ELSEVIER

Journal of Statistical Planning and Inference 96 (2001) 41–66

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Design issues for studies of infectious diseases

Niels G. Becker[a,*], Tom Britton[b]

[a] *National Centre for Epidemiology and Population Health, The Australian National University, Canberra, ACT 0200, Australia*

[b] *Department of Mathematics, Uppsala University, Box 480, S-751 06 Uppsala, Sweden*

Abstract

The design of infectious disease studies has received little attention because they are generally viewed as observational studies. That is, epidemic and endemic disease transmission happens and we observe it. We argue here that statistical design often provides useful guidance for such studies with regard to type of data and the size of the data set to be collected. It is shown that data on disease transmission in part of the community enables the estimation of central parameters and it is possible to compute the sample size required to make inferences with a desired precision. We illustrate this for data on disease transmission in a single community of uniformly mixing individuals and for data on outbreak sizes in households. Data on disease transmission is usually incomplete and this creates an identifiability problem for certain parameters of multitype epidemic models. We identify designs that can overcome this problem for the important objective of estimating parameters that help to assess the effectiveness of a vaccine. With disease transmission in animal groups there is greater scope for conducting planned experiments and we explore some possibilities for such experiments. The topic is largely unexplored and numerous open research problems in the area of statistical design of infectious disease data are mentioned.

Keywords: Assessing a vaccine; Basic reproduction number; Disease transmission rates; Efficient designs; Epidemic models; Household outbreaks; Incomplete data; Partially observed process; Planned veterinary experiments

1. Can infectious disease studies be planned?

There has been relatively little emphasis on the design of studies of infectious diseases. One reason for this is that it is unethical to induce disease transmission in a human community and this limits the scope for conducting planned experiments. Furthermore, studies of disease transmission in a community are usually thought of as

* Correspondence address. National Center for Epidemiology and Population Health, The Australian National University, Canberra, ACT 0200, Australia. Tel.: +61-2-6125-2312; fax: +61-2-6125-0740.

E-mail address: n.becker@anu.edu.au (N.G. Becker).

0378-3758/01/$ - see front matter
PII: S0378-3758(00)00323-2

observational studies, because we typically acquire data by observing the course of an epidemic that has arisen naturally. However, we argue here that there are many important design questions in the study of infectious diseases. These questions are mainly concerned with determining the type of data and the size of the data set to be collected. It is our aim to point out such design problems and to illustrate some of them in detail, with the hope that this will encourage further work in the area.

A general feature that makes it important to design studies of infectious diseases is that it is usually not feasible to observe disease transmission over the entire community. While complete observation is the intention of surveillance systems for notifiable communicable diseases, surveillance registers nearly always suffer from severe under reporting, due to both noncompliance by medical officers and the occurrence of subclinical infections. In practice it is often only feasible to achieve 'complete' observation in a subset of the community. The important questions of 'which subset' and 'how large a subset' require planning. This opens up a rich class of problems, since we have different types of infectious diseases, different types of communities and variety in the type of data that can be collected.

One distinguishing feature of infectious disease data is that often only parts of the infection process and disease progression are observed. We usually do not know the time when an infection occurred, nor which contact caused the infection. Neither do we observe when an individual's infectious period begins or ends. A consequence of this partial observation is that the data are sometimes inadequate for estimating central parameters of the transmission model. Identifying a type of data set that can overcome such non-identifiability of parameters and determining how well they do this are important design problems.

The control of disease transmission is of primary interest, so it is not surprising that many studies are concerned with the assessment of vaccines as a means of preventing disease transmission. In particular, there is interest in testing whether a proposed vaccine provides protection against infection and in the estimation of vaccine efficacy. An estimate of vaccine efficacy is made from data on how many of the vaccinated individuals are infected and how many of the unvaccinated individuals are infected. It is therefore necessary to design the study so that there are groups of vaccinated and unvaccinated individuals who are exposed to similar forces of infection over the time period of observation. It is also necessary to determine the group sizes and the duration of the study required for effective inference about the vaccine efficacy.

The effectiveness of a vaccine can be measured in a number of ways. The most common interpretation is in terms of the protection it offers against infection, relative to an individual who is not vaccinated. However, when a vaccinated individual does get infected he/she sometimes gets a milder form of the disease, which may mean that he/she is less infectious than an infected individual who was not vaccinated. A proper assessment of the effectiveness of the vaccine in the community therefore requires estimates of the rates of disease transmission between an infective and a susceptible, where each of the two individuals could be vaccinated or not vaccinated. Therefore up to four different transmission rates need to be estimated. It is difficult to estimate these

rates from the available data because we generally do not observe who infects whom. We only observe who gets infected. This means that some parameters are not estimable unless we carefully plan what we are going to observe, with an eye for situations in which we have some information about who is likely to have infected whom. Data sets that ensure the estimability of such between and within group transmission rates are discussed in Sections 3.2 and 6.2.

While studies involving animals must be approved by an ethics committee, they are generally not constrained to the same degree as studies of disease transmission among humans. It is therefore possible to conduct planned experiments with animals, and this increases the range of possible studies substantially. For example, it is possible to set up a group of animals and begin an outbreak in this group at a known time. One can then monitor the infection process to determine the times of infection and monitor disease progression in each infected animal. We discuss some of the unique features of statistical planning for such studies in Section 7.

2. Introduction to disease transmission models

For readers unfamiliar with epidemic models we introduce two simple models that form the basis of our discussion.

2.1. Disease transmission in continuous time

The first is an SIR model that describes disease transmission in calendar time. An epidemic model is said to be of the SIR type if, with respect to the disease, individuals begin by being susceptible to infection and upon infection they immediately enter an infectious stage, which is followed by a recovered state where they remain having acquired immunity from further infection.

With respect to some time origin, let $S(t)$, $I(t)$ and $R(t)$ denote the number of individuals in the susceptible, infectious and recovered states at time t, respectively. The population is assumed to be closed so that $S(t)+I(t)+R(t)=n$ for all t. The chance of disease transmission is described by

$$\Pr\{S(t+\mathrm{d}t)=s-1,\ I(t+\mathrm{d}t)=i+1 \mid S(t)=s,\ I(t)=i\} \simeq \frac{\beta si}{n}\,\mathrm{d}t. \tag{1}$$

An underlying assumption is that the community consists of individuals who, with regard to disease transmission, are homogeneous and mix uniformly with each other. The parameter β is the rate at which an individual has 'close contact' with others, a proportion $S(t)/n$ being with susceptibles so $\beta I(t)S(t)/n$ is the rate at which the $I(t)$ infectious individuals have close contact with susceptibles. By 'close contact' is meant a contact which results in infection if the contacted individual is susceptible. An alternative way of viewing the parameter β is as a product of the actual contact rate

and the probability of disease transmission given a contact with a susceptible, but here we are only interested in the product of these quantities.

The infectious period has a duration Y with distribution function F, mean μ and variance σ^2. The infectious periods are assumed to be mutually independent. If Y has an exponential distribution we obtain a Markovian model, which has been studied extensively under the label *general epidemic* model.

For many diseases there follows a latent period after infection, during which the infectious agent develops inside the host, with no potential to transmit the disease. Models with a latent period are often called SEIR models, where E stands for 'exposed'.

2.2. Epidemic chain models

An epidemic chain tracks the spread of disease in terms of generations. The initial generation consists of the introductory infectives. Their direct contacts lead to the infectives of generation 1, who make contacts giving the cases of generation 2, and so on. An epidemic chain denoted by $i_0 \to i_1 \to \cdots \to i_r \to 0$ has i_0 initial infectives, i_1 infectives in generation 1, and so on until generation $r+1$ which contains no cases and the chain stops at that generation. The $\to 0$ at the end of the notation for the chain is often omitted with the understanding that the chain stops at that point.

An epidemic chain model describes the spread of the disease in calendar time only if the disease has a long latent period and a short infectious period, with relatively little variation, since then cases of the same generation are clustered together at the same calendar time and cases from different generations are separated in time.

A chain binomial model assumes that each susceptible individual has the same probability of being infected and the events of escaping infection are independent for different susceptibles. For generation t it is specified by

$$\Pr\{S(t)=s-i_t,\ I(t)=i_t \mid S(t-1)=s,\ I(t-1)=i\}=\binom{s}{i_t} p_i^{i_t} q_i^{s-i_t}, \tag{2}$$

with $p_i+q_i=1$. Here q_i is the probability that a susceptible individual escapes infection when exposed to the i infectives of one generation for the duration of their infectious period.

With $q_i=q^i$ for every i, model (2) gives the well-known Reed–Frost chain binomial model for disease transmission, while $q_i=q$ if $i>0$ and $q=1$ if $i=0$, gives the Greenwood model; see Bailey (1975, Chapter 14) and Becker (1989, Chapter 2) for details of these models. The probabilities $\{q_i\}$ could also be generalised to allow variation between individuals and/or households leading to random effects models described in Becker (1989, Chapter 3).

