Frequency distribution of lymphatic filariasis microfilariae in human populations: population processes and statistical estimation

B. T. GRENFELL[2], P. K. DAS[1], P. K. RAJAGOPALAN[1] *and* D. A. P. BUNDY[3]

[1] *Vector Control Research Centre, Medical Complex, Indira Nagar, Pondicherry-605 006, India*
[2] *Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ*
[3] *Parasite Epidemiology Research Group, Department of Pure and Applied Biology, Imperial College, London SW7 2BB*

(*Accepted 7 June 1990*)

SUMMARY

This paper uses simple mathematical models and statistical estimation techniques to analyse the frequency distribution of microfilariae (mf) in blood samples from human populations which are endemic for lymphatic filariasis. The theoretical analysis examines the relationship between microfilarial burdens and the prevalence of adult (macrofilarial) worms in the human host population. The main finding is that a large proportion of observed mf-negatives may be 'true' zeros, arising from the absence of macrofilarial infections or unmated adult worms, rather than being attributable to the blood sampling process. The corresponding mf distribution should then follow a Poisson mixture, arising from the sampling of mf positives, with an additional proportion of 'true' mf-zeros. This hypothesis is supported by analysis of observed *Wuchereria bancrofti* mf distributions from Southern India, Japan and Fiji, in which zero-truncated Poisson mixtures fit mf-positive counts more effectively than distributions including the observed zeros. The fits of two Poisson mixtures, the negative binomial and the Sichel distribution, are compared. The Sichel provides a slightly better empirical description of the mf density distribution; reasons for this improvement, and a discussion of the relative merits of the two distributions, are presented. The impact on observed mf distributions of increasing blood sampling volume and extraction efficiency are illustrated via a simple model, and directions for future work are identified.

Key words: *Wuchereria bancrofti*, lymphatic filariasis, microfilariae, frequency distributions, truncated negative binomial.

INTRODUCTION

The fundamental measure of macroparasite abundance in the definitive host is the *frequency distribution* of adult parasites, either measured directly or via the production of transmission stages such as eggs or microfilariae. Apart from the fact that they underlie basic summary measures of infection, such as prevalence and intensity, these distributions are also of intrinsic epidemiological interest. In particular, the overdispersed frequency distributions which characteristically describe observed adult parasite abundance give a measure of the unevenness of parasite distribution between hosts (Anderson & May, 1985). In turn, this has important theoretical implications for the operations of non-linearities, such as density dependence and acquired immunity, in host–parasite relationships (Anderson & Gordon, 1982; Anderson & May, 1985; Pacala & Dobson, 1988).

When adult parasite abundance can only be measured indirectly, via egg or microfilarial counts, the interpretation of observed patterns is much more difficult. This is particularly so for filariasis, where levels of microfilariae (mf) must be measured in blood samples or skin snips (Hairston & Jachowski, 1968; Park, 1988; Sasa, 1976). The numbers of mf in peripheral night blood sampled from human populations where lymphatic filariasis is endemic will, in general, follow some discrete frequency distribution. This will reflect various heterogeneities, both within and between hosts (Park, 1988). A number of workers have used single night blood samples to examine the variation in microfilarial counts between individual hosts (Denham *et al.* 1971; Hairston & Jachowski, 1968; Hairston & de Meillion, 1968; Pichon *et al.* 1980; Park, 1988; Southgate, 1974). These studies have aimed to quantify both the prevalence and intensity of infection as measured by mf counts in humans, and the resulting potential transmission rate to the vector population (Hairston & Jachowski, 1968). Methodological factors, such as the volume of blood sampled and the efficiency of extraction methods, have also been shown to have a significant bearing on epidemiological predictions (Sasa, 1976; Southgate, 1974).

Observed mf frequency distributions have been analysed by fitting various statistical models, notably the Poisson (Hairston & Jachowski, 1968) and lognormal (Sasa, 1976). Recently, Park (1988) made the important point that the frequency distribution of mf in samples of peripheral blood can be described by a mixed Poisson distribution (Johnson & Kotz, 1969), reflecting Poisson sampling of an underlying mf 'density' distribution in the blood. In practice, Park adopted a gamma distribution for the mf blood

density, leading to a highly skewed negative binomial distribution as an empirical description of observed mf counts (Park, 1988). One problem with this approach is that it does not take explicit account of the population distribution and reproductive rate of adult *macrofilariae* from which the mf originate.

In this paper, we extend the quantitative description of observed mf frequency distributions to take account of the expected distribution of macrofilariae in the host population. The principal aim is to examine what the observed mf distribution can reveal about the underlying pattern of macrofilarial burdens, and in particular about the *prevalence* of adult worm infections. As in Park's (1988) analysis, this is achieved by fitting mixed Poisson distributions to observed mf counts, in order to estimate the density distribution of mf in peripheral blood. Since this mf 'concentration' is the important variable in determining the infection rate of vector populations (Park, 1988; Hairston & Jachowski, 1968), it is important to characterize the associated frequency distribution as accurately as possible. A subsidiary aim of the paper is therefore to identify the most efficient method for fitting observed mf frequency distributions (and therefore of estimating the mf density distribution). Specifically, we compare the fit of the negative binomial distribution, which is mathematically tractable (in terms of the ease of parameter estimation) and has parameters with some biological meaning, with a more general empirical descriptor of overdispersed frequency distributions, the Sichel distribution (Sichel, 1971; Stein, Zucchini & Juritz, 1987).

The paper is divided into four sections. We begin by deriving a simple statistical model for the distribution of mf in blood samples from human populations. The second section presents the results of fitting these models to observed microfilarial distributions. The third section outlines how changes in blood sampling volume and efficiency are likely to influence mf distributions, and the final section draws together the main conclusions of the paper.

