The Total Size of a General Stochastic Epidemic

Norman T. J. Bailey


Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28195306%2940%3A1%2F2%3C177%3ATTSOAG%3E2.0.CO%3B2-E

*Biometrika* is currently published by Biometrika Trust.
THE TOTAL SIZE OF A GENERAL STOCHASTIC EPIDEMIC

By NORMAN T. J. BAILEY

Nuffield Lodge, Regent's Park, London

1. Introduction

The early work on the mathematical theory of epidemics (e.g. Ross,* 1916 and later; Brownlee, 1918; Kermack & McKendrick, 1927 and later; Soper, 1929) was invariably of a 'deterministic' nature, and assumed that for given numbers of susceptible and infectious individuals, and given attack and removal rates, a certain definite number of fresh cases would occur in any specified time. However, it is widely realized that an appreciable element of chance enters into the conditions under which new infections or removals take place. The probability approach was fundamental to Greenwood's (1931, 1946) use of chain binomials in discussing the distribution of multiple cases of disease in households. Recent discussions of these problems (e.g. Bartlett, 1946, 1949; Bailey, 1950) have therefore turned to 'stochastic' models. In these we have, for any given instant of time, probability distributions for the total numbers of susceptible and infected individuals replacing the single point-values of the deterministic treatments. Stochastic models have a special importance in this context due to the fact that for epidemic processes stochastic means are not the same as the corresponding deterministic values. Although for large homogeneously mixing groups deterministic methods might be fairly adequate, it seems likely that in practice epidemics actually occur in several relatively small groups of friends and acquaintances, the epidemiological returns for an administrative unit being compounded of many such comparatively distinct processes. Moreover, when we are considering the distribution of cases of a disease in a household the size of the group is always so small as to demand a stochastic treatment.

In my 1950 paper I discussed a simple stochastic epidemic where none of the infected individuals was removed from circulation by death, recovery or isolation. This might well apply to some of the milder infections of the upper respiratory tract, and can also be used approximately to represent epidemics for which the time taken for removal from circulation is long compared with the time usually required for the epidemic to be completed. The present paper considers the more general problem of allowing for both infection and removal. The analytical difficulties present in the treatment of the simple epidemic appear here in a more acute form, though it has proved possible to compute the frequency distribution of the total size of the epidemic for moderate group size given the ratio of removal to infection rate. The results obtained may be compared with those described by Kermack & McKendrick (1927) in the deterministic case. No obvious analogue to the threshold theorem can be discerned in the stochastic models. An important application of these results is to the problem of the distribution of multiple cases of disease in a household, and a method is given for obtaining maximum-likelihood estimates of the ratio of the removal to infection rate.

* Although Ross started with the idea of probability his mathematical theory is essentially deterministic.

Biometrika 40
The total size of a general stochastic epidemic

2. Deterministic treatment

We must first glance briefly at the results obtained in the deterministic case. The following treatment, with constant infection and removal rates, is substantially that given by Kermack & McKendrick (1927), though with some slight alterations to their notation.

Consider a homogeneously mixing community of \( n \) individuals, of whom at time \( t \) there are \( x \) susceptibles, \( y \) infectious cases in circulation and \( z \) individuals who are isolated, dead, or recovered and immune. Thus we have

\[
x + y + z = n.
\]

Now suppose that there is a constant infection rate \( \beta \) and a constant removal rate \( \gamma \), so that the number of new infections in time \( dt \) is \( \beta xy \) and the number of removals from circulation is \( \gamma y dt \). We can choose our time scale so that \( t \) is replaced by \( \beta t \). Then it is easy to see that the course of the epidemic is represented by the differential equations

\[
\begin{align*}
\frac{dx}{dt} &= -xy, \\
\frac{dy}{dt} &= xy - py, \\
\frac{dz}{dt} &= py,
\end{align*}
\]

where \( p = \gamma/\beta \), the ratio of the removal to infection rate. Initially, when \( t = 0 \), we can assume that \( x \) is approximately equal to \( n \). It is then clear from (1) that unless \( p < n \) no epidemic can start to build up as this requires \( dy/dt > 0 \). Kermack & McKendrick obtained an approximate solution to (1) for epidemics of small magnitude and showed that, if \( p = n - \nu \), where \( \nu \) is small compared with \( n \), an epidemic of total size \( 2\nu \) will occur. This constitutes what may be called Kermack & McKendrick's Threshold Theorem, and can be interpreted by saying that if the initial density of susceptibles is \( n = p + \nu \) then the introduction of a few infected persons will give rise to an epidemic, after which the density of susceptibles is reduced to \( p - \nu \), a value as far below the threshold \( p \) as originally it was above it. Somewhat similar results can be obtained for the more general case of variable infection and removal rates, and an extension can be made to the situation where an intermediate host is involved (Kermack & McKendrick, 1927).