3. Sampling from a uniformly mixing population

In this section we treat a closed uniformly mixing population of size n, assumed to be fairly large. The spread of the infectious disease of interest is described by the SIR

model outlined in Section 2.1. We begin by assuming individuals are homogeneous and discuss the situation with heterogeneous individuals briefly in Section 3.2.

3.1. Homogeneous individuals

The parameter β denotes the contact rate per time unit and μ the mean duration of the infectious period, so $\theta = \beta\mu$ is the expected number of infections by one individual in a completely susceptible population. Often θ is referred to as the basic reproduction number and denoted R_0. Its value is of interest because the probability of a major outbreak is positive only when the product $\theta s_0 > 1$, where s_0 is the initial proportion of susceptible individuals in the population; (e.g. Ball, 1983). This means that major outbreaks can be prevented if s_0 is made sufficiently small. Specifically, the proportion of individuals that must be immunized to prevent major outbreaks is $1 - 1/\theta$, see Becker (1989, p. 8) for example. Clearly θ is a central parameter and we focus on its estimation. A general feature when making inference from one population is that a major outbreak, i.e., an epidemic, is necessary for consistent estimation. It is therefore assumed that a major outbreak has occurred in what follows. In applications this will be the case since a minor outbreak would usually not be detected.

Procedures for estimating θ have been considered for several types of data observed on the entire population; see Becker (1989), Rida (1991) and Becker and Hasofer (1997). However, it is expensive and time consuming to observe the whole population, making it relevant to consider inference procedures based on observations from a subset of the population. First we treat the estimation of θ when we collect data on a sample of individuals on two occasions, namely at a time before the epidemic season and at a time after the epidemic, when individuals infected during the epidemic have become immune. On each occasion every sampled individual is classified as 'susceptible' or 'immune'. A study of this type was conducted for the study of transmission of influenza A(H3N2) in Tecumseh, MI, see Addy et al. (1991) and references therein.

Assume that n_0 individuals are sampled before the epidemic season and that S_0 of them are found to be susceptible, while the remaining n_0–S_0 are immune and remain so over the epidemic season (here and in the sequel capital letters denote random variables). After the epidemic n_1 of the S_0 individuals who were susceptible at the first-sampling time are tested again and N_1 are found to have been infected during the epidemic season. Based on these data from the two samples we give estimates of $\theta = \beta\mu$, the basic reproduction number, and s_0, the proportion of the population that is initially susceptible.

The parameter s_0 is simply estimated by the corresponding sample proportion S_0/n_0. Similarly, the sample proportion $\hat{p} = N_1/n_1$ 'estimates' $\tilde{p} = N/s_0 n$, the proportion among the initially susceptible individuals who became infected. To come up with an estimate for θ we use a result from epidemic theory that relates θ to $\tilde{p}$. For a major outbreak, $\tilde{p}$ converges (for large n) to a normal distribution with mean p defined as the positive

solution to

$$1 - p = \exp(-\theta s_0 p) \tag{3}$$

and variance

$$\sigma_{\hat{p}}^2 = \frac{pq[1 + q(\theta s_0 \sigma/\mu)^2]}{s_0 n(1 - q\theta s_0)^2}, \tag{4}$$

where $q = 1 - p$ and, as before, μ and σ respectively denote the mean and standard deviation of the infectious period, see for example Ball (1983). We use (3) as an estimating equation for θ. The estimators for s_0 and θ are thus

$$\hat{s}_0 = \frac{S_0}{n_0} \quad \text{and} \quad \hat{\theta} = \frac{-\log(1 - \hat{p})}{\hat{s}_0 \hat{p}}. \tag{5}$$

Large sample inferences can now be based on the following result.

Theorem 1. *For the model defined above the estimator $\hat{s}_0$ is the unbiased ML-estimator for s_0, while $\hat{\theta}$ is asymptotically equivalent to the ML-estimator for θ, as n_0, n_1 and n tend to infinity. For the same limit, $(\hat{s}_0, \hat{\theta})$ has a bivariate normal distribution with mean (s_0, θ), variances σ_s^2 and σ_θ^2, and covariance $\sigma_{s,\theta}$ given by*

$$\sigma_s^2 = \frac{s_0(1 - s_0)}{n_0}\left(1 - \frac{n_0}{n}\right), \quad \sigma_{s,\theta} = -\frac{\theta(1 - s_0)}{s_0 n_0}\left(1 - \frac{n_0}{n}\right) \tag{6}$$

and

$$\sigma_\theta^2 = -\theta\sigma_{s,\theta} + \frac{1}{n_1}\frac{(1 - q\theta s_0)^2}{s_0^2 pq}\left(1 - \frac{n_1}{s_0 n}\right) + \frac{1}{s_0 n}\frac{1 + q(\theta s_0 \sigma/\mu)^2}{s_0^2 pq}. \tag{7}$$

The proof is outlined in the appendix.

The variances and covariance in (6) and (7) are estimated consistently by replacing the parameters by their estimates. However σ/μ, the coefficient of variation of the duration of the infectious period, must be known. All results rely on a major epidemic, which is only possible when $\theta s_0 > 1$, otherwise only few infections will occur and there is not enough information for consistent estimation. Whenever a positive fraction is infected in a large community the estimates will satisfy $\hat{\theta}\hat{s}_0 > 1$.

Remark. The variance and covariance estimates reduce to

$$\hat{\sigma}_s^2 = \frac{\hat{s}_0(1 - \hat{s}_0)}{n_0}, \quad \hat{\sigma}_{s,\theta} = -\frac{\hat{\theta}(1 - \hat{s}_0)}{n_0 \hat{s}_0} \text{ and } \hat{\sigma}_\theta^2 = -\hat{\theta}\hat{\sigma}_{s,\theta} + \frac{(1 - \hat{q}\hat{\theta}\hat{s}_0)^2}{n_1 \hat{s}_0^2 \hat{p}\hat{q}}, \tag{8}$$

when the sample sizes n_0, n_1 are small relative to the population size n. An advantage of these expressions is that we do not need to know the coefficient of variation σ/μ.

When designing a study, the sample sizes n_0 and n_1 should be chosen so that the estimators $\hat{s}_0$ and $\hat{\theta}$ have the desired precision according to (6) and (7), or (8) if appropriate, using some plausible values for unknown parameters (see Example 2).

Note that the parameters β and μ cannot be estimated separately from such data, only their product $\beta\mu = \theta$ can be estimated. This should not come as a surprise since the data provides no information about the development of the epidemic over time.

We now give two examples. The first example demonstrates that the proposed estimates from the sample data can indeed have a precision that is good enough to make the estimates useful.

Example 1. Suppose a sample of 200 individuals is drawn from a large community at a time prior to the epidemic season, and that 20 of them are found to be immune. After the epidemic season the 180 individuals who initially tested susceptible are tested again and it is found that 90 of them were infected since the first test. In the above notation $n_0 = 200$, $S_0 = 180$, $n_1 = 180$ and $N_1 = 90$. It is assumed that $n \gg 200$ so (8) may be used. This gives the estimates $\hat{s}_0 = 0.90$ (0.021), $\hat{p} = 0.50$ (0.037) and $\hat{\theta} = 1.540$ (0.065), the standard errors being given in parentheses. The critical immunity level $v_c = 1 - 1/\theta$ is estimated by $\hat{v}_c = 1 - 1/\hat{\theta} = 0.351$ and its standard error is $\hat{\sigma}_\theta/\hat{\theta}^2 = 0.027$, using the δ-method. The precision, as reflected by the standard errors, indicates that the estimates are clearly of practical value.

The next example illustrates how to determine the requisite sample size for a study aimed at estimating θ, or testing a hypothesis about θ.

Example 2. Assume that a vaccine is available which provides full protection against infection by the disease and that a vaccination coverage of 75% of all individuals is attainable in a large community. The proportion susceptible will then be 25%, making the reproduction number for the disease in the vaccinated community 0.25θ. Major epidemics are then prevented with probability 1 if $0.25\theta \leqslant 1$, so there is considerable interest in knowing whether or not the basic reproduction number θ exceeds 4. Consider therefore a study, of the type described above, that seeks to provide evidence that $\theta < 4$.

A sample of n_1 individuals is to be selected from those initially susceptible and these individuals are tested to determine whether they were infected during the last epidemic. Our task is to determine the sample size n_1. We set up the null hypothesis H_0: $\theta \geqslant 4$ and seek evidence against H_0. As an illustration, we choose n_1 by requiring the probability of rejecting the hypothesis H_0: $\theta \geqslant 4$, at the 5% significance level, to be at least 0.9 when the true value of θ is as low as 3. In other words, we want to determine the sample size required to give a power of 0.9 for the 5% significance test when θ is actually 3.