STATISTICAL MODELS AND METHODS

Generation of microfilarial distributions in the blood

The abundance of microfilariae in the blood of infected hosts derives ultimately from the distribution of adult macrofilariae in the lymphatics. It is useful to explore the implications of this relationship for observed mf counts by considering the following simple statistical model. Define the frequency distribution of macrofilarial burdens in the host populations in terms of a discrete random variable X with probability function $\psi(x)$ $(x = 0, 1, \ldots, \infty)$, which represents the probability that an individual harbours $X = x$ adult worms. In general, this is likely to be an overdispersed distribution, such as the negative binomial, in which a proportion of individuals (corresponding to $\psi(0)$) are uninfected (Pacala & Dobson, 1988). Assume that this distribution remains constant over the time period under consideration, and that the resulting distribution of microfilariae is defined by the random variable R, with probability function $\sigma(r)$ $(r = 0, 1, \ldots, \infty)$ for the probability of $R = r$ *total* mf in an individual's blood. The relationship between microfilarial (R, σ) and macrofilarial (X, ψ) distributions can then be set out in general terms as follows

$$\text{Zero mf } (r = 0) \qquad \sigma(0) = \psi(x = 0) + S + \Omega(0) \quad (1)$$

$$\text{Mf positive } (r > 0) \qquad \sigma(r) = \Omega(r). \quad (2)$$

Equations (1) and (2) express the following assumptions. (a) Mf-positives. The proportion of individuals who harbour various burdens of mf $(\sigma(r), r > 0)$ will be determined by the proportion of female macrofilariae which are mated and producing eggs, and the balance between their *per capita* reproductive rate (which controls the recruitment of new mf into the blood) and the death rate of mf in the blood. This complex stochastic process is represented by the general function Ω(r), which indicates the proportion of individuals with a burden of r mf generated by the adult population. (b) Mf-negatives. The proportion of completely mf-negative individuals can be divided into three components. Firstly, those hosts who are completely without macrofilarial infection ($\psi(0)$). Secondly, the proportion of individuals with macrofilarial infections which do not include reproducing females (S). This category will be dominated by single-sex and prepatent infections, the implications of which have been explored in detail for filarial infections by Hairston & Jachowski (1968). Thirdly, the recruitment/death process for mf encapsulated in Ω can also lead to mf-negatives (Tallis & Leyton, 1966), which are expressed in equation (1) as $\Omega(0)$.

A number of authors have considered the theoretical implications of immigration/death processes (Ω) and mating probabilities (inversely measured by S) for parasite population dynamics. The intricacies of parasite sex ratios, mating probabilities and strategies have been discussed by Macdonald (1965), May (1977), and Anderson (1982). The parasite distributions arising from a balance between infection (or reproduction) and death have generally been discussed in terms of the basic infection process (Anderson & Gordon, 1982; Pacala & Dobson, 1988), or the distribution of eggs or free-living stages (Tallis & Donald, 1964; Tallis & Leyton, 1966). However, for present purposes, the general formulation of equations (1) and (2) will serve as a description of these processes.

The important variable in terms of blood sampling

for microfilariae is not their total abundance in the blood, but the 'concentration' of mf in peripheral or venous blood available for night blood sampling (Sasa, 1976; Park, 1988; Wenk, 1986). The simplest way to allow for this complication is to replace the discrete probability function describing the mf burden in equation (2) ($\sigma(r), r > 0$) with a continuous probability density $h(m)$, describing the associated 'concentration' of mf (m per unit blood volume) available for sampling. In particular, we represent the mf-positive distribution (which, from equation (1), amounts to a proportion $1-\sigma(0)$ of the total population) by $h(m)$, such that $\int_a^b h(m)\,dm$ is the proportion of mf positive who have concentrations in the range a to b. $h(m)$ is a true density function, since $\int_0^\infty h(m)dm = 1$ represents the total proportion of mf positives. The overall distribution of mf in the host population is then described by the sum:

$$\underset{\sigma(0)\text{mf negative}}{} + \underset{\text{mf positive}}{(1-\sigma(0))\int_0^\infty h(m)\,dm} = 1. \tag{3}$$

The transition from the original discrete macrofilarial distribution to this mixed (discrete/continuous) description of the microfilarial distribution is illustrated diagrammatically in Fig. 1.

The sampling process

As pointed out by Park (1988), the procedure of sampling blood for mf can reasonably be represented by a Poisson process (Johnson & Kotz, 1969). In other words, assuming a constant density, m, of mf available for peripheral blood sampling, the probability of sampling i mf would be given by the simple Poisson term: $m^i \exp(-m)/i!$. However, given that m varies between individual hosts (as noted above), this added heterogeneity will lead to a Poisson mixture for the observed mf distribution

$$p(i) = \int_0^\infty g(m)\, m^i \exp(-m)/i!dm, \tag{4}$$

where $p(i)$ is the probability of observing $i(=0, 1, \ldots, \infty)$ mf in a blood sample of unit volume. Park (1988) assumed a gamma distribution,

$$g(m) = [(k/\mu)^k/\Gamma(k)]m^{k-1}\exp(-km/\mu) \tag{5}$$

(with parameters μ and k) to describe the variation in m across the complete host population. This leads, in turn, to a negative binomial distribution

$$p(i) = \frac{\Gamma(k+i)}{i!\,\Gamma(k)}\left[\frac{\mu}{\mu+k}\right]^i\left[1+\frac{\mu}{k}\right]^{-k} \tag{6}$$

(again with parameters μ and k) for $p(i)$. This model implicitly assumes that zero mf counts (i.e. $p(0)$) are due only to the random nature of the sampling process. However, as discussed above, this is not the *only* source of zeros; a large (and unknown) proportion of hosts will also be 'true' mf negatives (corresponding to $\sigma(0)$ in equation (1)).

