ELSEVIER

A Test of Homogeneity Versus a Specified Heterogeneity in an Epidemic Model

TOM BRITTON
Mathematical Statistics, Department of Mathematics, Stockholm University, S-106 91 Stockholm, Sweden

Received 8 February 1996; revised 30 August 1996

ABSTRACT

A two-parameter epidemic model allowing any specified heterogeneous contact structure is studied. With the use of recursive formulas for the final size distribution, as in (Addy, Longini and Haber, Biometrics, 47:961–974, 1991), the score test of the hypothesis that the heterogeneous structure is nonsignificant is derived. The test may be used if there is uncertainty about the spreading mechanism of an infectious disease and a known heterogeneous structure, such as geographical or social structure or both, would be apparent if the disease spread through person-to-person contacts. © Elsevier Science Inc., 1997

1. INTRODUCTION

During the last ten years, much of the research in epidemic theory was aimed at incorporating heterogeneous structures in the population and finding out how this heterogeneity affects certain properties of the model, such as the basic reproduction number and the probability for a major epidemic outbreak. For a fixed and finite population, a technique has been developed for computing the distribution of the final outcome for such models, a method based on recursive formulas [1–4].

In the present paper, these results will be used to construct a simple test of the hypothesis that the population mixes homogeneously, where the alternative is any specified heterogeneous contact structure. In a human population, such prespecified structures could, for example, be families, geographic structure, social structure, or a mixture of them. If the population consists of animals or plants, one might consider the "nearest neighbors" structure (the model assumes that animals/plants infect their neighbors at a higher rate).

If the disease we are interested in is known to spread through person-to-person contacts, then the hypothesis of homogeneity seems very unlikely—so the test should certainly reject the hypothesis and

655 Avenue of the Americas, New York, NY 10010
0025-5564/97/$17.00
PII S0025-5564(96)00170-8

hence not be very informative. However, for several infectious diseases, the exact mechanism with which it is spread is not known. Common alternatives to diseases spreading through person-to-person contacts are, for example, airborne diseases and diseases spreading through a common water source. The test presented in this paper can therefore be interpreted as a test to see whether a disease spreads through person-to-person contacts. It is only when this hypothesis is rejected that one might move on to estimation of certain parameters in the model. An investigation of procedures for estimating these parameters would be far from simple under the general alternative hypothesis treated in this paper.

The data needed for the test is a binary variable for each individual indicating whether or not he was infected at the end of the epidemic. The score statistic Equation (10) or its approximation Equation (11) is a linear combination of pairwise products of these variables and is thus straightforward to compute [at least Eq. (11)]. Because we are testing for homogeneity against heterogeneity, the total number of infected individuals carries no information in itself about the hypothesis. The information on which our test is based is *who* is infected given the total number infected. We will condition on this number when we derive the distribution of the score statistic, the total number being sufficient for a nuisance parameter under the hypothesis of homogeneity.

This is the plan for the rest of the paper. In Section 2, the model is defined, the recursive formula is presented, using the results of Addy et al. [1], and the hypothesis is stated. The score statistic of the hypothesis and the test procedure are given in Section 3. Unless the population is small, certain difficulties appear in the test procedure, which are treated in Section 4. Simulations of epidemics are presented in Section 5, to convey a sense of how severe the heterogeneous structure must be to be detected. Finally, in Section 6, the paper at hand is related to similar work found in the literature, with a discussion of possible extensions of the present work.

2. THE MODEL AND HYPOTHESIS

The epidemic model is a Susceptible, Infective and Removed (SIR) model with general distribution for the infectious periods and with arbitrary contact rates between each pair of individuals. Consider a closed population consisting of n individuals, labeled $1,\ldots,n$, in which an infectious disease is spread. At the start of the epidemic, we assume one individual, say u, has just become infected from some external source, and everyone else is susceptible. Individuals who are infected become infective immediately (no latent period) and remain so for a

time duration with distribution F. During the infectious period, individual i makes contact with other individuals according to independent Poisson processes; at rate $\lambda_{i,j}$, he makes contact with individual j, $j \neq i$ (see the end of the section for an interpretation of $\lambda_{i,j}$). If a contact is made with a susceptible individual, the individual becomes infected and infective; otherwise nothing happens. When i's infectious period has ended, he becomes immune and plays no further role in the epidemic—we say that i has been removed. All contact processes and infectious periods are defined to be mutually independent. The epidemic evolves until the first time at which there are no infective individuals in the population; when this happens, no one spreads the disease and no one can get infected. The epidemic stops, and the state of the population at this time is called the final outcome (i.e., which individuals have been removed and which are still susceptible).

Note that an i-to-j infection is possible only when i is infective and j is susceptible. At such times, there is no "opposite" contact process, so $\lambda_{j,i}$ need not equal $\lambda_{i,j}$ for the construction to be consistent. Thus the parameter $\lambda_{i,j}$ is the rate of transmission between i and j when i is infective and j is susceptible. Let $\Lambda = \{\lambda_{i,j}\}$ denote the contact matrix where we define $\lambda_{i,i} := 0$ for all i, because it is assumed that self-infection is impossible.

The distribution of the final outcome for this epidemic model has been derived by Addy et al. [1] and is given by recursive formulas that are presented below. First, we need some more notation. All vectors in this paper will be binary and have dimension n. The ith component of a vector $\mathbf{k}$ is denoted k_i, and $|\mathbf{k}| := \sum_i k_i$, the number of 1's. All matrices will be $n \times n$ with 0's on the diagonal unless otherwise stated. Let ϕ be the Laplace transform of the infectious period, $\phi(a) := \int_0^\infty e^{-ax}\, dF(x)$, and for vectors $\mathbf{a}$ and $\mathbf{k}$ write $\Phi(\mathbf{a})^{\mathbf{k}} := \prod_{i=1}^n \phi(a_i)^{k_i}$. Further, we write $\sum_{\mathbf{j}=\mathbf{0}}^{\mathbf{k}}$ for $\sum_{j_1=0}^{k_1} \cdots \sum_{j_n=0}^{k_n}$.

Suppose $\mathbf{k}^u$ is a binary vector such that the uth component, k_u^u, equals 0, and let $P^u(\mathbf{k}^u; \Lambda)$ denote the probability that exactly those individuals i with $k_i^u = 1$ were ultimately infected (besides individual u). Then these probabilities may be computed through the recursive formula

$$\sum_{\mathbf{j}=\mathbf{0}}^{\mathbf{k}^u} P^u(\mathbf{j}; \Lambda)\Phi\left[\Lambda(\mathbf{1} - \mathbf{e}_u - \mathbf{k}^u)\right]^{-(\mathbf{j}+\mathbf{e}_u)} = 1, \qquad \text{for } \mathbf{0} \leqslant \mathbf{k}^u \leqslant \mathbf{1} - \mathbf{e}_u, \quad (1)$$

where $\mathbf{e}_u$ is the uth unit vector and $\mathbf{1}$ is the vector with 1 in each component.

In this paper, I wish to make some inference on the contact matrix Λ after having observed the final outcome of an epidemic. For a general

matrix, there are $n(n-1)$ parameters and an n-dimensional vector is observed, so the number of parameters have to be reduced to have a well-posed statistical problem. What we do is to assume Λ to be of the form $\lambda_{i,i} = 0$ and, for $i \neq j$,

$$\lambda_{i,j} = \frac{\lambda}{n} + \delta c_{i,j} \tag{2}$$

where $C = \{c_{i,j}\}$ is some prespecified matrix with nonnegative elements and $c_{i,i} = 0$. (The factor $1/n$ has been inserted to simplify notation and for asymptotic reasons.) We have thus reduced the number of parameters to only two, λ and δ. Within this model, we will construct a test of the hypothesis

$$H_0\colon \delta = 0, \text{ against the alternative } H_A\colon \delta > 0, \tag{3}$$

which means that we have a homogeneous population that mixes uniformly instead of the alternative $\delta > 0$ when $i \to j$ contacts with large $c_{i,j}$ are more likely.