3. Stochastic treatment

Let us now consider the stochastic analogue of the deterministic treatment discussed in the previous section. We shall use the same definitions of \( x \), \( y \) and \( z \), and shall replace \( t \) by \( \beta t \) as before. Then on the assumption of homogeneous mixing of the susceptibles and infectious individuals in circulation the probability of one new infection taking place in time \( dt \) is \( xy dt \), while the probability of one infected person being removed from circulation in time \( dt \) is \( py dt \). Let \( p_{rs}(t) \) be the probability that at time \( t \) there are \( r \) susceptibles still uninfected and \( s \) infectious individuals in circulation. Let us assume that the epidemic is started by the introduction of \( a \) infectious cases into a population of \( n \) susceptibles. It is now easy to show
by the usual methods that the whole process can be characterized by the partial differential equation for the probability generating function $\Pi$:

$$\frac{\partial \Pi}{\partial t} = (v^2 - uv) \frac{\partial^2 \Pi}{\partial u \partial v} + \rho(1-v) \frac{\partial \Pi}{\partial v},$$  
(2)

where

$$\Pi = \sum_{r,s} u^r v^s p_{rs},$$  
(3)

with limits

$$0 \leq r + s < n + a, \quad 0 \leq r \leq n, \quad 0 \leq s \leq n + a.$$  
(4)

Equation (2) is substantially that given by Bartlett (1949, equation (49)), putting his immigration rate equal to zero.

Let us now use the Laplace transform and its inverse with respect to time given by

$$\phi^*(\lambda) = \int_0^\infty e^{-\lambda t} \phi(t) \, dt, \quad R(\lambda) > 0,$$

$$\phi(t) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} e^{\lambda t} \phi^*(\lambda) \, d\lambda,$$

where $\int_{c-i\infty}^{c+i\infty} = \lim_{w \to \infty} \int_{c-iw}^{c+iw}$, and $c$ is positive and greater than the abscissae of all the residues. Taking transforms of (2) and (3), and using the boundary condition

$$p_{na}(0) = 1,$$

we obtain

$$(v^2 - uv) \frac{\partial^2 \Pi^*}{\partial u \partial v} + \rho(1-v) \frac{\partial \Pi^*}{\partial v} - \lambda \Pi^* + w^a \rho^a = 0,$$  
(7)

and

$$\Pi^* = \sum_{r,s} u^r v^s p_{rs}^* = \sum_{r,s} u^r v^s q_{rs},$$  
(8)

where

$$q_{rs} = p_{rs}^* = \int_0^\infty e^{-\lambda t} p_{rs}(t) \, dt.$$  
(9)

Substituting (8) in (7), and equating coefficients of $w^a \rho^a$, yields the recurrence relations

$$(r + 1)(s - 1) q_{r+1,s-1} - (s(r + \rho) + \lambda) q_{rs} + \rho(s + 1) q_{r,s+1} = 0,$$

and

$$- (a(n + \rho) + \lambda) q_{na} + 1 = 0,$$

with

$$0 \leq r + s < n + a, \quad 0 \leq r \leq n, \quad 0 \leq s \leq n + a.$$  
(10)

Any $q_{rs}$ whose suffix falls outside the prescribed ranges is taken to be identically zero. It is evident from the form of the equations that, starting with $q_{na}$, all the quantities $q_{rs}$ could be calculated successively. Using the inverse of the Laplace transformation, we could then arrive at the required $p_{rs}$, exhibiting them as sums of exponential terms like $e^{-i(j + \rho)t}$. There seems to be considerable difficulty in handling such expressions in a compact and convenient way to give, for example, epidemic completion times or the stochastic epidemic curve showing the rate of change with respect to time of the average total number of removals at any instant. However, some progress is possible if we concentrate attention on the total size of the epidemic, i.e. the value of $n + a - x$ for $t = \infty$. As $t \to \infty$ all terms in the expansion of $p_{rs}(t)$ involving negative exponentials like $e^{-i(j + \rho)t}$ vanish unless $i = 0$. The non-vanishing term is the coefficient of $\lambda^{-1}$ in the partial fraction expansion of $q_{rs}$ in terms of $i(j + \rho) + \lambda)^{-1}$.  