For simplicity suppose that s_0, the proportion susceptible before the last epidemic, is known and that the community is large relative to the sample. By Theorem 1 the estimator $\hat{\theta}$ is Gaussian around the true θ and with a variance which is estimated by $\hat{\sigma}_\theta^2 = (1 - \hat{q}\hat{\theta}s_0)^2/s_0^2\hat{p}\hat{q}n_1$. To test H_0 against the one-sided alternative H_A: $\theta < 4$ we reject H_0 when $\hat{\theta} < 4 - 1.645\hat{\sigma}_\theta$, where 1.645 is the 95% quantile of the standard

Normal distribution and 4 corresponds to $\theta = 4$, the least favourable θ under the null hypothesis. At the design stage we have no estimate for the standard deviation, so below we replace $\hat{\sigma}_\theta$ by σ_θ with θ set at the stipulated value 3. Our desired power is achieved when $0.9 = \Pr(\hat{\theta} < 4 - 1.645\hat{\sigma}_\theta \mid \theta = 3) \approx \Phi(1/\sigma_3 - 1.645)$, where $\Phi(\cdot)$ is the distribution function of the standard Normal distribution. This leads to the equation $1.282 = 1/\sigma_3 - 1.645$, or

$$n_1 = 2.927^2 \left[\frac{(1 - q\theta s_0)^2}{s_0^2 pq} \right]_{\theta=3}, \tag{9}$$

where the $p = 1 - q$ values are computed from (3) for the specified values of θ.

For example, in a completely susceptible population ($s_0 = 1$) the solution to (3) is found to be $p = 0.941$ when $\theta = 3$ giving the required sample size as $n_1 = 104$. On the other hand, when the proportion susceptible before the epidemic is only $s_0 = 0.5$ and $\theta = 3$, so the actual reproduction number is 0.5×3, then the solution to (3) is $p = 0.583$ if $\theta = 3$ implying that the required sample size is $n_1 = 20$ from (9). An explanation for the surprising result that a smaller sample suffices when some individuals are initially immune, i.e. $s_0 < 1$, goes as follows. For a given value of θ, the value of p in the estimating equation (3) increases with s_0. Also, the standard deviation σ_θ, which measures the precision of our estimate for θ, increases with p and hence with s_0, for fixed θ and subject to estimating equation (3). Both of these increases are substantial for p near 1, as is the case for $s_0 = 1$ and $\theta = 3$, our value of interest. It can be shown that $s_0 = 1$ maximizes n_1 in (9), so when s_0 is unknown the value of n_1 computed for $s_0 = 1$ is a safe choice.

Consider now the case with partial observation of the epidemic over time. In principle, it is possible to observe the time of diagnosis for each infected individual. As the infectivity has often decreased by the time of diagnosis, and since social activity is reduced when an individual show symptoms, such data are approximately equivalent to observing the removal processes over time.

Maximum likelihood estimation for θ and μ when removal times are observed for the whole population and assuming exponentially distributed infectious periods, is treated by Bailey (1975, Section 6.83) and applied to smallpox data from an outbreak in Abakaliki, Nigeria. On the other hand, Becker and Hasofer (1997) address inference for such data by using martingale theory to construct estimating functions. In the present discussion our interest is in knowing if data on the removal times of infected individuals in a subset of the population allows us to estimate these parameters, and if we can determine the size of the subset required to achieve adequate precision. The first task towards this end is to adapt the methods of estimation to data on removal times in a subset of the population, and this seems feasible. The task of determining the size of the subset required to achieve adequate precision is more likely to be manageable for the approach based on martingale estimating functions because explicit expressions for standard errors are then available. These are open problems of considerable interest.

3.2. Heterogeneous individuals

Even in a uniformly mixing population individuals may differ in aspects that affect disease transmission. For example, such differences can occur because the state of the immune system depends on factors such as age, gender and/or vaccination status. One way to treat this heterogeneity is to classify individuals into a few homogeneous sub-classes (types), and to let the infection rate between pairs of individuals, β/n, depend on the types of the pair. Inference procedures based on a sample of individuals for this situation are yet to be derived. Identifiability problems may sometimes occur as was observed by Britton (1998), who considers data from the entire community. For example, if only the initial and final proportions infected are observed for each type, then it is not even possible to estimate the basic reproduction number R_0, an important parameter when designing vaccination programs. One way of addressing this identifiability problem is to observe outbreaks in a set of households containing different combinations of types; this idea is illustrated in Section 6.

4. Sampling isolated household outbreaks

Data on disease incidence in households have a long history, because such data are relatively easy to acquire and the uniform-mixing assumption, which greatly simplifies analysis, seems plausible within households. In this section we consider analyses for such data based on models that assume the force of infection acting from outside the household is negligible relative to the infection intensity within the household, when there is an infective in the household. In other words, once infection enters the household, its outbreak is assumed to evolve independently of disease transmission occurring in the rest of the population. Bailey (1975) and Becker (1989) use models derived from such assumptions to analyse data on household outbreaks of measles and the common cold. The following discussion is presented with reference to households of size two and three to keep the algebraic expressions simple, but in practice it is of course important to consider larger households. The analysis presented here is readily generalised to larger households, although expressions become increasingly complicated as the household size increases.

4.1. Epidemic chain data

Suppose we have data on the epidemic chains of outbreaks of an infectious disease in households of size three. Assume that each outbreak begins with one of the three initial susceptible individuals being infected by a contact with someone from outside the household. Observations on n such outbreaks give n_1, n_{11}, n_{111} and n_{12} epidemic chains of type 1, $1 \to 1$, $1 \to 1 \to 1$ and $1 \to 2$, respectively. Suppose that an epidemic chain binomial model of Reed–Frost type (described in Section 2.2) is believed to describe

the outbreaks. When we assume that all individuals are homogeneous, with respect to infectivity and susceptibility, we obtain the probability distribution

Epidemic chain	1	$1 \to 1$	$1 \to 1 \to 1$	$1 \to 2$	Total
Probability	q^2	$2pq^2$	$2p^2q$	p^2	1
Frequency	n_1	n_{11}	n_{111}	n_{12}	n

for the epidemic chains, where $q = 1 - p$ is the probability that a susceptible individual escapes infection when exposed to one infective for the duration of the infectious period. Suppose our main interest lies in making statistical inference about the parameter p.

The log-likelihood for the chain data is

$$\ell_c(p) = \text{constant} + n_1 \log(q^2) + n_{11} \log(2pq^2) + n_{111} \log(2p^2q) + n_{12} \log(p^2),$$

where the constant term does not depend on the parameter p. From $\ell_c(p)$ we find the maximum likelihood estimator for p to be

$$\hat{p} = \frac{n_{11} + 2n_{111} + 2n_{12}}{2n + n_{11} + n_{111}},$$

which is simply the proportion of all exposures leading to infection. The large-sample variance of $\hat{p}$ is the reciprocal of the expected information

$$i_c(p) = 2n\left(4 + \frac{q^2}{p} + \frac{p^2}{q}\right) = \frac{2n}{pq}(1 + pq). \tag{10}$$

Information (10) may be interpreted in terms of information per exposure. Each exposures is a Bernoulli trial with expected information $1/pq$ about the parameter. The expected number of exposures generated by the chains is $2n(1 + pq)$. In contrast, for households with two susceptible individuals, of which one is infected from outside, the expected number of exposures is n. Therefore, while outbreaks in households of size three have only twice as many susceptibles at the start of the outbreak, the expected information contained in data on these outbreaks has *more* than doubled. This is because all outbreaks in households of size three have two initial exposures, but the chains $1 \to 1$ and $1 \to 1 \to 1$ also have one secondary exposure. The latter two chains have a cumulative probability of $2pq$, which explains the additional term $2npq$ in expected number of exposures. This indicates, in particular, that data on outbreaks in 20 households of size three with one introductory case gives a more precise estimate than data on outbreaks in 40 households of size two having one primary case. It is clear that considerable gains can be achieved by designing studies carefully with regard to sizes of households and number of households to be included in the study; see Example 3 for more details.

There is another point of interest. The estimate $\hat{p}$ is like the estimate of a binomial success probability, except that the number of trials, the exposures, is random. It is therefore necessary, when planning the size of the study to make allowance for the chance element in the number of exposures that might be realized. Consider a study consisting only of outbreaks in households of size three and we wish to determine how many outbreaks are needed to achieve a desired precision. Viewing $\hat{p}$ as a sample proportion we can use standard methods to determine n_e, the requisite number of exposures to achieve this precision. The next task is to determine how many household outbreaks are needed so that the probability of them containing at least n_e exposures is high. In other words, we want to find n such that

$$\Pr(2n + n_{11} + n_{111} \geqslant n_e) = 0.95, \quad \text{say}, \tag{11}$$

which we can do by using the large-sample distribution $N[2n(1+pq), 2npq(1-2pq)]$ for the number of exposures $2n + n_{11} + n_{111}$ (the variance is obtained from the chain probabilities given above). This distribution depends on p, which is unknown. To overcome this problem we set $p = \frac{1}{2}$, which gives a conservative value for n, i.e. it tends to be larger than necessary, since both the mean $2n(1+pq)$ and the variance $2npq(1-2pq)$ assume their largest values for $p = \frac{1}{2}$.

Example 3. In contrast to Example 2 suppose that estimation is the main objective, rather than hypothesis testing. We wish to determine the number of households required in the sample so that the (approximate) 95% confidence interval for p has width at most 0.2.