The model contains the unknown nuisance parameter λ. Heuristically, this implies that the total number infected carries no information about δ. For this reason, we will condition on the observed number infected and base our test on *who* were infected, given this number. The formal statistical reason for conditioning on the total number infected is that this number is sufficient for λ under the null hypothesis.

The matrix C should be interpreted as some measure of "closeness." An example is $c_{i,j} = 1$ if i and j belong to the same family and $c_{i,j} = 0$ otherwise. Another example is $c_{i,j} = d(i,j)^{-1}$, where $d(\cdot,\cdot)$ is some geographic or social distance or both, with the interpretation that "close" individuals infect each other at a higher rate. In both of these examples, the matrix C, and hence Λ, is symmetric, but this must not always be the case. In fact, $\lambda_{i,j}$ should be thought of as the product of three factors $\lambda_{i,j} = \kappa_{i,j}\iota_i\sigma_j$, where $\kappa_{i,j}$ is the contact rate between i and j (naturally symmetric), ι_i is i's infectivity, and σ_j measures how susceptible j is. So, for example, if i is an adult and j is a child, we might have $\lambda_{i,j} > \lambda_{j,i}$ owing to the fact that children more easily catch a cold than adults do. This is in fact one question that we will be able to test by using the results of the next section.

3. TESTING FOR HOMOGENEITY

In this section, we will derive a test of the hypothesis stated in Equation (3). Because no uniformly most powerful test exists, we will derive the score test, the test that maximizes power for alternatives

close to the null hypothesis (i.e., small δ). In applications, it is rarely known which individual(s) started an epidemic. We will assume this to be unknown and act as if the starting individual were randomized; and, because all individuals are equally susceptible under the null hypothesis, we will randomize according to the uniform distribution. (A similar analysis can be performed if the initial infective is known or if some other distribution of the initial infective is preferred—this situation is treated in the remark after Theorem 3.1.)

Suppose we observe the final-state vector $\mathbf{k}$; that is, that exactly those individuals i with $k_i = 1$, were infected. Then the score test should be based on

$$\frac{\partial}{\partial \delta} \log \mathrm{P}(\mathbf{k}; \lambda, \delta)\Big|_{\delta = 0}, \tag{4}$$

where $\mathrm{P}(\mathbf{k}; \lambda, \delta)$ is the probability that $\mathbf{k}$ is the final outcome if the initial infective is drawn completely at random from the whole population. This means that $\mathrm{P}(\mathbf{k}; \lambda, \delta) = n^{-1} \Sigma_{u; \mathbf{k}_u = 1} \mathrm{P}^u(\mathbf{k} - \mathbf{e}_u; \lambda, \delta)$, using the notation in Equation (1) and associating λ, δ with the matrix Λ defined in Equation (2). As mentioned earlier, we will condition on the observed number of infected, $|\mathbf{k}|$, so Expression (4) should actually contain $-\frac{\partial}{\partial \delta} \log \mathrm{P}(|\mathbf{k}|; \lambda, \delta)|_{\delta = 0}$. This factor will depend only on data through $|\mathbf{k}|$, which is why it is omitted.

From the recursive formula in Equation (1), it is possible to show the following theorem. Its proof, which is given in Appendix 1, is somewhat technical although not deep. The main ingredients are basic combinatorics and changing the order of summation.

THEOREM 3.1

The log-derivative for the model defined in Section 2 is

$$\frac{\partial}{\partial \delta} \log P(\mathbf{k}; \lambda, \delta)\Big|_{\delta = 0} = \frac{\alpha_k}{k} \left(\mathbf{k}^T C \mathbf{k} - \beta_k \mathbf{k}^T C \mathbf{1} \right), \tag{5}$$

where α_k and β_k are coefficients depending only on $\mathbf{k}$ through $k = |\mathbf{k}|$. Further, these coefficients are defined through the recursive formulas

$$\sum_{j=1}^{k} \frac{\binom{k-1}{j-1}}{\binom{n-1}{j-1}} p_{j-1} \phi[\lambda(1-k/n)]^{k-j} \left\{ \alpha_j \frac{j-1}{k-1} + j \frac{\phi'[\lambda(1-k/n)]}{\phi[\lambda(1-k/n)]} \right\}$$

$$= 0, \qquad k = 2, \ldots, n, \tag{6}$$

$$\sum_{j=1}^{k} \frac{\binom{k-1}{j-1}}{\binom{n-1}{j-1}} p_{j-1}\phi[\lambda(1-k/n)]^{k-j}\left\{\alpha_j\beta_j + j\frac{\phi'[\lambda(1-k/n)]}{\phi[\lambda(1-k/n)]}\right\}$$

$$= 0, \qquad k = 1,\ldots,n, \tag{7}$$

where $\{p_j\}$ *are themselves defined recursively by*

$$\sum_{j=0}^{k-1} \frac{\binom{k-1}{j}}{\binom{n-1}{j}} p_j\phi[\lambda(1-k/n)]^{-(j+1)} = 1, \qquad k = 1,\ldots,n. \tag{8}$$

Remark. Theorem 3.1 was for the case where the initial infective was unknown and drawn completely at random from the whole population. More generally, if ω_u is the probability that u is the initial infective, then the log-derivative in the theorem becomes $(\Sigma_{u;\mathbf{k}_u=1}\omega_u)^{-1}$ $\Sigma_{u;\mathbf{k}_u=1}\omega_u r_{\mathbf{k}}^u$, where $r_{\mathbf{k}}^u = \alpha_k \mathbf{k}^T C(I - \mathbf{e}_u\mathbf{e}_u^t)\mathbf{k} - \alpha_k\beta_k\mathbf{k}^T C(\mathbf{1} - \mathbf{e}_u)$.

The distribution of Equation (5) depends on the nuisance parameter λ. Under the null hypothesis ($\delta = 0$), each realization with the same total number infected has equal probability by symmetry. Let $p_{k-1} = p_{k-1}(\lambda)$ denote the probability that $k-1$ individuals, besides the initially infective, are ultimately infected. We then have $\mathrm{P}^u(\mathbf{k} - \mathbf{e}_u, \lambda, 0) = p_{k-1}\binom{n-1}{k-1}^{-1}$ and Equation (8) is simply Equation (1) when $\delta = 0$. For the case where the initial individual is randomized, we have $\mathrm{P}(\mathbf{k}, \lambda, 0) = p_{k-1}\binom{n}{k}^{-1}$. This implies that $k = |\mathbf{k}|$ is sufficient for λ, so we condition upon this number, making α_k a constant term. We may therefore omit the factor α_k / k in Equation (5) when we construct the test statistic.

If $c_{i,\cdot} = c$ for all i, where $c_{i,\cdot} = \Sigma_j c_{i,j}$, then the term $\mathbf{k}^T C\mathbf{1}$ in Equation (5) becomes $k\beta_k c$ and is constant, given k. Thus we should use $\mathbf{K}^T C\mathbf{K} = \Sigma_{i,j} c_{i,j} K_i K_j$ as the test statistic of the hypothesis (we write uppercase letters for random vectors and variables). The interpretation of $c_{i,\cdot} = c$ is that the average number of contacts is the same for each individual—it is only *who* is contacted that may differ.