---

*Norman T. J. Bailey*
The total size of a general stochastic epidemic

Now the epidemic ceases to spread to fresh susceptibles as soon as \( s = 0 \). Thus the probability of an epidemic of total size \( w \) (not counting the initial \( a \) infectious persons) is

\[
P_w = \lim_{t \to \infty} p_{n-w,0}(t) \quad (0 \leq w \leq n),
\]

\[
= \lim_{\lambda \to 0} \lambda q_{n-w,0}
\]

\[
= \lim_{\lambda \to 0} \rho q_{n-w,1}, \quad \text{putting} \quad r = n - w \quad \text{and} \quad s = 0 \quad \text{in (10)},
\]

\[
= \rho f_{n-w,1},
\]

where

\[
f_{rs} = \lim_{\lambda \to 0} q_{rs}, \quad \begin{cases} 1 \leq r + s \leq n + a, & 0 \leq r \leq n, \quad 1 \leq s \leq n + a. \end{cases}
\]

The quantities \( f_{rs} \) evidently satisfy the following recurrence relations obtained from (10) by writing \( f_{rs} \) for \( q_{rs} \) and putting \( \lambda = 0 \),

\[
(r + 1)(s - 1)f_{r+1,s-1} - s(r + \rho)f_{r+1,s} + \rho(s + 1)f_{r,s+1} = 0
\]

and

\[
-a(n + \rho)f_{n+1,a+1} = 0,
\]

with same limits as in (12).

Some further simplification results from writing

\[
f_{rs} = \frac{n!(r + \rho - 1)!\rho^{a+s-r-s}}{sr!(n + \rho)!} g_{rs}.
\]

Substituting in (13) gives

\[
g_{r+1,s-1} - g_{rs} + (r + \rho)^{-1} g_{r,s+1} = 0
\]

and

\[
g_{na} = 1.
\]

I am indebted to Dr F. G. Foster for suggesting to me the alternative approach of considering the succession of population states represented by the points \( (r, s) \). Thus the progress of the epidemic can be regarded as a random walk from the point \( (n, a) \) to the points \( (n - w, 0) \) \( w = 0, 1, \ldots, n \), with an absorbing barrier at \( r = 0 \), and where the possible transitions from \( (r, s) \) are

\[
(r, s) \to (r - 1, s + 1), \quad \text{occurring with probability} \quad r/(r + \rho),
\]

and

\[
(r, s) \to (r, s - 1), \quad \text{occurring with probability} \quad \rho/(r + \rho).
\]

Foster’s general formula for \( P_w \) can now be written down almost immediately simply by considering the sum of the probabilities of all possible paths from \( (n, a) \) to \( (n - w, 0) \). Thus we have

\[
P_w = \frac{\rho^{a+w}}{\rho + n - w} \sum_{\boldsymbol{x}} \binom{n}{w} \frac{1}{(n + \rho)^w} \prod_{i=1}^{w} (\rho + n - i)^{-\alpha_i} (\rho + n - w)^{-\alpha_w},
\]

where the summation is over all compositions of \( a + w - 1 \) into \( w + 1 \) parts such that \( 0 \leq \alpha_i \leq a + i - 1 \) for \( 0 \leq i \leq w - 1 \) and \( 1 \leq \alpha_w \leq a + w - 1 \). However, for the purposes of computation there appears to be some advantage, especially if \( n \) is at all large, in calculating the quantities \( P_w \) from (11), (14) and (15), instead of from (16). The reason for this is that not only is the form of (16) not very suitable for computation, but also the partitional nature of
the summation may leave some doubt as to whether all relevant terms have been included in any specific instance. Using (11), (14) and (15), therefore, the $P_w$ have been calculated over a suitable range of values of $\rho$, for $n = 10$, 20 and 40, and taking $\alpha = 1$ as a standard initial condition. Some typical results are shown in Figs. 1, 2 and 3. It can be seen from Figs. 1, 2 and 3 that when the relative removal rate $\rho$ is large epidemics tend to be small, and

---

**Fig. 1.** Diagram showing the probability of the final total size of the epidemic for groups of ten susceptibles, starting with the introduction of one new infectious case.