(a) When all households are of size two, including the initial infective, the total number of exposures is n, the number of households. The standard error for $\hat{p}$ is $\sqrt{\hat{p}\hat{q}/n}$, so that the width of the confidence interval is $2 \times 1.96\sqrt{\hat{p}\hat{q}/n}$. At the time of determining sample size we do not have an estimate $\hat{p}$ and we use the fact that the largest width occurs when $\hat{p} = \hat{q} = 1/2$. Accordingly we determine n by the requirement $2 \times 1.96/\sqrt{4n} \leqslant 0.2$. This gives the required sample size for outbreaks in households of size two as $n = 96$.

(b) Suppose now that we sample only households of size three, including the initial infective. The total number of exposures is then a random variable. We wish to determine n, the required number of households so that the width of the (approximate) confidence interval for p is at most 0.2. One way to proceed is to base the confidence interval on the large-sample variance given by the reciprocal of the information (10). That is, determine n by the smallest value such that $2 \times 1.96\sqrt{\hat{p}\hat{q}/(2n + 2n\hat{p}\hat{q})} \leqslant 0.2$. As $\hat{p}$ is unknown we use the largest width, which obtains for $\hat{p} = \hat{q} = 1/2$. This gives $n = 39$ for the required number of households of size three.

Using (10) amounts to replacing the random number exposures by the mean number of exposures, which does not allow for the possibility that we might, by chance, have considerably fewer exposures in our study. With 39 outbreaks it is possible to have as few as 78 exposures. There are strong arguments to support the claim that the precision quoted for estimates should correspond to the number of exposures actually

arising in the study, rather than the number expected on average in such studies. We might therefore prefer to ensure that the desired 96 exposures, as computed in part (a), will occur with high probability. Using (11) with $n_e = 96$ and $p = 0.5$ gives $n = 41$ as the required number of outbreaks in households of size three. This does not differ greatly from the 39 computed above, but it may differ more in other examples. Note that we have used $p = 0.5$ in these calculations and this choice of p gives the largest value for the required sample size.

4.2. Size of outbreak data

It is generally easier to determine the size of an outbreak than it is to determine which epidemic chain occurred. A relevant question is therefore: Is it worth the extra effort to determine which chain of infection occurred? To help answer this question we need to determine how much more informative epidemic chain data are compared with data on the size of outbreaks.

Let m_1, m_2 and m_3 denote the observed number of household outbreaks with 1, 2 and 3 eventual cases, respectively, when every household is of size three and each outbreak had one primary case. Expressed in the notation of epidemic chains, we observe only $m_1 = n_1$, $m_2 = n_{11}$ and $m_3 = n_{111} + n_{12}$. The log-likelihood corresponding to the final size data m_1, m_2 and m_3 is

$$\ell_s(p) = \text{constant} + m_1 \log(q^2) + m_2 \log(2pq^2) + m_3 \log(2p^2q + p^2).$$

From $\ell_s(p)$ we find the maximum likelihood estimator

$$\hat{p} = \frac{6m_1 + 11m_2 + 12m_3 - \sqrt{(6m_1 + 7m_2)^2 + 48(m_1 + m_2)m_3}}{4(2m_1 + 3m_2 + 3m_3)}. \tag{12}$$

This estimator cannot be easily intepreted in terms of expected number of exposures since the number of exposures realized is not observable. The corresponding expected information is

$$i_s(p) = 2n\left(4 + \frac{q^2}{p} + \frac{2p^2}{1+2q}\right), \tag{13}$$

and our interest is in comparing this with $i_c(p)$, given by (10). A comparison of the expression (13) with the middle term of (10) helps to make the difference apparent. A comparison in terms of $r(p) = i_s(p)/i_c(p)$ is also instructive, since this is a ratio of the large sample variances of the ML-estimators for p. Fig. 1 shows the graph on the relative information $r(p)$ against p.

We see that $r(p)$ decreases as p increases, but very gradually for $p \in (0, 0.5)$. Having the complete epidemic chains is important only when p, the probability of disease transmission within a household, is large. This is explained by noting that epidemic chain data is more informative only in that it gives the relative size of n_{111}

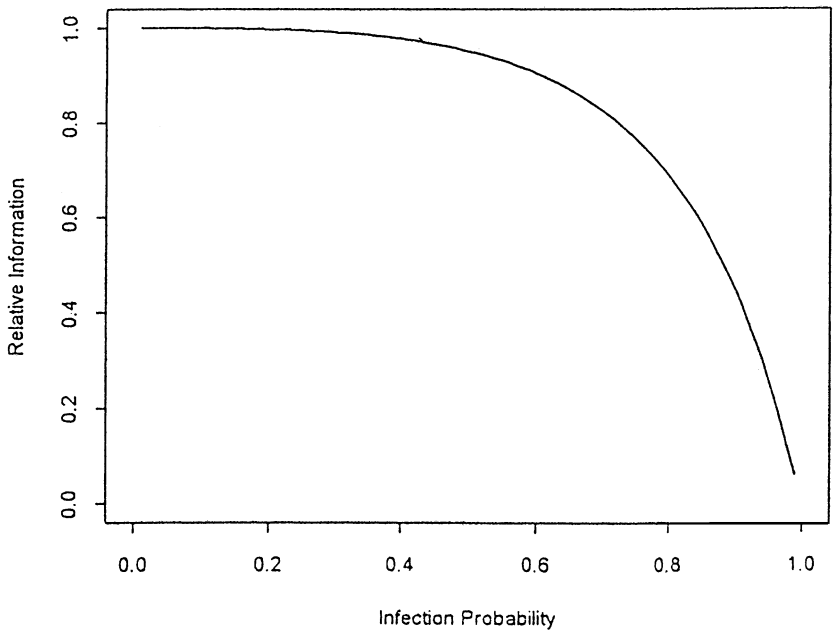

Fig. 1. Information in size of outbreak data relative to chain data.

to n_{12}, and the chains $1 \to 1 \to 1$ and $1 \to 2$ only occur with some frequency when p is large. Similar comparisons should be made for larger households.

4.3. Other study objectives

In the above discussion we made statistical inference about the parameter p the objective of the study. There are several other possible study objectives that might be used as the focus of the design of the study. For example, our main interest might be in testing the Reed–Frost assumption $q_i = q^i$. On the other hand we might wish to test the hypothesis that the parameter p is the same as we go from generation to generation in an epidemic chain. This hypothesis would tend to be rejected when susceptibility varies between individuals, because the more susceptible household members would tend to be infected in earlier generations. This indicates a large range of design problems that are worthy of further work.

5. Observing part of a community of households

The analyses described in Section 4 assume that for individuals in infected households the chance of acquiring infection from outside the household can be ignored. It is true that the chance of acquiring infection from a given infective member of the

household is usually much larger than the chance of acquiring it from a given infective not belonging to the household. However, in a large epidemic, there will be many infectives outside the household and the probability of making contact with at least one of them may not be negligible. Longini and Koopman (1982) propose an analysis, based on size of outbreak data for a sample of households, that allows disease transmission from outside the household in a simple and pragmatic way. We now consider this approach.

The type of study is as follows. A random sample of households is selected at a time prior to the epidemic season and every member of these households has a serological test to determine their status: susceptible or immune. After the epidemic season all individuals who were initially susceptible have a second test, to determine their status at that stage. In other words, we observe who has been infected since the first test. The advantages of making diagnoses via laboratory tests are that it enables subclinical cases to be detected and it ensures that case diagnosis is objective. In contrast to studies discussed in Section 4, the data may now include households in which no outbreak occurred.

The model used to analyse data from this study contains two parameters. The first parameter, p_b, is the probability of being infected from outside the household during the epidemic season, and individuals are infected from outside independently. The subscript 'b' indicates between-household transmission. The second parameter, p_w, is the probability of being infected by a given infected individual of the same household, and the events of being infected are independent for two separate individuals of the same household. The subscript 'w' indicates within-household transmission. The model can be described as the Reed–Frost model to which p_b, the probability of being infected by an external source, is added. Let p_{si} denote the probability that i of the s susceptibles present in the household at the start of the epidemic season are infected by the end of the epidemic season. These probabilities can be computed from the recursive formula

$$p_{si} = \binom{s}{i} p_{ii}(q_b q_w^i)^{s-i} \quad i = 0, 1, \ldots, s-1 \text{ and } p_{ss} = 1 - \sum_{i=0}^{s-1} p_{si}, \tag{14}$$

see Longini and Koopman (1982), where $q_b = 1 - p_b$ and $q_w = 1 - p_w$.