If, on the other hand, $c_{i,\cdot} \neq c_{j,\cdot}$ for some i and j, one has to compute β_k through Equations (6)–(8). From these equations, it is seen that β_k is a function of the nuisance parameter λ. The most natural way to proceed is to replace λ by a good estimator under the null hypothesis. When $\delta = 0$, the population mixes uniformly and the present model reduces to what is sometimes called the Generalized Epidemic Model.

Two well-known results for the Generalized Epidemic Model—for example, Martin-Löf [5]—state that, in a large population, a major outbreak can occur only if $\lambda > 1$ and that, in case of a major outbreak in a large population, the distribution of the final proportion infected is concentrated about τ, where

$$\tau \text{ is the positive solution of the equation } x = 1 - e^{-x\lambda}. \tag{9}$$

It follows that $\hat{\lambda} = -\hat{\tau}\log(1-\hat{\tau})$, where $\hat{\tau} = k/n$ is the observed proportion infected, is a good estimator for λ if the population is fairly large. Consequently, let $\hat{\beta}_k$ be defined from Equations (6)–(8) computed with $\lambda = \hat{\lambda}$.

This implies that Equation (5) is approximately equivalent to the statistic

$$T = \sum_{i,j} c_{i,j} K_i K_j - \hat{\beta}_k \sum_i c_{i,\cdot} K_i. \tag{10}$$

Numerical results indicate that, even in a small population, $\beta_k \approx k/n$ for rather arbitrary λ and distribution F (see Figure 1 for a population of size $n = 50$, $\lambda = 1.4$, and F the exponential distribution). This suggests that Equation (10) may be substituted by

$$T' = \sum_{i,j} c_{i,j} K_i K_j - \hat{\tau} \sum_i c_{i,\cdot} K_i = \sum_{i,j} c_{i,j} K_i \left(K_j - \hat{\tau}\right), \tag{11}$$

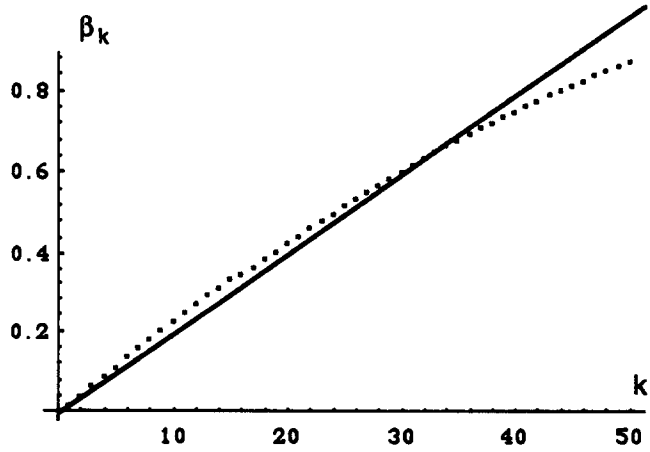

FIG. 1. β_k for $n = 50$, $\lambda = 1.4$, and $F \sim \exp(1)$ compared with the line k/n.

where $\hat{\tau}$ $(=\sum_i K_i/n)$ is the observed proportion infected. When we calculate the level of significance, we shall use the conditional null distribution of T' (or T), given the observed number of infected individuals k. This distribution is obtained from the underlying uniform distribution of the epidemic, which puts equal mass $\binom{n}{k}^{-1}$ to each configuration $\mathbf{k}=(k_1,\ldots,k_n)$ such that $|\mathbf{k}|=\sum_{i=1}^n k_i = k$. It has no closed form but, in a small population, it can, of course, be derived analytically. An alternative approach is to approximate the distribution by simulation—how this is done is discussed in Section 4.

If we observe an epidemic outbreak and want to test for homogeneity, we should use the following test procedure.

TEST PROCEDURE

Reject the hypothesis $\delta = 0$ *if the observed value* t *of Equation* (10) *is large compared with the distribution of* T, *which puts equal mass* $\binom{n}{k}^{-1}$ *to each realization* $(k_1,\ldots,k_n)$ *such that* $\sum_{i=1}^n k_i = k$ (k *is the observed total number infected*). *Replacing the statistic* T *by* T', *defined in Equation* (11), *gives an approximation that is simpler to calculate.*

4. LARGE POPULATION APPROXIMATIONS

When the population size n is fairly large—say, 100—two difficulties arise in the test procedure described in Section 3. The first is to calculate the exact value of $\hat{\beta}_k$ from Equations (6)–(8). In principle, this is easy, but the equations are numerically unstable (even a powerful computer may have problems in computing $\hat{\beta}_{50}$ when $n=100$, and this is a fairly small population). After $\hat{\beta}_k$ has been calculated, the second difficulty is to derive the exact conditional distribution of T in Equation (10), given the observed number infected $k=\sum_{i=1}^n k_i$.

These two problems motivate the need for approximations valid when n is fairly large, which is the case treated in this section. To emphasize the dependence on n, we will attach a superscript in our notation whenever necessary. It is assumed that, as $n\to\infty$,

$$\lambda^{(n)}\to\lambda, \qquad c_{i,j}^{(n)}\to c_{i,j}, \quad \text{and} \quad \sum_{j=1}^{n} c_{i,j}^{(n)}\to\eta_i\leqslant\eta, \tag{12}$$

where λ, $\{\eta_i\}$, and η are fixed positive constants. These assumptions imply that the average number of "homogeneous contacts" by an individual tends to a bounded limit and that, for a fixed positive δ, the number of contacts from the heterogeneous population structure is of the same order as the corresponding number from the homogeneous part.

In Appendix 2, it is shown that the approximation $\hat{\beta}_k^{(n)} \approx \hat{\tau} = k/n$ holds for "likely" k. This approximation justifies the use of T' defined in Equation (11) instead of T given in Equation (10), and this new statistic is very easily computed. The second problem remains: to derive an approximation for the conditional null distribution of T', given the observed number of infected k. This is the distribution that puts equal mass $\binom{n}{k}^{-1}$ to each $\mathbf{k}$ such that $|\mathbf{k}| = k$.

One way to approximate the conditional null distribution is by means of simulation. For fixed n and k, an outcome of the epidemic can be simulated simply by choosing k among the n individuals completely at random. When this has been done, it is straightforward to compute the value of T'. Through many repetitions of this procedure, the distribution of T' may be approximated by the induced empirical distribution.

A natural question is whether there exists a unique central limit theorem for the conditional distribution of T'. The answer is no. The asymptotic distribution of T' depends heavily on the contact matrix $C^{(n)}$. With different choices of $C^{(n)}$, satisfying Equation (12), a very broad class of limit distributions can be obtained. We will examine just two examples, with the Gaussian and χ^2 distributions as limits.