**Fig. 2.** Diagram showing the probability of the final total size of the epidemic for groups of twenty susceptibles, starting with the introduction of one new infectious case.

**Fig. 3.** Diagram showing the probability of the final total size of the epidemic for groups of forty susceptibles, starting with the introduction of one new infectious case.
The total size of a general stochastic epidemic

conversely. There is a fairly gradual transition between the two extremes, though for some intermediate values of $\rho$ most of the probability is accounted for by the two ends of the distribution. For example, with $\rho = 5$ for $n = 20$, there is a 20% chance of no additional cases and a 64% chance of 19 or 20. Again, there is only a gradual drop in the average size of the epidemic with increasing $\rho$. Specimen values for the range $\rho = 0 \rightarrow 25n \rightarrow 50n$ are set out in Table 1 below. There is no obvious analogue of the Threshold Theorem derived by Kermack & McKendrick (1927) for the deterministic case.

Table 1. Average total size of epidemic for various values of $\rho$ and $n$

<table>
<thead>
<tr>
<th>$\rho$</th>
<th>$n = 10$</th>
<th>$n = 20$</th>
<th>$n = 40$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10:00</td>
<td>20:00</td>
<td>40:00</td>
</tr>
<tr>
<td>0:25n</td>
<td>7:13</td>
<td>14:40</td>
<td>29:12</td>
</tr>
<tr>
<td>0:50n</td>
<td>4:33</td>
<td>7:97</td>
<td>15:31</td>
</tr>
<tr>
<td>0:75n</td>
<td>2:74</td>
<td>4:32</td>
<td>6:94</td>
</tr>
<tr>
<td>1:00n</td>
<td>1:89</td>
<td>2:62</td>
<td>3:58</td>
</tr>
<tr>
<td>1:25n</td>
<td>1:38</td>
<td>1:77</td>
<td>2:18</td>
</tr>
<tr>
<td>1:50n</td>
<td>1:08</td>
<td>1:30</td>
<td>1:50</td>
</tr>
</tbody>
</table>

The methods of the foregoing section may also be employed to investigate the distribution of multiple cases of a disease in households. This problem was first examined statistically by Greenwood (1931) who considered the hypothesis that, with a fairly infectious disease like measles, the first case in a family would arise from an outside contact while subsequent cases would occur through contacts within the family. The period of infectiousness is thought to be short for measles and if reduced, for the purpose of simplification, to an instant, Greenwood showed that the course of the intra-familial epidemic may be represented by a chain of binomial distributions. The frequencies of the final number of cases observed can then be found in terms of a parameter $\rho$, which is a measure of infectiousness. Such a model is quite adequate for measles, and satisfactory tests of goodness-of-fit were obtained for the data available, merely by equating observed and expected means. However, for diseases like diphtheria which have a more extended period of infectiousness there is probably some advantage in using the concepts of infection and removal employed in this paper. Equations (11), (14) and (15) can be used as before to calculate the quantities $P_{rn}$ for small values of $n$, e.g. 1 to 5, keeping in $\rho$ as a parameter and taking $a$ to be unity. There is some advantage, for simplicity in handling the algebra, in partially solving the recurrence relation in (15) to give $g_{rn}$ as a linear function of $g_{r+1,i}$, $i = (s - 1), ..., (n - r)$. The requisite formulae are easily found to be