For example, for households of size two (susceptibles) we find

$$p_{20} = q_b^2, \quad p_{21} = 2p_b q_b q_w, \quad p_{22} = p_b^2 + 2p_b q_b p_w,$$

whereas for households of size three we find

$$p_{30} = q_b^3, \quad p_{31} = 3q_b^2 p_b q_w^2, \quad p_{32} = 3p_b q_b q_w^2(2p_w q_b + p_b),$$
$$p_{33} = 1 - p_{30} - p_{31} - p_{32}.$$

In this model p_b, the probability of getting infected from the *global* source of infection, is viewed as a parameter, i.e. an unknown constant. In reality it is the probability of getting infected by an infective who is not a household member, which depends on the size of the epidemic in the community. It is therefore more correct to view p_b as the realization of a random variable. However, its treatment as a parameter in

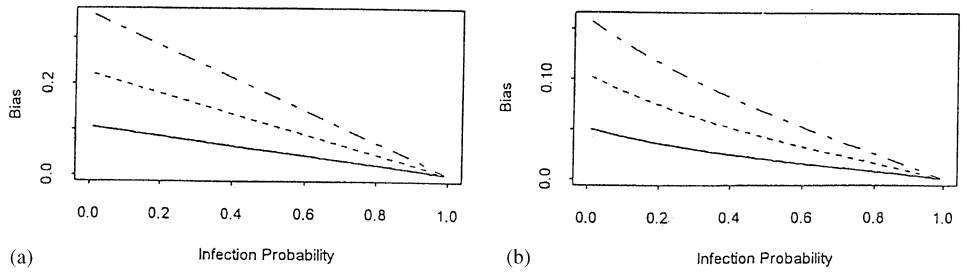

Fig. 2. Bias resulting when between-household transmission is ignored: (———) $p_b = 0.1$; (· · · · · · · · · · ·) $p_b = 0.2$; (— · — · —·) $p_b = 0.3$. (a) Household size 2, (b) Household size 3.

this model does provide a simple and pragmatic way of making allowance for the possibility of acquiring infection from outside the household when estimating p_w, the parameter of interest. Most of the information about p_b comes from knowledge about the households that are not infected, because we know that each of their members escaped the global force of infection which was the only force of infection to which they were exposed. Indeed, observing the households that escape infection is central to the effective estimation of p_b and p_w.

There are a number of design issues associated with this kind of study. They include determination of the requisite number of households in the sample and an assessment of which household sizes are best included for precise parameter estimation. Here we address two issues concerned with comparisons of the present study setting, where our sample may contain households that are not infected, with the setting of Section 4, where only infected households are sampled. For simplicity we illustrate these issues in the simple case where we have only households of size two and three. First, consider the bias that arises in the estimate of p_w when we focus on infected households and assume that the between-household transmission rate is negligible compared to the within-household rate.

Let h_{si} denote the number of households observed in which i of the s susceptibles present at the start of the epidemic season are infected by the end of the epidemic season. The model of Section 4 gives the ML-estimate $h_{21}/(h_{21} + h_{22})$ for q_w when we have only data on infected households of size two. Under the present model the expected value of this ML-estimate is approximately $p_{21}/(p_{21} + p_{22}) = 2q_w q_b/(1 + q_b)$. Therefore q_w is generally underestimated, but the bias is small when $p_b \approx 0$. The amount by which p_w is overestimated, on average, is $q_w p_b/(1 + q_b)$, which increases monotonically with p_b, and is shown in Fig. 2a as a function of p_w, for $p_b = 0.1$, 0.2 and 0.3. The bias can be substantial when p_w is small.

When we have only data on infected households of size three the ML-estimate of p_w is given by (12). This estimate has a large-sample expectation given by

$$\frac{6p_{31} + 11p_{32} + 12p_{33} - \sqrt{(6p_{31} + 7p_{32})^2 + 48(p_{31} + p_{32})p_{33}}}{4(2p_{31} + 3p_{32} + 3p_{33})}.$$

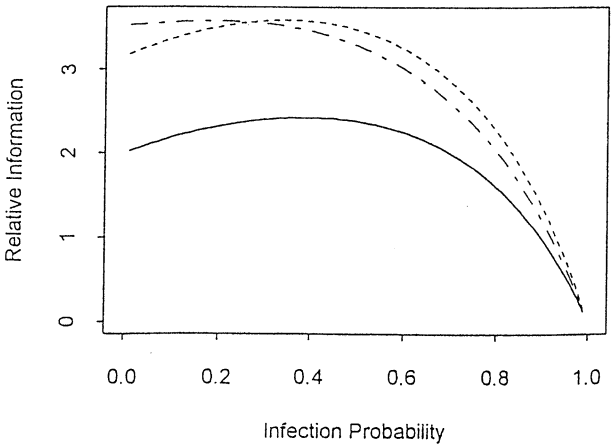

Fig. 3. Information in households of size three relative to that of size 2, (———) between-household transmission is ignored: (· · · · · · · · · · · ·) $p_b = 0.1$; (— · — · —·) $p_b = 0.3$.

Subtracting p_w from this gives the bias, which is shown in Fig. 2b as a function of p_w, for $p_b = 0.1$, 0.2 and 0.3. Again, the bias is larger for small values of p_w, but it is substantially smaller than for households of size 2.

The parameter of interest is p_w. In Section 4 we found that, for data on epidemic chains starting with a single primary case, the information about p_w more than doubled for household of size three compared with households of size two. Let us see if this remains true for data on household outbreaks when we allow for transmission from outside the household. We simplify the discussion by assuming that p_b is known. From the log-likelihood function

$$\ell(p_w|p_b) = \text{constant} + \sum_{s,i} h_{si} \log p_{si}, \quad p_w \in [0,1],$$

the expected information about p_w in n households of size s is computed to be

$$i_s(p_w|p_b) = n \sum_{i=1}^{s} \frac{(\partial p_{si}/\partial p_w)^2}{p_{si}}, \quad p_w \in [0,1].$$

Fig. 3 compares the graphs of $r(p_w|p_b) = i_3(p_w|p_b)/i_2(p_w|p_b)$, as functions of p_w, for $p_b = 0.1$ and 0.3, with the corresponding ratio for data on infected households and based on the probability model of Section 4. We find that the information in data on outbreaks in households of size three relative to outbreaks in households of size two is high for values of p_w as high as 0.7, but declines rapidly beyond that. In fact there is less information in households of size three for values of $p_w > 0.9$. This is in contrast to what we found for epidemic chain data in Section 4. It occurs for size of outbreak data because complete infection is most likely in infected households of size three when p_w is near 1. As a consequence, without the benefit of observing the individual epidemic chains, data on outbreaks in households of size three are not so well suited for distinguishing values of p_w near 1.

The relative information in households of size three is substantially higher when we allow for disease transmission from outside the household, but it does not depend greatly on the value of p_b.

6. Effectiveness of a vaccine

When the population consists of some individuals who are vaccinated and others who are not we simply have different types of individual. The comments in Section 3.2 are relevant, but the case where heterogeneity is due to vaccination warrants separate discussion for two reasons. Firstly, vaccine studies are easily the most common type of studies associated with infectious diseases. Secondly, we now have additional information about the types of individuals and a more specific purpose. The purpose is often to assess the effectiveness of the vaccine for both protecting an individual against infection and restricting the spread of the disease through the community. This is necessary because vaccines are often not fully effective, so that vaccinated individuals can still get infected, albeit at a lower rate. Furthermore, when infected, a vaccinated individual may react to infection less severely than an unvaccinated individual, and may consequently have less potential to transmit the disease to others.

Here we focus on a parametric specification of transmission rates, known as proportionate mixing, to obtain parameters that directly reflect the reduction in susceptibility induced by the vaccine and the reduction in infectivity it induces.

6.1. Observations from a uniformly mixing population

Consider a single fairly large community and suppose that, with respect to disease transmission, individuals differ only in that some are vaccinated and others are not. Every infected individual recovers after an infectious period and is then permanently immune from further infection. Let us label individuals v or u depending on whether they are vaccinated or not, respectively. We stipulate a disease transmission rate between a given infectious unvaccinated individual and a given susceptible unvaccinated individual of $\alpha_u\beta_u$, and similarly $\alpha_u\beta_v$, $\alpha_v\beta_u$ and $\alpha_v\beta_v$ for the transmission rates between the three remaining possible pairs. The model is over-parameterized since $\alpha\beta = \alpha'\beta'$, with $\alpha' = c\alpha$ and $\beta' = \beta/c$, for any positive c. A constraint needs to be imposed on the four parameters to overcome this identifiability problem. In the interest of having parameters with direct epidemiological interpretations it is useful to write the rates λ, λr_s, λr_i and $\lambda r_i r_s$, respectively. Our main interest lies in $r_i = \alpha_v/\alpha_u$ and $r_s = \beta_v/\beta_u$, the proportionate reduction in infectivity and susceptibility, respectively, for a vaccinated individual relative to an unvaccinated one. The parameter r_s is closely linked to the concept known as *vaccine efficacy*; see Halloran et al. (1992). However, r_i is also important for specifying how effective the vaccine is for controlling the spread of disease. Suppose vaccinated and unvaccinated individuals mix uniformly and we wish to make inferences about r_i and r_s from data on an epidemic in this large community.

Britton (1998) treats methods for making such inference for various types of data from the whole community, using estimating equations derived from martingale theory. The relative susceptibility r_s has the simple estimator $\hat{r}_s = \log(1 - \tilde{p}_v)/\log(1 - \tilde{p}_u)$, where $\tilde{p}_u$ is the community proportion infected of those unvaccinated and initially susceptible and $\tilde{p}_v$ is the community proportion infected among the vaccinated susceptible individuals. The relative infectivity, r_i, is not estimable when only the eventual numbers infected in the two groups are observed. If, on the other hand, the infection and/or removal processes of the two sub-populations are observed continuously over time, then an estimator for r_i is available (Britton, 1998) and the estimator is consistent except for the rather unlikely situation where vaccination reduces infectivity but does not alter the susceptibility.