An extension to Park's analysis, which takes these factors into account, is as follows. Assume that single night blood samples of unit volume are taken from a total of N individuals. Following equation (3), a proportion ($\sigma(0)$) of these will be 'true' mf negatives, and the remainder ($1-\sigma(0)$) have a distribution of mf 'concentrations' available for sampling (m) described by the density function $h(m)$. Applying equation (4), the associated discrete probability function ($p(i)$), $i = 0, 1, \ldots m\,\infty$), describing the distribution of mf counts in blood samples, becomes

$$i = 0\colon p(0) = \sigma(0) + (1-\sigma(0))\,p'(0) \tag{7}$$

$$i > 0\colon p(i) = (1-\sigma(0))\,p'(i). \tag{8}$$

Here,

$$p'(i) = \int_0^\infty h(m)m^i \exp(-m)/i!\,dm \quad (i = 0, 1, \ldots, \infty) \tag{9}$$

and

$$\sum_{i=0}^{\infty} p'(i) = 1 \tag{10}$$

is the probability of observing i mf in the proportion $(1-\sigma(0))$ of the population who are mf-positive. The $p'(i)$ are therefore multiplied by the factor $(1-\sigma(0))$ in equations (7) and (8) and, in particular, $(1-\sigma(0))\,p'(0)$ is the expected proportion of mf-zeros due entirely to the sampling process. Note that, for clarity, we shall subsequently use the definitions: *mf-negative* for the proportion of 'true' mf-negatives ($\sigma(0)$) and *mf-zero* for the total observed proportion of zeros in blood samples ($p(0)$). Mf-zeros are therefore the sum of the true mf-negatives plus those samples from mf-positives where no microfilariae are counted during the sampling process. This analysis of the blood sampling process is illustrated in Fig. 1.

Estimation procedures

In practice, we cannot partition the total proportion of mf-zeros in equations (5) ($p(0)$) between mf-negatives ($\sigma(0)$) and sampling zeros. The mf density distribution, $h(m)$, can therefore only be characterized by reference to the non-zero mf distribution, $p'(i)$, $i = 0, 1, \ldots, \infty$. Since (from equation (10))

$$\sum_{i=1}^{\infty} p'(i) = 1 - p'(0), \tag{11}$$

the series

$$\frac{1}{1-p'(0)}\{p'(1)+p'(2)+\ldots\} = 1 \tag{12}$$

represents a zero-truncated probability distribution

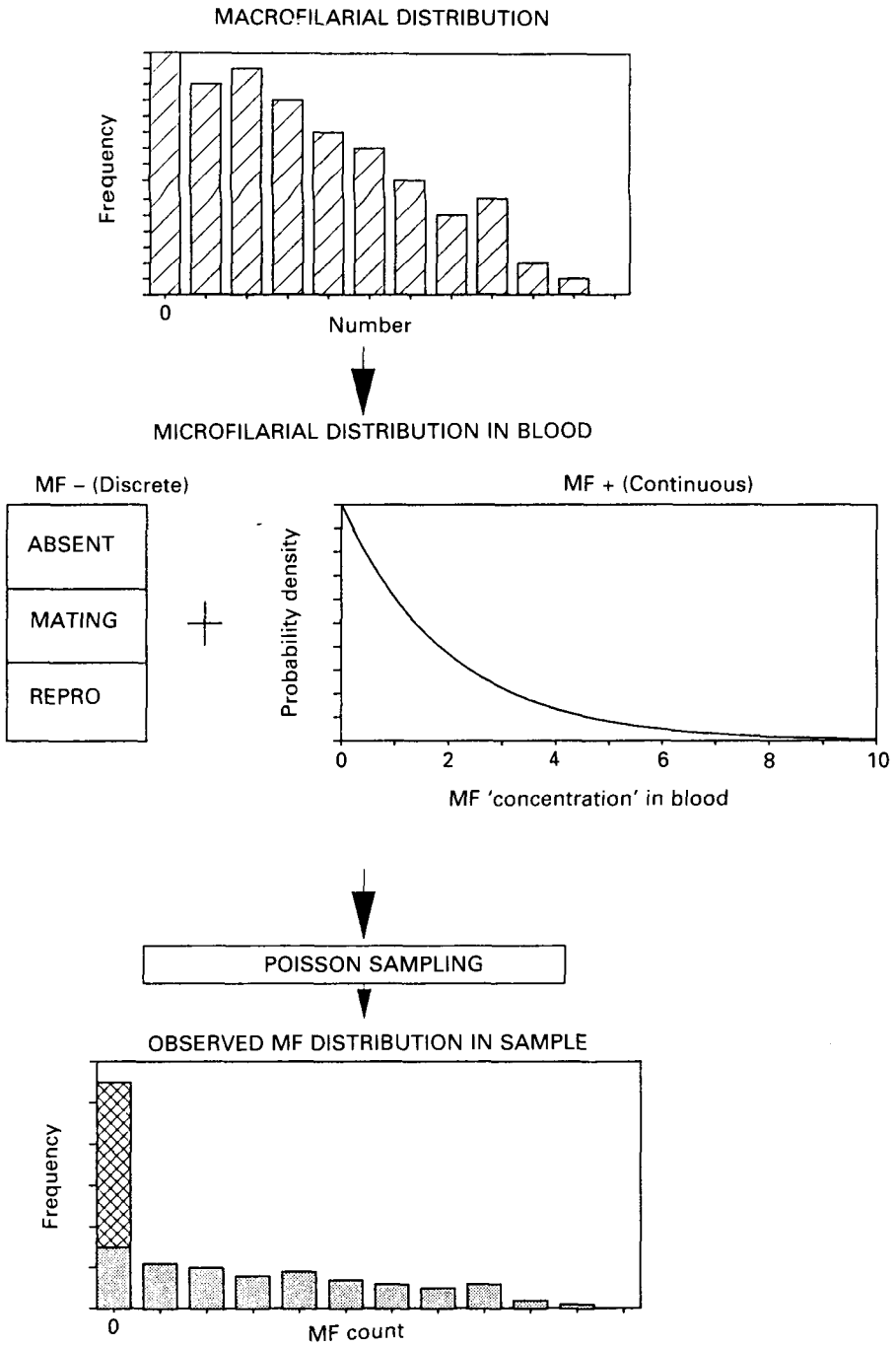

Fig. 1. Schematic illustration of the relationship between macro- and microfilarial frequency distributions in human populations, and the sampling process which leads to observed mf blood counts. The proportion of 'true' mf-negatives (MF−) is partitioned according to the macrofilarial infection status, between: individuals with no macrofilariae (ABSENT), those with pre-patent infections and/or unmated female worms (MATING), and those with mated worms but no mf arising from the stochastic balance between mf production and death processes (REPRO). The process labelled 'Poisson sampling' refers to sampling from a compound Poisson distribution (see text for more explanation). ▨, sampling of MF+; ⊠, 'true' MF−.