$$g_{rn} = \sum_{i=s-1}^{n-r} (r + \rho)^{s-i-1} g_{r+1,i} \quad (s > 1),$$

with

$$g_{r1} = g_{r2}/(r + \rho),$$

and

$$g_{n1} = 1.$$
<table>
<thead>
<tr>
<th>n = 1:</th>
<th>$P_0 = \rho/(\rho + 1)$</th>
<th>$\hat{\rho} = a_0/a_1$</th>
<th>$I_\rho = N/\rho(\rho + 1)^2$</th>
<th>$P_1 = 1/(\rho + 1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>n = 2:</td>
<td>$P_0 = \rho/(\rho + 2)$</td>
<td>$\hat{\rho} = a_0/a_1$</td>
<td>$I_\rho = N/\rho(\rho + 1)^2$</td>
<td>$P_1 = 2(2\rho + 1)/(\rho + 2) (\rho + i)^2$</td>
</tr>
<tr>
<td>n = 3:</td>
<td>$P_0 = \rho/(\rho + 3)$</td>
<td>$\hat{\rho} = a_0/a_1$</td>
<td>$I_\rho = N/\rho(\rho + 1)^2$</td>
<td>$P_1 = 3\rho^3/\rho(\rho + 3) (\rho + 2)^2$</td>
</tr>
<tr>
<td>n = 4:</td>
<td>$P_0 = \rho/(\rho + 4)$</td>
<td>$\hat{\rho} = a_0/a_1$</td>
<td>$I_\rho = N/\rho(\rho + 1)^2$</td>
<td>$P_1 = 4\rho^4/(\rho + 4) (\rho + 3)^2$</td>
</tr>
<tr>
<td>n = 5:</td>
<td>$P_0 = \rho/(\rho + 5)$</td>
<td>$\hat{\rho} = a_0/a_1$</td>
<td>$I_\rho = N/\rho(\rho + 1)^2$</td>
<td>$P_1 = 5\rho^5/(\rho + 5) (\rho + 4)^2$</td>
</tr>
</tbody>
</table>
The total size of a general stochastic epidemic

On the other hand, with values of $n$ as small as 1 to 5, it is probably just as easy to derive the $P_n$ straight from Foster's formula (16).

We cannot expect to be able to estimate $\beta$ and $\gamma$ separately as the asymptotic distribution of epidemic size for infinite time yields no information about the time scale. For this we should require data giving the time intervals between successive infections in families with two or more cases.

Having calculated the $P_n$ for any given family size, we can throw the results into a form suitable for the maximum likelihood estimation of $\rho$. $N$ is the total number of families of a given size; and $a_w$ is the observed number of families with a total of $w$ cases in addition to the first one. The case $n = 0$ is trivial. For $n = 1$ there are simple expressions for the amount of information as well as the maximum likelihood estimate, while for $n \geq 2$, the information functions become increasingly awkward to handle, and the simplest procedure is to use the well-known method of calculating the observed amount of information from sufficiently close values of the score. The values of the $P_n$ and the corresponding score for $n = 1, 2, 3, 4$ and 5, are set out below in Table 2.

The values of $P_n$ can be conveniently checked by ensuring that their sum for any given $n$ is unity. Although the scores for the larger $n$ contain some awkward looking polynomials in $\rho$, there is little difficulty in practice, with the aid of Barlow's Tables and a calculating machine, in computing the score at a few trial values of $\rho$ for the purposes of inverse interpolation. However, if such methods were to be used at all extensively it would be worth while considering the construction of special tables. The scores are linear functions of the observations and the coefficients of these observational quantities could be tabulated over a wide range of values of $\rho$.

Suitable data for the application of the above methods do not seem to be available, or at any rate are not readily accessible, apart from the material on the 1926 St Pancras measles epidemic used by Greenwood (1931). As already mentioned, Greenwood found that the chain-binomial model gave a satisfactory fit and so we should hardly expect the present model to give a very adequate description for measles, though it might for other diseases. This is in fact the case. I have carried out the maximum likelihood estimation of $\rho$ on Greenwood's data for families up to total size 5, and in no case is a satisfactory fit obtained. It is therefore hardly worth giving the details of the calculations. The epidemiological implications are, however, important, for it has thus now been shown that in the case of measles a satisfactory fit is to be obtained neither by the simple binomial distribution to be expected if the disease were not highly infectious within families (Greenwood, 1931) nor by the distribution expected when there is both infection and removal of the types discussed in the present paper. On the other hand, the chain binomial model used by Greenwood (1931), appropriate to very short periods of high infectivity, is adequate.

5. Summary and conclusions

An investigation has been made of the total size, i.e. for infinite time, of a general stochastic epidemic involving both infection and removal by recovery, death or isolation. For small homogeneously mixing groups no analogue has been found of the Threshold Theorem derived by Kermack & McKendrick (1927) for the deterministic case. In stochastic models wide variations in the size of an epidemic can occur purely by chance with fixed infection and removal rates. This may have important consequences for the interpretation
of epidemiological data. An application to the problem of the distribution of multiple cases of disease in a household is also considered. It is shown how maximum likelihood estimates of the ratio of removal to infection rate can be obtained from suitable data, and the appropriate maximum likelihood scores are given for families up to a total size of 5 (not including the first case). The model under discussion is not suitable for diseases like measles involving short periods of high infectivity, but its adequacy for other infections requires to be tested.

I am indebted to Miss Eva Rowland for undertaking the computations, on which Figs. 1, 2 and 3 and Table 1 were based.

REFERENCES