EXAMPLE 1. WITHIN-FAMILY INFECTIVITY

Testing procedures for this example have been studied previously [6–9]. The alternative hypothesis here is that there is an increased rate of infecting members of one's own family. This increased rate may or may not depend on the size of the family. For example, Schork [6] treats a slightly different model in which he assumes $c_{i,j} = 1$ if i and j belong to the same family, *independent* of the family size. Britton [7] considers the case $c_{i,j} = 1/(f-1)$ if i and j are in the same family whose size is f. This choice simplifies the test statistic because $c_{i,\cdot} = 1$ for all individuals (except singles). Here we treat the general case and let $c_{i,j} = \alpha_f$ if i and j belong to the same family of size f, where $\{\alpha_f\}$ are specified nonnegative coefficients. Let $N_{f,i}$ denote the number of families of size f with i infected individuals at the end of the epidemic, $0 \leqslant i \leqslant f \leqslant f_{max}$, where f_{max} is the size of the largest family. With this new notation we have $n = \sum_{\{0 \leqslant i \leqslant f \leqslant f_{max}\}} fN_{f,i}$ and $k = \sum_{\{0 \leqslant i \leqslant f \leqslant f_{max}\}} iN_{f,i}$, and Equation (11) becomes

$$T' = \sum_{f=2}^{f_{max}} \sum_{i=1}^{f} \alpha_f i(i-1) N_{f,i} - \hat{\tau} \sum_{f=2}^{f_{max}} \sum_{i=1}^{f} (f-1)\alpha_f i N_{f,i}. \qquad (13)$$

We seek the asymptotic conditional distribution of T', given $\hat{\tau}$. It can be shown [10] that, if the proportion of different family sizes as well as $\hat{\tau}$ converge, then the conditional distribution of the vector $\mathbf{N}^{(n)} = \{N_{f,i}^{(n)}\}$

converges in distribution to a singular Gaussian vector. This implies that the conditional distribution of T', a linear combination of $\mathbf{N}^{(n)}$, converges to a normal distribution. It can also be shown that the conditional normalizing moments are

$$\mathrm{E}(T' \mid \hat{\tau}) = 0$$

and

$$\mathrm{var}(T' \mid \hat{\tau}) = 2\hat{\tau}^2(1-\hat{\tau})^2 \sum_f \alpha_f^2 f(f-1) n_f + \hat{\tau}^3(1-\hat{\tau}) \left\{ \sum_f \alpha_f^2 f(f-1)^2 n_f - \frac{\left[\Sigma_f \alpha_f f(f-1) n_f\right]^2}{n} \right\},$$

where n_f denotes the number of families with size f.

For the specific case $\alpha_f = 1$, considered by Schork [6], T' may also be written as $\Sigma_{f,i}\, i[i-1-\hat{\tau}(f-1)]N_{f,i}$. If, instead, we have $\alpha_f = 1/(f-1)$, as in Equation [7], we get $T' = \Sigma_{f,i \geqslant 2}\, i(i-1)N_{f,i}/(f-1) + \hat{\tau}N_{1,1} - n\hat{\tau}^2$ (note that α_f has to be defined only for $f \geqslant 2$ because, in single individual families, there is no one to infect further). The variance expression for this choice of $\{\alpha_f\}$ is $2\hat{\tau}^2(1-\hat{\tau})^2 \Sigma_f f n_f/(f-1) + \hat{\tau}^3(1-\hat{\tau})(n_1 - n_1^2/n)$.

EXAMPLE 2. MULTITYPE EPIDEMIC

Here we consider a population consisting of two groups (or types) of equal size where it is suspected that there might be an increased contact rate between individuals belonging to the same group. The same example is analyzed further in Appendix 2. We define the contact matrix C by $c_{i,j} = 1/n$ if i and j are in the same group and $c_{i,j} = 0$ otherwise. Let $\overline{M}_1$ and $\overline{M}_2$ denote the proportion infected in the two groups. Using Equation (11), we get $T' = -(1-\hat{\tau})/2 + n(\overline{M}_1 - \overline{M}_2)^2/(8\hat{\tau})$, and, given k (i.e., given $\hat{\tau}$), the randomness comes solely from $(\overline{M}_1 - \overline{M}_2)^2$. Because $(\overline{M}_1 + \overline{M}_2)/2 = \hat{\tau}$, this test is equivalent to the χ^2 test. The test procedure is thus to reject the null hypothesis if $n(\overline{M}_1 - \overline{M}_2)^2/[4\hat{\tau}(1-\hat{\tau})]$ is large compared with the χ^2 distribution with 1 degree of freedom.

More generally, if we have k subpopulations of equal size, we get a χ^2 test with $k-1$ degrees of freedom. If the sizes are different, one also will obtain a χ^2 test if the model is suitably parametrized.

5. SIMULATIONS OF EPIDEMICS

To get a sense of the magnitude of heterogeneity (i.e., the size of δ in comparison with λ) that is necessary for detection by the test, I have performed some simulations. The answer will, of course, depend on the

type of heterogeneity considered. For the case in which the heterogeneity is due to families or households, the reader is referred to Britton [7], where a special case of the present test is also illustrated with several sets of real data. In the present section, we simulate an epidemic modeling the spread of a disease among cows in a cattle house, say, where it is assumed that there is a higher risk of infecting the right and left "neighbor" if the disease spreads from cow to cow. In this situation, the infected cows should tend to cluster, meaning that infected cows will stand close to each other. If, on the other hand, the disease is caused by food or water, then there would be no reason for the clustering to appear. In this made-up example, we assume for simplicity that all cows have exactly two neighbors, one to the left and one to the right, meaning that they stand in a circle. We label the cows $1,\ldots,n$ so that the heterogeneous contact matrix C has nonzero elements $c_{i,i-1} = c_{i,i+1} = 1$ (for $i = 1$, this is interpreted as $c_{1,n} = c_{1,2} = 1$) and the remaining elements of C are 0. Because $c_{i,\cdot} = c_{i,i-1} + c_{i,i+1} = 2$ for each i, the second term in Equation (10) is a constant. The test statistic that we should use is hence only the first term of Equation (10), which is equivalent to

$$T = c_{1,n}K_1K_n + \sum_{i=1}^{n-1} c_{i,i+1}K_iK_{i+1}. \tag{14}$$

I have simulated epidemics for different population sizes; the sizes were chosen as $n = 25$, 50, 100, and 200. For each of these sizes, epidemics with various magnitudes of heterogeneity were simulated. In all simulations, $\lambda + \delta$ was kept fixed equal to 1.25, resulting in outbreak sizes with, on the average, 30–50% of the whole population. The degree of heterogeneity, denoted by γ in Table 1, was measured by the relative proportion of δ; that is, $\gamma = \delta/(\delta + \lambda) = \delta/1.25$. The choices of γ

TABLE 1

Simulations of an Epidemic in a Cattle House: Percentage of Simulations Rejecting the Hypothesis of Homogeneity at the 5% Level (and 1% level within parenthesis) for Different Population Sizes and Degrees of Heterogeneity

Population Size (n)	Degree of Heterogeneity (γ)					
	0	0.01	0.05	0.10	0.20	0.50
25	1.6 (0.5)	2.5 (1.0)	3.7 (1.1)	5.5 (1.7)	11.9 (4.9)	39.7 (22.1)
50	2.5 (0.3)	3.8 (0.5)	9.1 (2.3)	15.4 (5.7)	31.1 (13.4)	82.4 (63.7)
100	5.4 (0.2)	6.1 (0.9)	15.1 (4.3)	28.3 (11.3)	57.2 (30.4)	98.5 (93.8)
200	4.8 (1.5)	9.0 (3.7)	21.0 (9.4)	45.2 (26.4)	82.0 (67.1)	99.6 (99.5)

were 0%, 1%, 5%, 10%, 20%, and 50%. The epidemic was initiated by one single infected cow (which cow being irrelevant owing to symmetry), and the distribution of the infectious periods was chosen to be the exponential distribution (μ was normed to 1). For each combination of n and γ, 1000 simulations resulting in outbreaks of more than five infected cows were performed (the reason for this restriction is given below).