(Johnson & Kotz, 1969) for the positive counts. The parameters of $h(m)$ can therefore be estimated by a maximum likelihood fit of the truncated distribution (equation (12)) to non-zero mf frequency data (as described in the Appendix). Given estimates of the observed proportion of zeros ($p(0)$) and the expected proportion due to sampling ($p'(0)$), we can calculate the proportion of 'true' mf negatives from the following rearrangement of equation (7)

$$\sigma(0) = \frac{p(0) - p'(0)}{1 - p'(0)}. \tag{13}$$

Empirical descriptions of the mf distribution

Thus far, the analysis has been based on a general mf density function, $h(m)$. In practice, we compare the fits of two Poisson mixtures, the negative binomial and the Sichel distribution. Note that the zero-truncated negative binomial has also been adopted as an empirical model of mf count data by Pichon *et al.* (1980). As indicated above (equation (6)), the negative binomial is based on a gamma distribution for $h(m)$, with parameters μ and k. These parameters have the practical advantage of a simple intuitive interpretation in terms of the observed distribution; μ is the mean of the positive counts and k is an inverse measure of overdispersion. The Sichel distribution (Sichel, 1971; Stein *et al.* 1987) has three parameters (ξ, α and γ), and is based on a generalized inverse Gaussian distribution for $h(m)$. It provides a successful statistical description of an extremely wide range of overdispersed frequency data, and can mimic the behaviour of a range of other theoretical frequency distributions, including the negative binomial (Sichel, 1971). The mf density function, $h(m)$, corresponding to the Sichel fit therefore provides a very flexible empirical description of the distribution of mf concentration. This allows a rigorous assessment of the performance of the negative binomial in reflecting observed mf distributions. The disadvantages of using the Sichel are the fact that its parameters do not have any simple biological interpretation (Stein *et al.* 1987) and the relative complexity of fitting the distribution by maximum likelihood (this procedure is described in the Appendix).

The following section tests the ideas presented above by comparing the fit of full and zero-truncated Poisson mixtures to observed frequency distributions for bancroftian filariasis. The performance of the negative binomial and Sichel distributions in describing the distribution of positive counts is also compared.

DATA ANALYSIS

Data sources

Our analysis is mainly based on a large data set collected by the Vector Control Research Centre (VCRC) of the Indian Council of Medical Research before and during a vector control programme against bancroftian filariasis in Pondicherry, S. India (full details have been given by Rajagopalan & Das (1987). The frequency distribution analysed is from the main pre-control data set: a sample of 24946×20 mm^3 peripheral night blood-smears; the sex- and age-distribution of this sample have been analysed by Das *et al.* (1990).

For comparison, we also analyse two published mf frequency distributions for bancroftian filariasis. The first is from a survey of 28885 individuals from the Amami Islands of Japan, using 30 mm^3 thick blood smears (Sasa, 1976; Park, 1988). For comparison, the second survey (of 366 individuals from Fiji) was based on much larger (1 ml) blood samples and the more efficient millipore filtration method for mf extraction (Southgate, 1974).

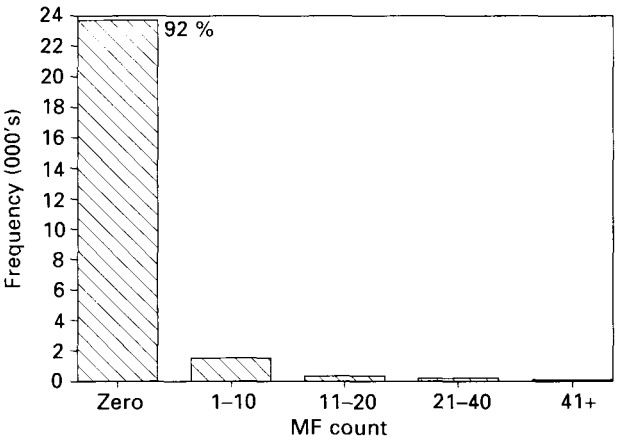

Fig. 2. Observed frequency distribution of mf in 20 mm^2 blood samples from the VCRC data set (see main text), illustrating the characteristic preponderance of zero samples.

Basic fits

All the observed distributions show the pattern illustrated for the VCRC data in Fig. 2. Typically, for mf samples (Park, 1988), this is a highly skewed distribution, with a large preponderance (92 %) of zeros. Table 1 presents the results of fitting negative binomial and Sichel distributions, as described above, to the VCRC, Amami and Fiji data. The results for each distribution are presented as parameter estimates with approximate standard errors (see Appendix), estimates of the proportion of mf negatives ($\sigma(0)$), and χ^2 goodness-of-fit tests. Comparisons of observed and expected positive mf counts for each data set are also presented in Fig. 3.

The VCRC data. The most reliable fits are for the VCRC data, where the raw data (and therefore the full tail of the mf distribution) were available. These results are displayed in Fig. 3A, which clearly shows that the full (i.e. non-truncated) negative binomial distribution is a much poorer fit to the observed distribution. This lack of fit is confirmed by the χ^2 analysis in Table 1, and arises because the full distribution is much too skewed to represent the distribution of mf-positive counts. The negative binomial parameter k, which is an inverse measure of this overdispersion, is an order of magnitude lower for the full distribution fit than for the zero-truncated estimate (Table 1). This comparison bears out the theoretical argument advanced above. The non-truncated fit is very overdispersed compared to the truncated analysis because it is dominated by the

Table 1. Details of fitting the negative binomial distribution (full and zero-truncated) and the zero-truncated Sichel distribution to the observed mf count frequency distributions described in the text

(Numbers in parentheses are approximate standard errors for the appropriate parameters. The calculation of $\sigma(0)$, the proportion of 'true' mf zeros, is described in the text.)