Estimation procedures based on samples of the community have so far not been addressed in the literature. Of course, point estimates for r_i and r_s can be obtained from Britton (1998) only replacing community proportions by sample proportions. However, the uncertainty of such estimates, important when determining sample sizes of a design, is not known and deserves to be studied.

The difficulty with estimating the relative infectivity r_i in certain circumstances stems from not knowing who infects whom. It is therefore useful to look for situations which contain information about who is likely to have infected whom. Such information is contained in data on outbreaks in small groups, perhaps households, where the groups are comprised of varying numbers of vaccinated and unvaccinated individuals. We now illustrate this possibility in the simple setting where we have data on outbreaks in matched pairs of individuals residing in the same dwelling.

6.2. Observations from matched pairs

Consider a study of disease transmission in pairs of individuals, for example sibling studies with parents assumed to be immune. The pairs could include vaccinated and/or unvaccinated individuals and the design problem lies in determining how many pairs should have both unvaccinated, how many with both vaccinated and how many pairs with one of each. The objective of the study is to assess the effectiveness of the vaccine by obtaining estimates of r_s and r_i, where these parameters have the same interpretation as in Section 6.1.

Recall that individuals are labeled v or u depending on whether they are vaccinated or not. We need the probabilities

$$\Pr(\text{individual } y \text{ escapes infection by partner } x) = q_{\mathrm{w}xy} \quad \text{and}$$

$$\Pr(x \text{ escapes infection from outside during the study period}) = q_{\mathrm{b}x},$$

where $x, y \in \{\mathrm{u}, \mathrm{v}\}$. Label a pair ij if it consists of i vaccinated and j unvaccinated individuals. The number of pairs of type ij in the study is n_{ij}. Label an outbreak $k\ell$ if eventually k vaccinated and ℓ unvaccinated individuals are infected in the pair. The number of observed outbreaks of type $k\ell$ arising in pairs of type ij is $n_{ij,k\ell}$. Then a model, which generalizes that of Longini and Koopman (1982) to two types

of individual and is a particular case of the model defined by Addy et al. (1991), is given by

Pair	Outbreak	Frequency	Probability
20	00	$n_{20,00}$	q_{bv}^2
20	10	$n_{20,10}$	$2p_{\mathrm{bv}}q_{\mathrm{bv}}q_{\mathrm{wvv}}$
20	20	$n_{20,20}$	$p_{\mathrm{bv}}^2 + 2p_{\mathrm{bv}}q_{\mathrm{bv}}\,p_{\mathrm{wvv}}$
11	00	$n_{11,00}$	$q_{\mathrm{bv}}q_{\mathrm{bu}}$
11	10	$n_{11,10}$	$p_{\mathrm{bv}}q_{\mathrm{bu}}q_{\mathrm{wvu}}$
11	01	$n_{11,01}$	$p_{\mathrm{bu}}q_{\mathrm{bv}}q_{\mathrm{wuv}}$
11	11	$n_{11,11}$	$p_{\mathrm{bu}}\,p_{\mathrm{bv}} + p_{\mathrm{bu}}q_{\mathrm{bv}}\,p_{\mathrm{wuv}} + p_{\mathrm{bv}}q_{\mathrm{bu}}\,p_{\mathrm{wvu}}$
02	00	$n_{02,00}$	q_{bu}^2
02	01	$n_{02,01}$	$2p_{\mathrm{bu}}q_{\mathrm{bu}}q_{\mathrm{wuu}}$
02	02	$n_{02,02}$	$p_{\mathrm{bu}}^2 + 2p_{\mathrm{bu}}q_{\mathrm{bu}}\,p_{\mathrm{wuu}}$

Our objective is to estimate the relative infectivity r_{i} and the relative susceptibility r_{s}. To this end we introduce the parameterization $q_{\mathrm{bu}} = \mathrm{e}^{-\theta_{\mathrm{b}}}$, $q_{\mathrm{bv}} = \mathrm{e}^{-\theta_{\mathrm{b}}r_{\mathrm{s}}}$, $q_{\mathrm{wuu}} = \mathrm{e}^{-\theta_{\mathrm{w}}}$, $q_{\mathrm{wvu}}=\mathrm{e}^{-\theta_{\mathrm{w}}r_{\mathrm{i}}}$, $q_{\mathrm{wuv}}=\mathrm{e}^{-\theta_{\mathrm{w}}r_{\mathrm{s}}}$ and $q_{\mathrm{wvv}}=\mathrm{e}^{-\theta_{\mathrm{w}}r_{\mathrm{i}}r_{\mathrm{s}}}$. This gives a smaller class of models, but it is one which retains the desirable property of being expressed in terms of parameters with clear epidemiological interpretations. Briefly, the parameter θ_{w} depends on the rate of disease transmission within pairs and on the duration of the infectious period, whereas the parameter θ_{b} depends on the duration of the study period and on the global force of infection assumed to be acting on each individual during that period. The way r_{i} and r_{s} have been introduced assumes that the reduction in susceptibility is the same within and between households and that the reduction in infectivity and susceptibility acts multiplicatively on the force of infection.

Note that the observations $n_{20,00}$, $n_{11,00}$ and $n_{02,00}$ contain no information about the parameters θ_{w} and r_{i}, but they contain most of the information about the parameters θ_{b} and r_{s}. This is seen by noting that they are observations on independent $\mathrm{Bin}(n_{20}, \mathrm{e}^{-2\theta_{\mathrm{b}}r_{\mathrm{s}}})$, $\mathrm{Bin}(n_{11}, \mathrm{e}^{-\theta_{\mathrm{b}}(r_{\mathrm{s}}+1)})$ and $\mathrm{Bin}(n_{02}, \mathrm{e}^{-2\theta_{\mathrm{b}}})$ random variables, respectively. This provides a basis for determining the number of pairs required for precise estimates of θ_{b} and r_{s}. It is worth noting that the n_{11} pairs of type 11 are not sufficient, by themselves, to estimate all four parameters, because the four possible outcomes in such pairs have only three degrees of freedom.

As uninfected pairs provide no information about θ_{w} and r_{i}, it is necessary to ensure that the number of infected pairs is adequate to provide precise estimates of θ_{w} and r_{i}. When the infectivity of the disease is such that a large fraction of pairs escapes infection, we may need a huge number of pairs in our sample to guarantee an adequate number of infected pairs. It might be impractical to conduct such a large study. In that situation we should determine the number of pairs required in the sample to estimate

θ_b and r_s with adequate precision, and then add a separate sample of infected pairs to boost the precision of estimates of θ_w and r_i. Of course, an outcome for one of the additional infected pairs contributes to the estimation of θ_w and r_i via the conditional probability of the outcome given that it is an infected pair.

The use of standard methods to determine the number of pairs required to achieve adequate precision in estimates of four parameters is tedious and it is preferable to use an approximate indirect argument. We will be guided in the determination of sample size by making calculations under the assumption that for each infection it can be determined whether it resulted from a within pair contact, or not. Then, for example, outbreaks of type 20 for households of type 20 can be divided into those where both partners were infected by the external force of infection and the rest. Similarly the outbreaks of type 11 in households of type 11 can be broken up into three sub-types and outbreaks of type 02 in households of type 02 can be broken up into two sub-types. The consequence is that we then have count data for 14 types of outbreak, instead of 10. There is then just one probability term for each outbreak, i.e. none of the probabilities is given by a sum of terms. As a result, maximum likelihood estimation for the parameters q_{bu}, q_{bv}, q_{wuu}, q_{wuv}, q_{wvu} and q_{wvv}, ignoring their dependence on θ_b, θ_w, r_i and r_s, is just like estimating parameters of the binomial distribution, and is therefore straightforward. We exploit this observation by noting that the relationships $r_s=\ln(q_{bv})/\ln(q_{bu})$, $r_i=\ln(q_{wvu})/\ln(q_{wuu})$ and $r_i=\ln(q_{wvv})/\ln(q_{wuv})$ indicate that precise estimation of q_{bu}, q_{bv}, q_{wuu}, q_{wuv}, q_{wvu} and q_{wvv} leads to precise estimation of r_s and r_i.

For the augmented data set the maximum likelihood estimates of q_{bu} and q_{bv} are binomial proportions with $2n_{02}+n_{11}$ and $2n_{20}+n_{11}$ trials, respectively. Similarly, estimates for q_{wuu}, q_{wuv}, q_{wvu} and q_{wvv} are proportions with the expected number of exposures given by $2n_{02}\,p_{bu}q_{bu}$, $n_{11}\,p_{bu}q_{bv}$, $n_{11}\,p_{bv}q_{bu}$ and $2n_{20}\,p_{bv}q_{bv}$, respectively. Then the large sample standard deviation for each estimate of an 'escape infection' probability q is given by $\sqrt{pq/(\text{expected number of exposures})}$. We suggest that suitable values for n_{20}, n_{11} and n_{20} are determined by requiring that $\text{s.e.}(\hat{q})\leqslant\varepsilon$ for every $\hat{q}$, for some small positive ε.

