in order to facilitate the calculations. This has now been done for $n = 2$, 3, 4 and 5. The standard theory of this approach is briefly reviewed in the next section, which is then followed by an illustrative example.

(2) MATHEMATICAL NOTE

Let there be a total of N households, containing n susceptibles besides the primary case. Of these a_w, $0 \leqslant w \leqslant n$, produce w additional cases of the disease. Let m_w be

https://doi.org/10.1017/S0022172400027595 Published online by Cambridge University Press

the expectation of a_w, where m_w is a function of ρ. Then the maximum-likelihood score for ρ is

$$S(\rho) = \sum_{w=0}^{n} a_w\left(\frac{1}{m_w}\frac{dm_w}{d\rho}\right) = \sum_{w=0}^{n} a_w S_w, \qquad (1)$$

where

$$S_w = \frac{1}{m_w}\frac{dm_w}{d\rho}. \qquad (2)$$

The score coefficients, S_w, have been calculated on EDSAC at the Cambridge University Mathematical Laboratory for

$$n = 2,\ 3,\ 4 \text{ and } 5, \text{ over the range } \rho = 1{\cdot}00\ (0{\cdot}10)\ 10{\cdot}00.$$

Copies of these tabulations may be obtained from the author. The solution of the maximum-likelihood equation, $S(\rho) = 0$, is then easily effected by calculating $S(\rho)$ for a few trial values of ρ until we have scores of opposite sign for two adjacent values, ρ_1 and ρ_2. Inverse interpolation then gives the required root $\hat{\rho}$. The amount of information, $I(\rho)$, may then be obtained to a fair degree of approximation from the rate of change of the score, i.e.

$$I(\rho) \doteqdot \{S(\rho_1) - S(\rho_2)\}/(\rho_2 - \rho_1), \qquad (3)$$

where $\rho_2 > \rho_1$. For more accurate work, four adjacent values can be used with four-point interpolation. The method recommended by Fisher & Yates (1948, p. 14) may be followed. Having estimated $\hat{\rho}$ the frequencies P_w can then be calculated, for the purpose of examining the goodness-of-fit, from the expressions given in my earlier paper (1935*a*). This is admittedly a trifle laborious and could be made easier by having additional tables, which might need however to be tabulated for finer subdivisions of ρ. On the other hand, the frequencies require to be calculated only once for a given set of data and this procedure is self-checking in that the probabilities must sum to unity. The provision of such tables was accordingly felt to be an unnecessary refinement.

(3) WORKED EXAMPLE

As an illustration of the use of the tables let us consider some data given by Wilson, Bennett, Allen & Worcester (1939) on scarlet fever, a disease involving an extended period of infection, which may be more suitably analysed by the present continuous infection model than by the chain-binomial approach. We have the following data shown in Table 1.

Table 1. *Size of epidemics of scarlet fever in households of three*

No. of secondary cases	No. of households	Expected number
0	172	169·5
1	42	46·0
2	21	19·5
Total	235	235·0

Using the tables for $n = 2$ to calculate the score,

$$S(\rho) = 172 S_0(\rho) + 42 S_1(\rho) + 21 S_2(\rho),$$

https://doi.org/10.1017/S0022172400027595 Published online by Cambridge University Press

at $\rho = 5\cdot 1$ and $5\cdot 2$, we find

$$S(5\cdot 1) = +0\cdot 19168 \text{ and } S(5\cdot 2) = -0\cdot 04645,$$

with an observed amount of information, $I = 2\cdot 3813$. We then easily obtain the maximum-likelihood estimate of ρ by linear inverse interpolation between the scores. Thus

$$\hat{\rho} = 5\cdot 18 \pm 0\cdot 66,$$

where the standard error following the ± sign is given by $I^{-\frac{1}{2}}$. In this case the results obtainable by four-point interpolation make little difference at the level of accuracy required. It can be shown that the estimate of $\hat{\rho}$ is unchanged and that the standard error is reduced to 0·65. The expected numbers have been calculated from the expressions given in my previous paper (1953*a*), and it is clear from Table 1 above that a good fit is obtained, with $\chi^2_{(1)} = 0\cdot 500$. It is interesting to observe that Wilson *et al.* (1939) obtained a significant deviation from expectation when fitting a chain-binomial. Their value of P was about 4 %, and this would have been slightly smaller if they had used maximum-likelihood estimation.

It is a pleasure to acknowledge my indebtedness to Dr J. C. P. Miller of the Cambridge University Mathematical Laboratory, who arranged for the tables described to be computed by EDSAC, and to Miss Margaret O. Lewin, who undertook the actual programming of the work.

REFERENCES

BAILEY, N. T. J. (1953*a*). The total size of a general stochastic epidemic. *Biometrika*, **40**, 177.

BAILEY, N. T. J. (1953*b*). The use of chain-binomials with a variable chance of infection for the analysis of intra-household epidemics. *Biometrika*, **40**, 279.

FISHER, R. A. & YATES, F. (1948). *Statistical Tables* (3rd. ed.). London: Oliver and Boyd.

GREENWOOD, M. (1931). On the statistical measure of infectiousness. *J. Hyg., Camb.*, **31**, 336.

GREENWOOD, M. (1949). The infectiousness of measles. *Biometrika*, **36**, 1.

LIDWELL, O. M. & SOMMERVILLE, T. (1951). Observations on the incidence and distribution of the common cold in a rural community during 1948 and 1949. *J. Hyg., Camb.*, **49**, 365.

WILSON, E. B., BENNETT, C., ALLEN, M. & WORCESTER, J. (1939). Measles and scarlet fever in Providence, R.I., 1929–34 with respect to age and size of family. *Proc. Amer. phil. Soc.* **80**, 357.

(*MS. received for publication* 9. XII. 53)

https://doi.org/10.1017/S0022172400027595 Published online by Cambridge University Press