Distribution	Parameters/ statistics	Data set		
		VCRC	Amami	Fiji
Non-truncated Neg. binomial	μ	3·62 (1·2)	3·63 (0·8)	3·6 (0·7)
	k	0·019 (0·0004)	0·005 (0·0002)	0·054 (0·0004)
	χ^2, D.F., p	397, 83, $< 10^{-10}$	489, 13, $< 10^{-10}$	1908, 8, $< 10^{-10}$
Zero-truncated Neg. binomial	μ	6·62 (0·36)	0·62 (0·045)	89·5 (15·4)
	k	0·34 (0·03)	0·021 (0·007)	0·24 (0·062)
	$\sigma(0)$	0·86	0·54	0·12
	χ^2, D.F., p	83·4, 49, 0·0016	37·1, 12, 0·00021	12·21, 15, 0·63
Non-truncated Sichel	ξ	—	—	0·03 (0·025)
	α	—	—	0·0021 (0·002)
	γ	—	—	1·7 (0·17)
	χ^2, D.F., p	—	—	12·4, 14, 0·57
Zero-truncated Sichel	ξ	4·08 (0·88)	0·22 (0·07)	26·7 (17)
	α	1·61 (0·27)	0·08 (0·0002)	1·82 (0·84)
	γ	−0·11 (0·096)	0·014 (0·0036)	0·0015 (0·001)
	$\sigma(0)$	0·9	0·92	0·29
	χ^2, D.F., p	57·5, 49, 0·19	36·5, 1, 0·00014	7·58, 13, 0·869

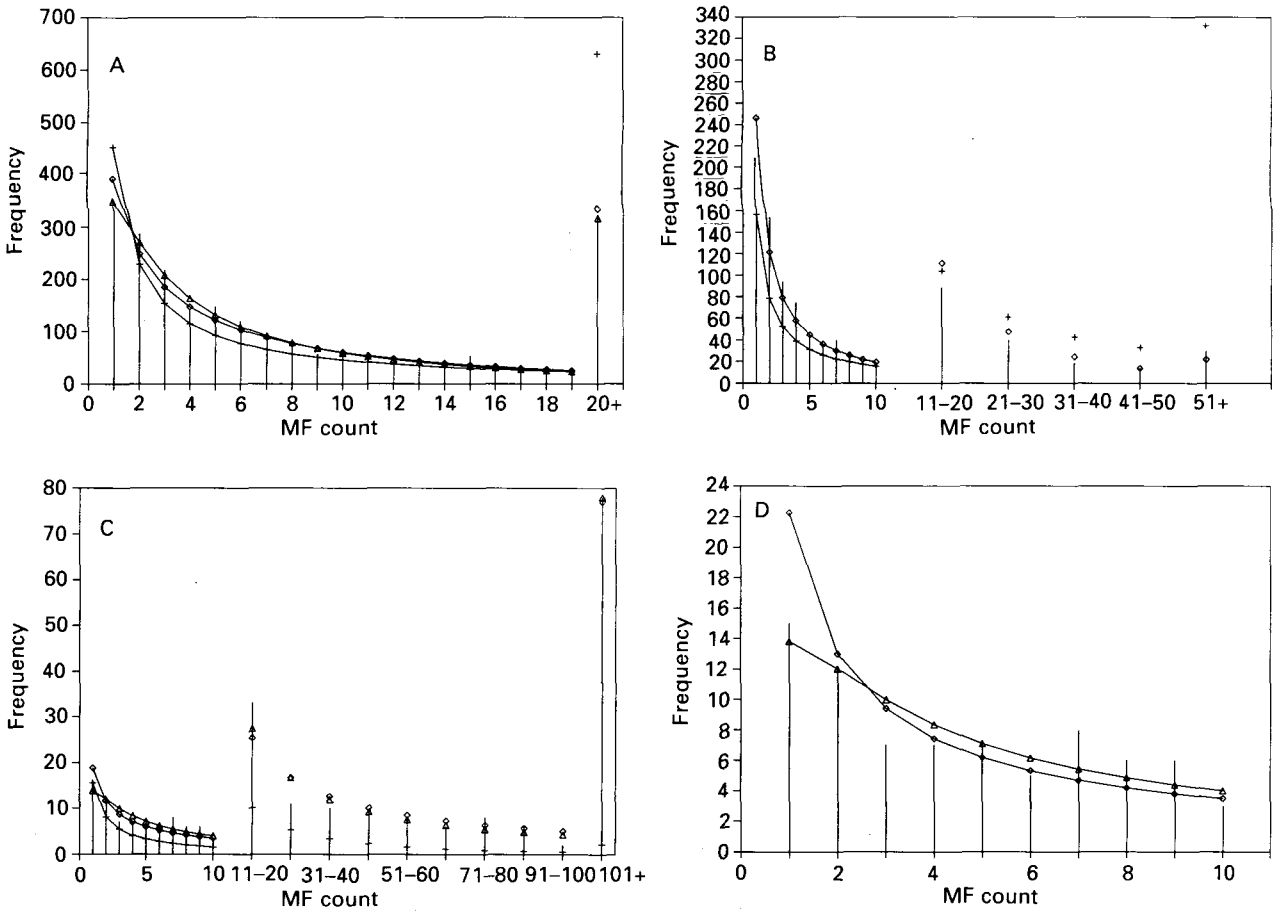

Fig. 3. Observed and expected mf-positive frequency distributions arising from the fits documented in Table 1. (A)–(C) fits of the non-truncated negative binomial (+), zero-truncated negative binomial (◇) and zero-truncated Sichel distributions (△) to (A) the VCRC data, (B) the Amami data and (C) the Fiji data (where the two truncated distributions have almost identical expected curves, only the negative binomial is shown). (D) Comparison of non-truncated (◇) and zero-truncated (△) Sichel distribution fits for the Fiji data (see text for explanation).

large preponderance of observed zeros (Park, 1988). The truncated distribution mimics the spread of positive mf counts much more closely, bearing out the hypothesis (embodied in equations (7) and (8)) that a large proportion of the observed zeros are 'true' mf negatives, rather than simply being attributable to the sampling process.