Because of the time consumption required to compute the exact conditional null distribution of T in Equation (14) for different values of n and outbreak sizes, the distributions were approximated by using the normal distribution. This is, of course, not recommended in real situations when the population is small. However, in the present context, the only purpose of the simulations is to show what combinations of γ and n are necessary to get a test with descent power. If the total outbreak size is very small, the normal approximation is not even approximately true. For this reason, simulations resulting with five or fewer infected were neglected. Because T is integer-valued, $1/2$ was used as a continuity correction to improve the normal approximation. Table 1 gives the percentage of the simulations that were rejected at the 5% significance level, with the corresponding percentage at the 1% level given in parenthesis. From Table 1, one concludes that, if the degree of heterogeneity is less than 10%, a population of several hundred is needed for the test to have a reasonable chance of rejecting the hypothesis of homogeneity; whereas, if the degree of heterogeneity is larger, it may be detected with fewer than 100 cows. For $\gamma = 0$, we see that the size of the tests seems approximately correct, at least when n is fairly large.

6. DISCUSSION

The case in which the heterogeneous structure of interest is the presence of families has been studied previously by several authors. With the choice of $c_{i,j} = 1/(f-1)$ if i and j belong to the same family whose size is f (see Example 1), the test procedure using T' of this paper coincides with the test suggested by Britton [7], originally from [10]. In the paper at hand, the score statistic is obtained by using exact methods; only the approximation $\beta_k \approx \hat{\tau}$ relies on a large population. Britton [10] derives a central limit theorem for the final state of the population and uses this to approximate the likelihood. The score statistic is then obtained by differentiating the approximate log likelihood. Walter [8] proposes a test statistic, which is the first term of Equation (13) with $\alpha_f = 1$, for the hypothesis that the disease does not aggregate in families. Leaving out the negative term in Equation (13)

has the effect that large families influence the statistic more than do small families. To avoid this effect, Fraser [9] suggests a statistic that is the first term of Equation (13) with $\alpha_f = 1/f$. Commenges et al. [11] derive the score test for the hypothesis of homogeneity between groups (or, equivalently, families) when the alternative hypothesis is that random effects, in a logistic regression model, of individuals belonging to the same group are positively correlated. The resulting score statistic for their model is of the same type as the present one. If all explanatory variables are identical, the score statistic of their model is like Equation (13), with $\alpha_f = 1$ in the first term and $\alpha_f = 2$ in the second term. This means that an equivalent version of the statistic is $\sum_{f=1}^{f_{max}} \sum_{i=0}^{f} i(i - 2f\bar{\tau}) N_{f,i}$. Although the test statistic proposed in the present paper is similar to the one in Commenges et al. [11], the technique used to derive the test differs. The reason for this is that the two models are fundamentally different. The model in Commenges et al. [11] is designed to allow dependencies within families due to random effects—for example, explaining genetic similarities—and all individuals behave *independently* conditional on the random effects. In the present model, individuals infect each other, so they are dependent in a more genuine way.

Test procedures, based on models for infectious diseases, suggested for testing homogeneity when the alternative is a general specified heterogeneous structure have received less attention in the literature. However, in Chapter 5 of Becker [12], such testing procedures are discussed. The main emphasis there is on epidemics observed over time in which "close" individuals can be separated into disjoint groups, but an example in which "close" is the same as "living close to each other," which does not separate the population, also is discussed. The test statistics presented in Chapter 5 of Becker [12] either are similar to the Mantel-Haenzel statistic or are quadratic forms. The present statistic, Equation (10), share several properties with these statistics—for example, the first term is also a quadratic form. In the present paper, the statistic is derived as the score statistic in a parametrized epidemic model. Recently and independently, a similar test of homogeneity was constructed by Commenges and Jacqmin-Gadda [13]. They treat a random-effect model describing the occurrence of, for example, noninfectious diseases. For this model, they derive the score test for the hypothesis of homogeneity, when the alternative hypothesis is that certain parameters, associated with the survival distribution functions, of different individuals are correlated in a rather arbitrary way. Just as for the family (or group) structure mentioned above, the model and the technique used to derive the test are different even though the resulting test statistic is similar.

The alternative hypothesis in the present paper is rather general but, unfortunately, the null hypothesis is not. It would therefore be of great interest to derive the corresponding test statistic when the null hypothesis is generalized. For example, one might want to allow children to have a higher infection risk than adults when testing for family effects. With a more general null hypothesis, one would also be able to test for certain heterogeneous structures sequentially, in more and more complex models until the hypothesis of homogeneity is accepted. For example, if the family effect is significant, one might want to test the effect of schools or other "semilocal" structures. In principle, a test statistic under such a generalized null hypothesis may be derived from the recursive formula of Equation (1); however, it seems to be difficult to do so in practice. The resulting test statistic will most likely be more complicated than the present statistic [Eq. (10) or its approximation, Eq. (11)].

APPENDIX 1. PROOF OF THEOREM 3.1

First, we show Equation (5). Fix a vector $\mathbf{k}$. For u such that $k_u = 1$, let $r_{\mathbf{k}}^u := \mathrm{P}^{u'}(\mathbf{k}-\mathbf{e}_u; \lambda, 0)/\mathrm{P}^u(\mathbf{k}-\mathbf{e}_u; \lambda, 0)$, where $\mathrm{P}^{u'}$ denotes the derivative with respect to δ. Then

$$\begin{aligned}
&\frac{\partial}{\partial\delta}\log \mathrm{P}(\mathbf{k}; \lambda, \delta)\Big|_{\delta=0} \\
&\quad = \frac{\partial}{\partial\delta}\log\left[\frac{1}{n}\sum_{u:\, k_u=1}\mathrm{P}^u(\mathbf{k}-\mathbf{e}_u; \lambda, \delta)\right]\Bigg|_{\delta=0} \\
&\quad = \frac{\sum_{u:\, k_u=1}\mathrm{P}^{u'}(\mathbf{k}-\mathbf{e}_u; \lambda, 0)}{\sum_{u:\, k_u=1}\mathrm{P}^u(\mathbf{k}-\mathbf{e}_u; \lambda, 0)} \\
&\quad = \frac{1}{k}\sum_{u:\, k_u=1}\frac{\mathrm{P}^{u'}(\mathbf{k}-\mathbf{e}_u; \lambda, 0)}{\mathrm{P}^u(\mathbf{k}-\mathbf{e}_u; \lambda, 0)} = \frac{1}{k}\sum_{u:\, k_u=1} r_{\mathbf{k}}^u.
\end{aligned}$$

The third equality is true because all k terms in the denominator sum are equal. This follows because, when $\delta = 0$, we have a homogeneously mixing population. So, by symmetry, $\mathrm{P}^u(\mathbf{j}; \lambda, 0) = \binom{n-1}{j}^{-1} p_j$, where p_j is the probability that j individuals, besides individual u, will get infected, and $j = |\mathbf{j}|$. By comparing the preceding equation with Equation (5), we see that the first part of the theorem is proved if we can show that $\sum_{\{u:\, k_u=1\}} r_{\mathbf{k}}^u = \alpha_k(\mathbf{k}^T C\mathbf{k} - \beta_k \mathbf{k}^T C\mathbf{1})$. It follows from Equation (1) that

$$k = \sum_{u;\, k_u=1}\ \sum_{\mathbf{j}=\mathbf{0}}^{\mathbf{k}-\mathbf{e}_u} \mathrm{P}^u(\mathbf{j}; \lambda, \delta)\Phi[\Lambda(\delta)(\mathbf{1}-\mathbf{k})]^{-(\mathbf{j}+\mathbf{e}_u)} \qquad \forall\, \delta \geqslant 0, \quad \text{(A1)}$$

where $\Lambda(\delta)$ is the matrix with off-diagonal elements given by Equation (2). It is straightforward to show that $\Phi[\Lambda(0)(\mathbf{1}-\mathbf{k})]^{-(\mathbf{j}+\mathbf{e}_u)} = \phi(\lambda(1-k/n))^{-(j+1)}$ and that

$$\frac{\partial}{\partial\delta}\Phi[\Lambda(\delta)(\mathbf{1}-\mathbf{k})]^{-(\mathbf{j}+\mathbf{e}_u)}\Big|_{\delta=0} = -\frac{\phi'[\lambda(1-k/n)]}{\phi[\lambda(1-k/n)]^{j+2}}(\mathbf{j}+\mathbf{e}_u)^T C(\mathbf{1}-\mathbf{k}).$$