Example 4. As an illustration of the above method for obtaining guidance on suitable values for n_{20}, n_{11} and n_{20} we choose $\varepsilon=0.1$ for each $\hat{q}$. First note that

$$\text{s.e.}(\hat{q}_{bu})=\sqrt{\frac{p_{bu}q_{bu}}{2n_{02}+n_{11}}}\leqslant\frac{1}{2\sqrt{2n_{02}+n_{11}}}\leqslant 0.1$$

leads to $2n_{02}+n_{11}\geqslant 25$. Similarly, $\text{s.e.}(\hat{q}_{bv})\leqslant 0.1$ leads to $2n_{20}+n_{11}\geqslant 25$. Next, note that

$$\text{s.e.}(\hat{q}_{wuu})=\sqrt{\frac{p_{wuu}q_{wuu}}{2n_{02}\,p_{bu}q_{bu}}}\leqslant\frac{1}{2\sqrt{2n_{02}\,p_{bu}q_{bu}}}\leqslant 0.1$$

leads to $n_{02}\geqslant 12.5/(p_{bu}q_{bu})$. Similarly, by requiring $\text{s.e.}(\hat{q}_{wvv})$, $\text{s.e.}(\hat{q}_{wuv})$ and $\text{s.e.}(\hat{q}_{wvu})$ to be 0.1, or less, we find $n_{20}\geqslant 12.5/(p_{bv}q_{bv})$, $n_{11}\geqslant 25/(p_{bu}q_{bv})$ and $n_{11}\geqslant 25/(p_{bv}q_{bu})$, respectively. It is now necessary to substitute plausible values, or bounds, for q_{bu} and

q_{bv}. For example, if $q_{bu} = 0.7$ and $q_{bv} = 0.9$ are considered appropriate values, then we obtain the inequalities

$$n_{02} \geqslant 60, \quad n_{20} \geqslant 139, \quad n_{11} \geqslant 93 \text{ and } n_{11} \geqslant 358.$$

In practice the value $n_{11} = 358$ is likely to be unacceptably large. This value of n_{11} is needed for the precise estimation of q_{wvu}. However, while $\hat{q}_{wvu}$ contributes to the estimation of r_i, the relationship $r_i = \ln(q_{wvv})/\ln(q_{wuv})$ reveals that it is enough to have precise estimates of q_{wvv} and q_{wuv}. We deduce that the sample sizes $n_{02} = 60$, $n_{20} = 139$ and $n_{11} = 93$ enable precise estimation of θ_b, θ_w, r_i and r_s.

7. Planning veterinary experiments

There is great interest in controlling infectious diseases in animal populations (e.g. Bouma et al., 1997) and here there is scope for controlled experiments. It is possible to determine both the amount and the type of data to be observed. For example, we can plan to have a closed community of a given size and can decide on the number of primary cases and fix the time of their (induced) infection. We can then regularly check the animals to determine, at least approximately, the times of their infection and monitor disease progression in infected individuals. We might then ask: How many animals should be used in the experiment? How should the animals be grouped? What details of disease progress should be observed? How should different types, perhaps distinguished by breed or vaccination status, be combined into subgroups?

In Section 7.1 we look at a specific design problem involving estimation of infection probabilities between two types of animal. In Section 7.2 we discuss some open problems.

7.1. Designs with two types of individual

Suppose we want to design a study for efficient estimation of the four different infection probabilities between two types of individual. At our disposal we have n animals, $n/2$ of each type say. The design question is to decide how the animals should be grouped into smaller isolated units and which animals should be chosen as primary cases. In particular we compare efficiency when animals are grouped into isolated pairs to the case when there are three animals in each group. It is shown that the latter design is generally more efficient although the resulting analysis is more complicated (a similar result for the case with homogeneous individuals was derived in Section 4). It is not necessarily true, however, that more is gained by using even larger groups: the analysis becomes very complicated and the parameter estimates become more and more confounded.

Assume that infection spreads according to a Reed–Frost-type model. Label the two types of individual a and b, and let $q_{xy} = 1 - p_{xy}$ denote the probability that a susceptible x-type escapes infection from an infected y-type of the same unit, where $x, y \in \{a, b\}$.

Here we permit four distinct transmission rates, in contrast to the proportionate mixing assumption used earlier. As the units are isolated there are no probabilities for between group transmission.

7.1.1. Pair-design

We begin with the design consisting of pairs of animals, one initially infectious and one susceptible. Let n_{ab} denote the number of pairs initially consisting of a susceptible a-type and an infectious b-type, and let Y_{ab} denote the number of these pairs in which the susceptible animal becomes infected. Different pairs are independent because they are isolated so $Y_{ab} \sim \mathrm{Bin}(n_{ab}, p_{ab})$. Similar results hold for the pair combinations aa, ba and bb. For simplicity we identify aa, ab, ba and bb with 1, 2, 3 and 4, respectively. It is easily shown that ML-estimates, variances and covariances are given by

$$\hat{p}_i^{(2)} = \frac{y_i}{n_i}, \quad \mathrm{Var}(\hat{p}_i^{(2)}) = \frac{p_i(1-p_i)}{n_i} \quad \text{and} \quad \mathrm{Cov}(\hat{p}_i^{(2)}, \hat{p}_j^{(2)}) = 0. \tag{15}$$

The variance is estimated consistently when p_i is replaced by $\hat{p}_i^{(2)}$. Prior information about the parameters and the desired precision for estimators influence the number of pairs of each type that should be used in the design. Without prior knowledge or precision preferences there should be equally many pairs of each type, i.e. $n_1 = \cdots = n_4 = n/8$.

7.1.2. Triplet design

Suppose now that each group initially is comprised of two susceptible animals and one newly infected animal. There are six possible ways to arrange such a triplet. These can be denoted $a20$, $b20$, $a02$, $b02$, $a11$ and $b11$, where the letter indicates the type of the initial infective and the digits indicate the initial numbers of susceptible a- and b-types, respectively. We only consider the first four types of triplets because the outcome probabilities of the remaining two are more complicated. To simplify notation we label these triplets 1, 2, 3 and 4, respectively and, as in Section 7.1.1, the probabilities p_{aa}, p_{ab}, p_{ba} and p_{bb} are denoted p_1, p_2, p_3 and p_4.

Let n_k denote the number of triplets of type k and $X_k(i)$ the random number of k-type triplets with i infected at the end. Then $(X_k(0), X_k(1), X_k(2))$ has a trinomial distribution so the log-likelihood is

$$\ell(p_1, p_2, p_3, p_4) = \text{constant} + \sum_{k=1}^{4} \sum_{i=0}^{2} X_k(i) \log [P_k(i)]. \tag{16}$$

The probabilities $P_k(i)$ of ending up with i infected are obtained in terms of $p_1, \ldots, p_4$ by enumerating the possible epidemic chains and accumulating the associated probabilities, as in Section 4.1. For example, $P_1(0) = (1-p_1)^2$, $P_1(1) = 2p_1(1-p_1)^2$ and $P_1(2) = 1 - P_1(0) - P_1(1)$, and for $P_2(\cdot)$ we have $P_2(0) = (1-p_2)^2$, $P_2(1) = 2p_2(1-p_2)(1-p_1)$ and $P_2(2) = 1 - P_2(0) - P_2(1)$. Thus, $P_1(\cdot)$ and $P_2(\cdot)$ only depend on p_1 and p_2. Similarly $P_3(\cdot)$ and $P_4(\cdot)$ only depend on p_3 and p_4, so these pairs will have

estimators independent of each other. By symmetry the estimator for p_4 will have the same form as the estimator for p_1 and the p_3 estimator has the same form as the p_2 estimator, only replacing the indices $k=1$ by 4 and $k=2$ by 3 throughout. Only estimation procedures for p_1 and p_2 are hence treated below.

The ML-estimates, derived by equating the partial derivatives of the log-likelihood (16) to zero, have no closed form. However, straightforward calculations on (16) lead to the expected (symmetric) information matrix I with elements

$$i_{11}(p_1,p_2)=n_1\frac{2(3+4p_1-p_1^2)}{p_1(3-2p_1)}+n_2\frac{2p_2q_2(2-p_2)}{q_1(2p_1q_2+p_2)}, \tag{17}$$

$$i_{22}(p_1,p_2)=n_2\frac{2(p_1q_2+2q_1p_2)}{p_2q_2(2p_1q_2+p_2)} \quad \text{and} \quad i_{12}(p_1,p_2)=n_2\frac{2p_2}{2p_1q_2+p_2}. \tag{18}$$

The large sample variance matrix for the ML-estimates $(\hat{p}_1^{(3)},\hat{p}_2^{(3)})$ is the inverse of the 2×2 matrix I with elements defined above.