As described above, the Sichel distribution allows a closer examination of the negative binomial fit. In terms of the truncated fits, the three parameter Sichel is a rather better representation of the positive counts than the negative binomial (Table 1). In particular, although the latter fits the tail of the observed distribution well, it is too skewed to match the frequency of low counts (Fig. 3A). As discussed by Das *et al.* (1990), this effect is probably due to the presence of significant heterogeneities in the mf distribution as a function of host age (Dietz, 1982*b*). The probability levels quoted in Table 1 are best viewed as relative rather than absolute measures of goodness-of-fit, since the χ^2 statistic is conservative in this respect for moderate to large sample sizes (Park, 1988). The analysis based on both truncated negative binomial and Sichel distributions indicates that the proportion of 'true' zeros ($\sigma(0)$ in Table 1) may be very large, amounting to around 90% of the total population.

The very overdispersed distribution with a low mean generated by including the observed mf zeros (Fig. 2) produces a fit for the non-truncated Sichel distribution which approximates to a modified log-series distribution (Johnson & Kotz, 1969; Stein *et al.* 1987). This corresponds to the limit $\alpha \to 0$ (Sichel, 1971), which produces a numerically unstable fit (see Appendix). The full distribution fit is therefore omitted from the results reported in Table 1, for both the VCRC and Amami data (which have low mean counts due to the small sampling volumes used). For positive expected frequencies, the log-series distribution, which is the appropriate limit of the Sichel distribution, is almost indistinguishable from the negative binomial for low μ and k (Fig. 3A).

As described above, the observed mf distribution in positive individuals can be thought of as arising from sampling from a compound Poisson distribution with an underlying continuous distribution $h(m)$ of microfilarial 'concentration' (m) in the blood. Fig. 4 compares the shape of the $h(m)$ distribution derived from the truncated and non-truncated fits of the negative binomial distribution to the VCRC data. Here, $h(m)$ takes the form of a gamma distribution (equation (4)), with parameters μ and k (Table 1). Fig. 4 presents the probability distribution of m arising from $h(m)$. It illustrates the much less skewed mf density distribution deriving from the truncated fit (the non-truncated and truncated density distributions have coefficients of skewness of 14·5 and 3·4 respectively). The non-truncated analysis indicates that almost 90% of mf-positives have an mf concentration of less than 0·25/20 mm^3, compared with 25% in this range for the truncated fit. The upper tail (where the truncated distribution is a much better fit to the observations) comprises around 60% of positive individuals above $m =$ 1/20 mm^3 for the truncated fit, and only 10% for the non-truncated. The implications of these results are discussed below.

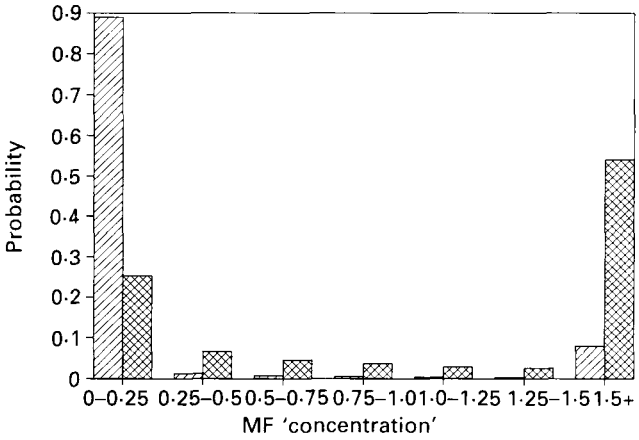

Fig. 4. Expected probability distribution ($h(m)$) of mf peripheral blood 'concentration' (m) in mf-positive individuals, calculated from the non-truncated (▨) and zero-truncated (⊠) negative binomial distribution fits to the VCRC frequency data. Probabilities were calculated by numerically integrating equation (4) over the appropriate limits for m. This discretized version of the distributions is presented for comparison with the results of Park (1988).

The Amami data. Fig. 3B and Table 1 again indicate that the mf counts from the Amami Islands are much better fitted by zero-truncated compound Poisson distributions, which allow for a proportion of mf zeros in excess of those due to Poisson sampling. The non-truncated negative binomial is much more overdispersed than the truncated form (as shown by the estimates of k in Table 1), and greatly overestimates the length of the upper tail of the observed distribution (Fig. 3B and χ^2 in Table 1). The truncated negative binomial and Sichel fits generate very similar expected non-zero counts, representing the tail of the observed distribution much more closely and slightly overestimating the degree of skewness of the low mf counts (Fig. 3B, Table 1). However, the truncated negative binomial generates a higher expected proportion of zeros due to 'sampling' ($p'(0)$) than the Sichel, leading to a higher proportion of true mf negatives ($\sigma(0) = 0{\cdot}54$ versus $\sigma(0) = 0{\cdot}92$; Table 1). This discrepancy, which illustrates the danger of extrapolating exact proportions of 'true' mf zeros from simple truncated fits, will be discussed below.

The Fiji data. These are characterized by a much longer tail to the mf frequency distribution than the

VCRC and Amami data (Fig. 3C). This reflects the larger blood volume and more efficient sampling method used (Southgate, 1974). The negative binomial fits (Fig. 3C, Table 1) again bear out the improvement in fit achieved with a truncated distribution, which mimics the upper tail of the distribution much more effectively. As with the VCRC data set, the truncated Sichel distribution produces a slightly better fit to the positive counts (Table 1). The flexibility of the Sichel distribution is further apparent from a successful fit of the non-truncated form to the full data set (Table 1). However, the truncated form is a more successful mimic, particularly of the lower end of the mf-positive distribution (Fig. 3D). The truncated negative binomial and Sichel fits indicate proportions of 'true' mf-negatives of 12 and 29% (Table 1). These rather low percentages partly reflect the large sampling volume and efficiency, although the prevalence of mf was high even in small blood samples from this population (Southgate, 1974).

Overall, all three data sets support the theoretical prediction that zero-truncated Poisson mixtures provide a better (although not a perfect) empirical description of observed mf blood counts, when compared with the equivalent distributions including zeros. However, only the results for the VCRC data (where the full tail of the distribution was available for fitting) can be treated with real confidence. Despite this, the analysis of the Fiji data underlines the potentially dramatic impact on observed mf counts of increases in blood volume and diagnostic efficiency. These ideas are explored theoretically in the next section.