Thus, differentiating both sides of Equation (A1) with respect to δ, setting $\delta = 0$, and changing the order of summation gives us

$$0 = \sum_{j=0}^{k-1}\left\{\binom{n-1}{j}^{-1} p_j \phi[\lambda(1-k/n)]^{-(j+1)} \times \sum_{\substack{\mathbf{j};\mathbf{j}\leqslant\mathbf{k}\\|\mathbf{j}|=j}}\sum_{\substack{u;k_u=1\\j_u=0}}\left(r_{\mathbf{j}}^u - \frac{\phi'[\lambda(1-k/n)]}{\phi[\lambda(1-k/n)]}(\mathbf{j}+\mathbf{e}_u)^T C(\mathbf{1}-\mathbf{k})\right)\right\}. \quad \text{(A2)}$$

For $0 \leqslant j < k$, let

$$s_{\mathbf{k}}^{(k-j)} := \sum_{\substack{\mathbf{j};\mathbf{j}\leqslant\mathbf{k}\\|\mathbf{j}|=j}}\sum_{\substack{u;k_u=1\\j_u=0}} r_{\mathbf{j}}^u,$$

so, for fixed j, $s_{\mathbf{k}}^{(k-j)}$ is the first part of the inner sum in Equation (A2). With this definition, we have $s_{\mathbf{k}}^{(1)} = \sum_{\{u:k_u=1\}} r_{\mathbf{k}}^u$, which is the quantity that we want to show equals $\alpha_k(\mathbf{k}^T C\mathbf{k} - \beta_k \mathbf{k}^T C\mathbf{1})$. We now state and prove a lemma.

LEMMA A.1

For $2 \leqslant i \leqslant k$,

$$s_{\mathbf{k}}^{(i)} = \sum_{\substack{\mathbf{j};\mathbf{j}\leqslant\mathbf{k}\\|\mathbf{j}|=i-1}} s_{\mathbf{k}-\mathbf{j}}^{(1)}. \quad \text{(A3)}$$

Proof. By the definition of $s_{\mathbf{k}}^{(k-j)}$, the right- and left-hand sides of Equation (A3) are equal to

$$\sum_{\substack{\mathbf{j};\mathbf{j}\leqslant\mathbf{k}\\|\mathbf{j}|=k-i}}\sum_{\substack{u;k_u=1\\j_u=0}} r_{\mathbf{j}}^u \quad \text{and} \quad \sum_{\substack{\mathbf{j};\mathbf{j}\leqslant\mathbf{k}\\|\mathbf{j}|=i-1}}\sum_{u;(\mathbf{k}-\mathbf{j})_u=1} r_{\mathbf{k}-\mathbf{j}-\mathbf{e}_u}^u,$$

respectively. The first of these two sums contains $\binom{k}{k-i} i$ terms and the latter has $\binom{k}{i-1}(k-i+1) = \binom{k}{k-i} i$, that is, the same. Further, each term in the first sum is also found in the second. This completes the proof. ■

If we use the identity $\sum_{\substack{\mathbf{j};\mathbf{j}\leqslant\mathbf{k} \\ |\mathbf{j}|=j}} \sum_{\substack{u;k_u=1 \\ j_u=0}} (\mathbf{j}+\mathbf{e}_u)^T = (j+1)\binom{k-1}{j}\mathbf{k}^T$, which is easy to show, and rearrange in Equation (A2), it follows that

$$s_{\mathbf{k}}^{(1)} = k\frac{\phi'[\lambda(1-k/n)]}{\phi[\lambda(1-k/n)]}\mathbf{k}^T C(\mathbf{1}-\mathbf{k}) - \frac{\binom{n-1}{k-1}}{p_{k-1}} \sum_{j=1}^{k-1} \left\{ \binom{n-1}{j-1}^{-1} p_{j-1}\phi[\lambda(1-k/n)]^{k-j} \times \left(s_{\mathbf{k}}^{(k-j+1)} - j\frac{\phi'[\lambda(1-k/n)]}{\phi[\lambda(1-k/n)]} \binom{k-1}{j-1} \mathbf{k}^T C(\mathbf{1}-\mathbf{k}) \right) \right\}. \quad \text{(A4)}$$

We are finally able to prove that $s_{\mathbf{k}}^{(1)} = \alpha_k(\mathbf{k}^T C\mathbf{k} - \beta_k \mathbf{k}^T C\mathbf{1})$, which we do by induction. Suppose $\mathbf{k}$ satisfies $|\mathbf{k}| = 1$. From Equation (A4), it then follows that $s_{\mathbf{k}}^{(1)} = \{\phi'[\lambda(1-1/n)]/\phi[\lambda(1-1/n)]\}\mathbf{k}^T C\mathbf{1}$ because $\mathbf{k}^T C\mathbf{k} = 0$ when $|\mathbf{k}| = 1$ because the diagonal elements of C are all 0). So, with $\alpha_1\beta_1 = -\phi'[\lambda(1-1/n)]/\phi[\lambda(1-1/n)]$, Equation (A4) is satisfied.

Assume now that $s_{\mathbf{k}'}^{(1)} = \alpha_{k'}(\mathbf{k}'^T C\mathbf{k}' - \beta_{k'}\mathbf{k}'^T C\mathbf{1})$ for all $\mathbf{k}'$ such that $k' = |\mathbf{k}'| < k$ and let $\mathbf{k}$ satisfy $|\mathbf{k}| = k$. Then, for $1 \leqslant j \leqslant k-1$,

$$s_{\mathbf{k}}^{(k-j+1)} = \sum_{\substack{\mathbf{v};\mathbf{v}\leqslant\mathbf{k} \\ |\mathbf{v}|=k-j}} s_{\mathbf{k}-\mathbf{v}}^{(1)} = \sum_{\substack{\mathbf{v};\mathbf{v}\leqslant\mathbf{k} \\ |\mathbf{v}|=k-j}} \alpha_j(\mathbf{k}-\mathbf{v})^T C(\mathbf{k}-\mathbf{v}) - \sum_{\substack{\mathbf{v};\mathbf{v}\leqslant\mathbf{k} \\ |\mathbf{v}|=k-j}} \alpha_j\beta_j(\mathbf{k}-\mathbf{v})^T C\mathbf{1}.$$

The first equality is Lemma A.1 and the second follows from the induction assumption. Basic combinatorics implies that the first sum on the right-hand side is equal to $\alpha_j\binom{k-2}{k-j}\mathbf{k}^T C\mathbf{k}$ (this is true only because the diagonal elements of C are all 0). The second sum equals $\alpha_j\beta_j\binom{k-1}{k-j}\mathbf{k}^T C\mathbf{1}$. This implies that

$$s_{\mathbf{k}}^{(k-j+1)} = \alpha_j\binom{k-1}{j-1}\left(\frac{j-1}{k-1}\mathbf{k}^T C\mathbf{k} - \beta_j\mathbf{k}^T C\mathbf{1}\right). \quad \text{(A5)}$$

If we insert this into Equation (A4), we see that $s_{\mathbf{k}}^{(1)}$ may clearly be written as $\alpha_k(\mathbf{k}^T C\mathbf{k} - \beta_k \mathbf{k}^T C\mathbf{1})$, which completes the induction step.