In order to specify the design we have to specify the number of each type of triplet in the study, that is, assign values to $n_1,\ldots,n_4$. Without prior knowledge about the parameters and parameters considered equally important, $n_1,\ldots,n_4$ should be chosen to make the variances of $\hat{p}_1^{(3)},\ldots,\hat{p}_4^{(3)}$ equal when the corresponding true values are equal. Symmetry arguments then imply that $n_1=n_4$ and $n_2=n_3$ and the relation between n_1 and n_2 (and hence between n_4 and n_3) is obtained by replacing p_1 and p_2 in (17) and (18) by a common value p and solving the equation $i_{11}(p,p)=i_{22}(p,p)$. It can be shown that this implies $n_2=n_1(3+4p-p^2)/(3-2p+p^2)$, a relation which depends on p. If p is small there should be approximately equally many of the two triplets, but if p is large up to three times more $b20$ triplets than $a20$ triplets are needed. This is intuitively clear because when p is small at most one contact will occur in a triplet, so $b20$ triplets only carry information about p_{ab} and this information is identical to the information on p_{aa} from an $a20$ triplet. On the other hand, if p is large there will be $a\to a$ contacts in some $b20$ triplets, so $b20$ triplets also carry information on p_{aa}. Since $a20$ only carry information on p_{aa} less of these triplets are required for the total information on p_{aa} and p_{ab} to be equal.

7.1.3. Comparison between pair and triplet designs

If the true parameters are of the same size p and equally many $(=n/8)$ of each pair-type are used, then the estimators from the pair design, $\{\hat{p}_i^{(2)}\}$ given by (15), have variances and covariances

$$V^{(2)}(p)=\frac{8p(1-p)}{n} \quad \text{and} \quad C^{(2)}(p)=0. \tag{19}$$

For the triplet-design the estimators are implicit solutions to ML-estimating equations. The estimators $(\hat{p}_1^{(3)},\hat{p}_2^{(3)})$ and $(\hat{p}_4^{(3)},\hat{p}_3^{(3)})$ are independent of each other and the variance matrix of $(\hat{p}_1^{(3)},\hat{p}_2^{(3)})$ is the inverse of information matrix I defined by (17) and (18), and similarly for $(\hat{p}_4^{(3)},\hat{p}_3^{(3)})$. If the true parameters are of the same size p and

the number of each type of triplet is chosen as discussed in the previous subsection, then all variances are equal, as are the two non-zero covariances $\mathrm{Cov}(\hat{p}_1^{(3)}, \hat{p}_2^{(3)})$ and $\mathrm{Cov}(\hat{p}_4^{(3)}, \hat{p}_3^{(3)})$. These variances and covariances are

$$V^{(3)}(p) = \frac{1}{n}\frac{18p(3-2p)}{9+9p-7p^2+p^3} \quad \text{and} \quad C^{(3)}(p) = -\frac{p}{3}V^{(3)}(p). \tag{20}$$

It is easily shown that $V^{(2)}(p) > V^{(3)}(p)$ for $0 \leqslant p \leqslant 0.71$. For example, if $p = 1/2$ then $V^{(2)}(0.5) = 2/n$ whereas $V^{(3)}(0.5) = 1.51/n$, i.e. a variance reduction of approximately 25%, and if $p = 1/4$ the variance reduction is 31%. The triplet-design is hence more efficient than the pair-design, at least when the disease of interest is not thought to be highly infectious. Note that the comparison is based on assuming that n, the total number of animals used, is the same in the two designs.

7.2. Some open problems in veterinary experiments

There is a wide range of open-design problems for veterinary experiments. Here we briefly mention a few.

Suppose the spread of a disease within a group of k animals is to be studied. Assume for simplicity that the individuals are homogeneous and that the disease spreads according to the Reed–Frost model with parameter p, see Section 2.2. A relevant question is: How many of the animals should be initially infected for efficient estimation of p? In Bouma et al. (1997) for example, experiments with $k = 10$ pigs are performed in which it was decided to infect 5 pigs at the start of the study. As discussed in Section 4, the precision of inference on p depends on the potential number of infective-to-susceptible contacts, so we want this number to be large. If few individuals are initially infected further infections may not occur resulting in few potential contacts. On the other hand, with many initial infectives we have fewer susceptibles who can accumulate exposures. The optimal balance between these two competing phenomena, a balance which will depend on the actual value of p, is not yet available.

There is scope to observe the epidemic process in greater detail in a controlled experiment. Besides knowing the time of induced infection for the primary cases it might be possible to observe, at least approximately, the time of infection for other infected animals, and the time of first symptoms for each infected animal. In particular, with such data for groups of size two, it seems feasible to estimate the transmission probabilities, the distribution of the infectious period and the relative infectivity during the infectious period. Specifically, a latency period can be detected from such an analysis.

Estimation of the basic reproduction number R_0 is another important problem in veterinary experiments. However, inference for this parameter might not be transferable if the experimental environment differs substantially from the natural environment. For precise estimates it is necessary to know the population density and other relevant components of the natural habitat.

Acknowledgements

The main part of the work was carried out while Tom Britton visited the School of Statistical Science at La Trobe University. He is grateful for the hospitality of the School, and to The Swedish Natural Science Research Council for financial support. Niels Becker gratefully acknowledges support from the Australian Research Council.

Appendix. Proof of Theorem 1

The estimator $\hat{s}_0$ of the proportion s_0 is based on a simple random sample from a finite population, and the results relating to it are well known.

A similar result holds for $\hat{p}$ as an estimator for $\tilde{p}$, but we need to acknowledge that $\tilde{p}$ is determined by the outcome of an epidemic, and is therefore a random variable whose distribution depends on θ and s_0. As mentioned, $\tilde{p}$ converges in probability to p, defined in (3), as $n \to \infty$, so asymptotically $\hat{p}$ is also the ML-estimate of p. Again by (3), $\theta = -\log q/(s_0 p)$ from which it follows that $\hat{\theta}$ is consistent and asymptotically equivalent to the ML-estimator. Applying the δ-method to $\hat{\theta}$ and using the independence of $\hat{s}_0$ and $\hat{p}$ we obtain

$$\sigma_\theta^2 = \left(\frac{\log q}{s_0^2 p}\right)^2 \sigma_s^2 + \left(\frac{1}{s_0 pq} + \frac{\log q}{s_0 p^2}\right)^2 \sigma_{\tilde{p}}^2 = \left(\frac{\theta}{s_0}\right)^2 \sigma_s^2 + \left(\frac{1 - q\theta s_0}{s_0 pq}\right)^2 \sigma_{\tilde{p}}^2$$

$$\sigma_{s,\theta} = \frac{\log q}{s_0^2 p}\sigma_s^2 = -\frac{\theta(1 - s_0)}{s_0 n_0}\left(1 - \frac{n_0}{n}\right), \tag{A.1}$$

where $\sigma_{\tilde{p}}^2$ is the variance of $\hat{p}$. The formula $\mathrm{Var}(\hat{p}) = \mathrm{E}(\mathrm{Var}(\hat{p}|\tilde{p})) + \mathrm{Var}(\mathrm{E}(\hat{p}|\tilde{p}))$ now gives

$$\sigma_{\hat{p}}^2 = \left(1 - \frac{n_1}{s_0 n}\right)\frac{pq - \sigma_{\tilde{p}}^2}{n_1} + \sigma_{\tilde{p}}^2 \approx \left(1 - \frac{n_1}{s_0 n}\right)\frac{pq}{n_1} + \sigma_{\tilde{p}}^2, \tag{A.2}$$

where $\sigma_{\tilde{p}}^2$ is defined in (4). This formula separates the variance into a term for sample error and a term arising from the randomness in the epidemic outbreak. Combining the results gives expressions for $\sigma_{s,\theta}$ and σ_θ^2 in (6) and (7), as required. □

References

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

Bailey, N.T.J., 1975. The Mathematical Theory of Infectious Diseases and its Applications. Griffin, London.

Ball, F., 1983. The threshold behaviour of epidemic models. J. Appl. Probab. 20, 227–241.

Becker, N.G., 1989. Analysis of Infectious Disease Data. Chapman & Hall, London.

Becker, N.G., Hasofer, A.M., 1997. Estimation in epidemics with incomplete observations. J. Roy. Statist. Soc. B 59, 415–429.

Bouma, A., De Jong, M.C.M., Kimman, T.G., 1997. The influence of maternal immunity on the transmission of pseudorabies virus and on the effectiveness of vaccination. Vaccine 15, 287–294.

Britton, T., 1998. Estimation in multitype epidemics. J. Roy. Statist. Soc. B 60, 663–679.

Halloran, M.E., Haber, M., Longini, I.M., 1992. Interpretation and estimation of vaccine efficacy under heterogeneity. Am. J. Epidemiol. 136, 328–342.

Longini, I.M., Koopman, J.S., 1982. Household and community transmission parameters from final distributions of infections in households. Biometrics 38, 115–126.

Rida, W.N., 1991. Asymptotic properties of some estimators for the infection rate in the general stochastic epidemic. J. Roy. Statist. Soc. B 53, 269–283.