THE EFFECTS OF SAMPLING VOLUME AND DIAGNOSTIC EFFICIENCY

The above analysis implicitly assumes that a unit volume of blood is sampled for mf with unit efficiency. We explore the implications for observed patterns of varying these parameters via the following simple model.

(a) *Diagnostic efficiency*

Assume that a unit volume of blood is sampled for mf from an individual with a constant mf density (m) available for sampling. Defining the proportional efficiency of detecting microfilariae in the sample by E ($0 < E < 1$), the effective density sampled is mE. If variations in mf concentration at the population level are described, as before, by the density function $h(m)$, the resulting probability distribution of observed counts from positive individuals, $p'(i)$, $i = 0, 1, \ldots, \infty$ (equation (9)) then becomes

$$p'(i) = \int_0^\infty h(m)(mE)^i \exp(-mE)/i!\, dm. \qquad (14)$$

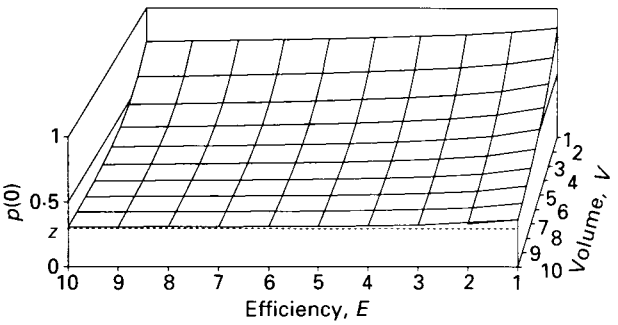

Fig. 5. The effect of varying blood sampling volume and percentage mf detection efficiency on the proportion of observed zeros ($p(0)$), assuming a proportion of 'true' mf zeros, $\sigma(0) = 0{\cdot}3$ and a negative binomial distribution, with parameters $m = 1$ and $k = 0{\cdot}1$, for the remainder of the population (see text and equation (15)).

(b) *Sampling volume*

By contrast, if we take a volume V times unit volume, the simplest sampling assumption is that this is equivalent to adding V samples of unit volume, each of which has an identical mf frequency distribution ($p(i)$, $i = 0, 1, \ldots, \infty$; equations (7) and (8)).

The interpretation of these models is particularly simple for the negative binomial distribution arising if $h(m)$ is a gamma distribution with parameters μ and k. For unit sampling volume and efficiency, the negative binomial distribution of observed positive counts will also have parameters μ and k. Increasing sampling efficiency by a factor E will increase the mean in proportion to μE, whereas enlarging the volume by a factor V will increase both parameters, to μV and kV respectively (Pacala & Dobson, 1988). Incorporating these assumptions into equations (4) and (7), the total proportion of mf zeros predicted by *this simple model becomes*

$$p(0) = \sigma(0) + (1 - \sigma(0))\, p'(0), \qquad (15)$$

where $p'(0) = [1 + (\mu EV)/(Vk)]^{-(Vk)}$ is the zero probability of the negative binomial distribution (equation (6)), allowing for sampling volume and efficiency.

Fig. 5 illustrates the effects of varying blood sampling volume (V) and efficiency (E) on the proportion of observed mf zeros as described by the simple model embodied in equation (15). The figure compares the impact of varying sampling volume and efficiency over the same range (one order of magnitude), assuming a negative binomial distribution for the counts arising from positive individuals and a proportion of 'true' mf zeros, $\sigma(0) = 0{\cdot}3$. As is observed in empirical studies (Denham *et al.* 1971; Southgate, 1974), increasing both V and E decreases the proportion of observed zeros. This is due to a decrease in the mf zeros due to sampling as V and E increase. In particular, for very large V, equation (15) indicates that the creation of 'false' zeros by the sampling zeros will fall to low levels, so

that the observed zeros, $p(0)$, no longer underestimate the 'true' zeros, $\sigma(0)$ (Southgate, 1974).

Fig. 5 also illustrates the result that increases in sampling volume have a much greater potential effect in reducing sampling zeros than increases in efficiency. This is because (as discussed above) a larger volume increases both the mean (μ) and the parameter k of the negative binomial sampling distribution, whilst a larger efficiency only increases μ. From equation (15), enlarging V has a much larger effect in reducing the sampling zeros, since (by increasing k) it decreases the degree of overdispersion of the sampling distribution. This effect will be magnified by the fact that sampling volume can be increased over a much wider range than diagnostic efficiency (which is a proportion, bounded at 100%). In practice, the situation is much more complex, in that large sampling volumes are based on venous blood, which may have an intrinsically different concentration of mf (Sasa, 1976; Southgate, 1974). Nevertheless, this simple model does provide a conceptual framework for evaluating the effects of sampling volume and efficiency.

In principle, the estimation of microfilarial prevalence and intensity could be significantly improved by utilizing the extra information provided by parallel samples of a range of blood volumes from the same population (Southgate, 1974). However, preliminary studies along these lines indicate that the resulting distributions are too heterogenous to be modelled in a multivariate analysis. This may partly be due to the variabilities introduced by lumping mf distributions across age classes (Das *et al.* 1990). An ideal data set for the purposes of this analysis would comprise parallel age-structured samples utilizing a consistent extraction method at a range of blood volumes.

DISCUSSION

A central question in the epidemiology of filarial infections is the relationship between macrofilarial burdens (which cannot in general be assessed) and their indirect estimation in terms of mf counts (Dietz, 1982*a*; Hairston & Jachowski, 1968). The main aim of this paper is to use simple models to consider the implications of macrofilarial distributions (and in particular the *prevalence* of adult worm infections) for the distribution of microfilarial counts. Our conclusion, that there is a proportion of 'true' mf zeros in excess of those due to sampling (Park, 1988), is largely supported by the improved fit of zero truncated mixed Poisson distributions to a variety of empirical data sets, as compared with the fit of the equivalent distributions including zero counts. This result has also been established by Pichon *et al.* (1980). However, it is important to note that not all the observed distributions are closely fitted by these simple statistical models, and that even a successful fit indicates only consistency with the model rather than a *proof* of its assumptions. Nevertheless, these results suggest the important practical implication that individuals who are apparently uninfected on the basis of blood diagnosis may either be true mf-zeros, or only apparently negative as a result of the sampling process (Southgate, 1974). The proportion of false negatives is also critically dependent upon the blood volume examined. Our simple model for the effects of sampling volume and efficiency on the observed mf distribution confirms the empirical finding (Southgate, 1974) that both will increase the reliability of estimating the proportion of 'true' mf zeros (and therefore the prevalence of infection, Das *et al.* 1990).