In the theorem, it was also claimed that α_j, β_j, and p_j satisfy Equations (6)–(8), which we now prove. That p_j are defined by Equation (8) is a direct consequence of Equation (1) with $\delta = 0$, which can also be deduced from Equation (A1) with $\delta = 0$. To show the defining recursive formulas for α_j and β_j, we use Equation (A4), with $s_{\mathbf{k}}^{(k-j+1)}$ replaced by the expression in Equation (A5). Extend the summation in Equation (A4) to $j = k$ and write the same term with opposite sign in front of the summation:

$$
\begin{aligned}
s_{\mathbf{k}}^{(1)} = {} & \alpha_k\left(\mathbf{k}^T C\mathbf{k} - \beta_k \mathbf{k} C\mathbf{1}\right) \\
& - \mathbf{k}^T C\mathbf{k} \frac{\binom{n-1}{k-1}}{p_{k-1}} \sum_{j=1}^{k} \frac{\binom{k-1}{j-1}}{\binom{n-1}{j-1}} p_{j-1} \phi\left[\lambda(1-k/n)\right]^{k-j} \\
& \times \left\{ \alpha_j \frac{j-1}{k-1} + j \frac{\phi'\left[\lambda(1-k/n)\right]}{\phi\left[\lambda(1-k/n)\right]} \right\} \\
& + \mathbf{k}^T C\mathbf{1} \frac{\binom{n-1}{k-1}}{p_{k-1}} \sum_{j=1}^{k} \frac{\binom{k-1}{j-1}}{\binom{n-1}{j-1}} p_{j-1} \phi\left[\lambda(1-k/n)\right]^{k-j} \\
& \times \left\{ \alpha_j \beta_j + j \frac{\phi'\left[\lambda(1-k/n)\right]}{\phi\left[\lambda(1-k/n)\right]} \right\}.
\end{aligned}
$$

Because both sums must be identically 0, we see that α_j and β_j are defined as was claimed.

APPENDIX 2. VERIFICATION OF LARGE POPULATION APPROXIMATION

As indicated in Figure 1, a natural guess is that $\beta_k^{(n)} \approx k/n$ for all k or, more strictly, $\beta_{[nx]}^{(n)} \to x$, for $x \in [0,1]$ as $n \to \infty$ ($[nx]$ denotes the integer part of nx). This is a conjecture that the author has not been able to prove. However, under H_0 for fixed $\lambda > 1$ and in the case of a major epidemic, the proportion infected, $|\mathbf{K}^{(n)}|/n$, converges in probability to τ, defined in Equation (9). This means that it suffices to know $\beta_k^{(n)}$ for $k \approx n\tau$. In this appendix, we therefore show that, in a large

population, the approximation $\beta_k^{(n)} \approx \hat{\tau} = k/n$ holds, where k now is the observed number infected. This justifies the use of T' instead of T, defined by Equations (11) and (10), respectively, in a fairly large population.

Equations (6)–(8) defining $\beta_k^{(n)}$ are independent of the contact matrix $C^{(n)}$. We may thus choose $C^{(n)}$ of a simple form for which asymptotic results for the distribution of $\mathbf{K}^{(n)}$ are known. Our choice is to separate the population into two subpopulations and to assume increased infectivity within each subpopulation. Let $C^{(n)}$ have elements

$$c_{i,j}^{(n)} = \begin{cases} 1/n & \text{if } \max\{i,j\} \leqslant n/2 \text{ or } \min\{i,j\} > n/2 \\ 0 & \text{otherwise} \end{cases} \tag{A6}$$

and define, as before, $c_{i,i}^{(n)} = 0$. This means that we have a symmetric two-type epidemic, which falls under a more general model treated by Ball and Clancy [4].

For the asymptotics, it turns out to be convenient to use a somewhat different parametrization. Assume $C^{(n)}$ to be of the form in Equation (A6), but let $\Lambda^{(n)} = \Lambda^{(n)}(\delta)$ have off-diagonal elements $\lambda_{i,j}^{(n)}(\delta) = (\lambda - \delta/2)/n + \delta c_{i,j}^{(n)}$, where λ is considered to be fixed but δ may vary in the interval $[0, 2\lambda]$. The reason for choosing this parametrization is that the law-of-large-number limit for the overall proportion infected remains constant as we vary δ. Thus, the final proportion infected will contain information only about λ and not about δ, the parameter of interest. This means that, asymptotically, the final proportion infected is an ancillary statistic and the Conditionality Principle tells us to condition on the observed value of this proportion [14]. Note that there is a one-to-one correspondence between this parametrization and the one in Equation (2) and that the hypothesis remains unchanged, so we have the same model.

We should look for an approximation of $\mathrm{P}(\mathbf{k}; \lambda - \delta/2, \delta)$ for the contact matrix defined by Equation (A6). Assume, for simplicity, that n is even so that $n/2$ is an integer, and define $m_1 = \sum_{i=1}^{n/2} k_i$ and $m_2 = \sum_{i=n/2+1}^{n} k_i$. This means that m_1 is the number infected in the first subpopulation and m_2 is the corresponding number in the second. Let $\mathrm{P}(m_1, m_2; \lambda - \delta/2, \delta)$ denote the probability that m_1 individuals in the first subpopulation and m_2 in the second were infected. By symmetry, we then have

$$\mathrm{P}(\mathbf{k}; \lambda - \delta/2, \delta) = \binom{n/2}{m_1}^{-1} \binom{n/2}{m_2}^{-1} \mathrm{P}(m_1, m_2; \lambda - \delta/2, \delta).$$

If M_1 and M_2 are the corresponding random variables and $\overline{M}_i = M_i/(n/2)$ with $i=1,2$, denote the proportion infected, Equation (4.9) in Ball and Clancy [4] states that, in case of a major epidemic,

$$\sqrt{n}\begin{bmatrix}\overline{M}_1^{(n)}-\tau\\ \overline{M}_2^{(n)}-\tau\end{bmatrix} \xrightarrow{\mathscr{D}} N\left[\begin{pmatrix}0\\0\end{pmatrix}, 2S^{-1}\Xi S^{-1}\right]. \tag{A7}$$

The matrices S and Ξ are 2×2 and doubly symmetric. Matrix S has diagonal elements $S_{1,1}=S_{2,2}=1-(1-\tau)(\lambda/2+\delta/4)$ and off-diagonal elements $S_{1,2}=S_{2,1}=-(1-\tau)(\lambda/2-\delta/4)$. The matrix Ξ has diagonal elements $\Xi_{1,1}=\Xi_{2,2}=\tau(1-\tau)[1+(1-\tau)r^2(2\lambda^2+\delta^2/4)/4]$ and the other two elements are $\Xi_{1,2}=\Xi_{2,1}=\tau(1-\tau)^2(\lambda^2-\delta^2/4)/2$, where r^2 is the variance of the infectious period, $r^2 := \int_0^\infty x^2\,dF(x)-[\int_0^\infty x\,dF(x)]^2$, assumed to be finite.

For a sequence $\{m_1^{(n)}, m_2^{(n)}\}$ such that $m_1^{(n)}, m_2^{(n)} \in (\tau n/2 - a\sqrt{n}, \tau n/2 - a\sqrt{n})$ for fixed a, Equation (A7) justifies the following approximation of the log likelihood:

$$\begin{aligned}&\log \mathrm{P}\left[m_1^{(n)}, m_2^{(n)}; \lambda-\delta/2, \delta\right]\\ &\quad\approx \log\left(\sqrt{\left|\frac{n}{2}S\Xi^{-1}S\right|}\Big/2\pi\right)\\ &\qquad -\frac{n}{4}\left[\overline{m}_1^{(n)}-\tau \quad \overline{m}_2^{(n)}-\tau\right]S\Xi^{-1}S\left[\overline{m}_1^{(n)}-\tau \quad \overline{m}_2^{(n)}-\tau\right]^T.\end{aligned}$$

Differentiating the right-hand side with respect to δ and then setting $\delta=0$ gives—after some algebra, preferably done using a computer program—exactly the right-hand side in Equation (A8). We hence approximate the log derivative by

$$\frac{\partial}{\partial\delta}\log \mathrm{P}\left[m_1^{(n)}, m_2^{(n)}; \lambda-\delta/2, \delta\right]\Big|_{\delta=0} \approx -\frac{1-\tau}{2}+\frac{n}{8\tau}\left[\overline{m}_1^{(n)}-\overline{m}_2^{(n)}\right]^2. \tag{A8}$$

To find an approximation of β_k, we now compare this large-population approximation with the corresponding exact result for this particular choice of C. The exact log derivative can be obtained from results of Section 3. First note that

$$\begin{aligned}&\frac{\partial}{\partial\delta}\log \mathrm{P}(m_1, m_2; \lambda-\delta/2, \delta)\Big|_{\delta=0}\\ &\quad=\frac{\partial}{\partial\delta}\log \mathrm{P}(m_1, m_2; \lambda, \delta)\Big|_{\delta=0}\\ &\qquad-\frac{1}{2}\frac{\partial}{\partial\lambda'}\log \mathrm{P}(m_1, m_2; \lambda', 0)\Big|_{\lambda'=\lambda}.\end{aligned} \tag{A9}$$

The first term is given by Equation (5). The second term also can be obtained from Equation (5) with a different choice of C (we may choose a different C because $\delta = 0$). Let $C = n^{-1}[\mathbf{1}\mathbf{1}^T - \text{diag}(\mathbf{1})]$. With this choice of C, the contact rates are, for $i \neq j$, $\lambda_{i,j} = \lambda/n + \delta/n$, so $P(m_1, m_2; \lambda + h, 0) = P(m_1, m_2; \lambda, h)$. The negative term on the right-hand side of Equation (A9) hence equals Equation (5) multiplied by $-1/2$ with $C = n^{-1}[\mathbf{1}\mathbf{1}^T - \text{diag}(\mathbf{1})]$. Using Equation (5) with these choices of C and observing that $m_1 + m_2 = k$, we get

$$\begin{aligned}
&\frac{\partial}{\partial \delta} \log P(m_1, m_2; \lambda - \delta/2, \delta)\Big|_{\delta = 0} \\
&\quad = \frac{\alpha_k}{k}\left\{\frac{m_1(m_1-1) + m_2(m_2-1)}{n} - \beta_k\left(\frac{1}{2} - \frac{1}{n}\right)k\right. \\
&\qquad\qquad \left. - \frac{1}{2}\left[\frac{k(k-1)}{n} - \beta_k \frac{n-1}{n}k\right]\right\} \\
&\quad = \frac{\alpha_k}{k}\frac{k}{n}\left[-\frac{1 - k/n}{2} + \frac{n}{8(k/n)}(\bar{m}_1 - \bar{m}_2)^2\right] \\
&\qquad - \frac{\alpha_k}{k}\left(\frac{k}{n} - \beta_k\right)\frac{k}{2n}. \qquad \text{(A10)}
\end{aligned}$$

If we compare the last row in Equation (A10), which is an exact result, with the large-population approximation [Eq. (A8)] and remember that $k/n = \hat{\tau} \approx \tau$ asymptotically, we see that we must have $\beta_k \approx \hat{\tau}$. We also note that $\alpha_k \approx n$, as a by-product.

It is worth emphasizing that we have only motivated the approximation $\beta_k \approx \hat{\tau} = k/n$ for $k \approx \tau n$, because the central limit theorem of Ball and Clancy [4] justifies approximations of $P(m_1, m_2; \delta)$ only when $m_1, m_2 \in (\tau n/2 - a\sqrt{n}, \tau n/2 - a\sqrt{n})$. Fortunately, the probability that we will observe such m_1 and m_2 can be made arbitrarily close to 1 by choosing a large, a consequence of the same theorem.

This paper is part of a PhD thesis written under the guidance of Åke Svensson, who is gratefully acknowledged. The author was supported in part by The Bank of Sweden Tercentenary Foundation.

REFERENCES

1 C. L. Addy, I. M. Longini, and M. Haber, A generalized stochastic model for the analysis of infectious disease final size data. *Biometrics* 47:961–974 (1991).

2 F. Ball, A unified approach to the distribution of the total size and total area under the trajectory of infectives in epidemic models. *Adv. Appl. Prob.* 18:289–310 (1986).

3 P. Picard and C. Lefèvre, A unified analysis of the final size and severity distribution in collective Reed-Frost epidemic processes. *Adv. Appl. Prob.* 22:269–294 (1990).
4 F. Ball and D. Clancy, The final size and severity of a generalized stochastic multitype epidemic model. *Adv. Appl. Prob.* 25:721–736 (1993).
5 A. Martin-Löf, Symmetric sampling procedures, general epidemic processes, and their threshold limit theorems. *J. Appl. Prob.* 23:265–282 (1986).
6 N. J. Schork, Sampling guidelines for testing secondary attack rates associated with short-latency infectious diseases. *Stat. Med.* 13:1563–1573 (1994).
7 T. Britton, Tests to detect clustering of infected individuals within families. In Epidemics with heterogeneous mixing: stochastic models and statistical tests, PhD Thesis, Department of Mathematics, Stockholm University (1996). *Biometrics* 53:50–61 (1997).
8 S. D. Walter, On the detection of household aggregation of disease. *Biometrics* 30:525–538 (1974).
9 D. W. Fraser, Clustering of a disease in population units: an exact test and its asymptotic version. *Am. J. Epidemiol.* 118:732–739 (1983).
10 T. Britton, Limit theorems and tests for within family clustering in epidemic models. In Epidemics with heterogeneous mixing: stochastic models and statistical tests, PhD Thesis, Department of Mathematics, Stockholm University (1996), in press.
11 D. Commenges, L. Letenneur, H. Jacqmin, T. Moreau, and J.-F. Dartigues, Test of homogeneity of binary data with explanatory variables. *Biometrics* 50:613–620 (1994).
12 N. G. Becker, *Analysis of infectious disease data*. Chapman and Hall, London, 1989.
13 D. Commenges and H. Jacqmin-Gadda, Generalized score test of homogeneity based on correlated random effects model. *J. R. Stat. Soc. B* 59: in press (1997).
14 D. R. Cox and D. V. Hinkley, *Theoretical statistics*. Chapman and Hall, Cambridge, 1974.