The implications of these results for observed age-prevalence curves based on mf counts are discussed further by Das *et al.* (1990). They also have a significant bearing on our assessment of the overall transmission dynamics of the infection. In particular, Fig. 4 shows that constructing an expected blood density distribution based on a non-truncated Poisson mixture may severely underestimate the proportion of individuals in the upper tail of the distribution, i.e. those with high densities of mf available for infecting mosquitos. The implications of these results will be explored, in terms of an overall model for transmission dynamics, in a subsequent paper.

A subsidiary aim of this study was to assess the suitability of the negative binomial distribution as an empirical descriptor of observed mf counts, by comparison with the Sichel distribution. Overall, the Sichel distribution provides a slightly better description of the observed data than the negative binomial. This underlines its great flexibility as an empirical statistical model for overdispersed data (Sichel, 1971; Stein *et al.* 1987). As described in the accompanying paper (Das *et al.* 1990), the significantly better fit of the Sichel distribution to the VCRC data (Table 1), may be due to heterogeneities in the mf distribution as a function of host age. In practice, the negative binomial may be a better simple model for mf data (particularly when this is subdivided by age), due to its greater ease of use, and the fact that its parameter can be interpreted biologically.

One important caveat which emerges from the analysis of the Amami data (Table 1) is that truncated distributions which generate similar expected distributions for mf positives may indicate significantly different proportions of 'true' mf zeros ($\sigma(0)$). This is because the estimation of $\sigma(0)$ is essentially an extrapolation from an empirical fit, and should therefore be treated as only a very approximate figure. A potential strategy for improving these estimates is to use the additional information provided by sampling a range of blood

volumes. More generally, the relationship between macro- and microfilarial distributions can only be clarified by a combination of theoretical, field and laboratory studies. These should be particularly fruitful given the current interest in developing novel methods for immunodiagnosis of different stages of filarial parasites (Forsythe *et al.* 1990).

The Vector Control Research Centre is an institute of the Indian Council of Medical Research. B.G. and D.A.P.B. were supported by the Rockefeller Foundation and the Wellcome Trust. We thank an anonymous referee for very helpful comments.

APPENDIX

This Appendix describes the maximum likelihood methods used to produce the fits of the full and zero-truncated negative binomial and Sichel distributions to mf count data, as described in the main text.

(*a*) *Full (non-truncated) distributions*

The non-truncated negative binomial was fitted by standard maximum likelihood methods (Johnson & Kotz, 1969). The Sichel distribution is much less easily dealt with, but can be efficiently fitted in the following parameterization, introduced by Stein *et al.* (1987).

$$p'(i) = \frac{(\omega/\alpha)^{\gamma}}{K_{\gamma}(\omega)} \frac{(\xi\omega/\alpha)^{i}}{i!} K_{i+\gamma}(\alpha). \qquad \text{(A 1)}$$

Here $p'(i)$ $(i = 0, 1, \ldots, \infty)$ is the probability function, representing the probability of observing i individuals. It is defined in terms of three parameters: $\xi(>0)$, $\alpha(>0)$ and $\gamma(-\infty < \gamma < \infty)$, (with the definition $\omega = (\xi^2+\alpha^2)^{0.5} - \xi$). The main numerical complexity in equation (A 1) involves the calculation of K, which is a modified Bessel function of the second kind. This function is defined in general as $K_{\nu}(z)$, with order ν and argument z. Since the order, as defined in equation (A 1), will generally not be an integer, we calculate K by numerical solution of the integral

$$K_{\nu}(z) = \int_0^{\infty} \exp(-\cosh(t)\, z) \cosh(\nu t)\, dt \quad \text{for } |\nu| > 1. \qquad \text{(A 2)}$$

and then use the recurrence relationship

$$K_{\nu+1}(z) = \frac{2\nu}{z} K_{\nu}(z) + K_{\nu-1}(z) \quad \text{for } |\nu| > 1. \qquad \text{(A 3)}$$

Full details of the calculation of Bessel functions have been given by Abramowitz & Stegun (1972).

Given numerical estimates of $p'(i)$, we can calculate the log-likelihood function

$$L = \sum_{i=0}^{\infty} \log\{p'(i)\}\, n_i \qquad \text{(A 4)}$$

(wherein n_i is the frequency of observing i mf counts in a sample), as a function of ξ, α and γ. Maximum likelihood estimates of these parameters can then be obtained by numerically maximizing equation (A 4). We adopt the simplex method for this purpose, calculating approximate standard errors from a quadratic approximation (Nelder & Mead, 1965). The method works well, except near the limit $\alpha \to 0$ when, as described in the main text, the highly skewed observed distribution approximates to a modified log series distribution. An alternative method based on Newton–Raphson iteration has been given by Stein *et al.* (1987).

(*b*) *Zero-truncated distributions*

From equation (12), the general log-likelihood function for the zero-truncated distribution is

$$L = \sum_{i=1}^{\infty} \log\{p'(i)/[1-p'(0)]\}\, n_i. \qquad \text{(A 5)}$$

Maximum likelihood parameter estimates can then be obtained by substituting in equation (A 5) the appropriate expression for $p'(i)$ (equation (6) for the negative binomial distribution and equation (A 1) for the Sichel), and then maximizing this function as described above (Schenzle (1979), Johnson & Kotz (1969) and Pichon *et al.* (1980) described and reviewed specific mathematical likelihood estimates for the negative binomial). When published data are only available in blocked form (e.g. reporting only summed frequencies over a range of mf counts as for the Amami and Fiji data)), the expected probabilities within blocks are summed before being entered into equation (A 5). The resulting fits are approximate and should be treated cautiously.

REFERENCES

ABRAMOWITZ, M. & STEGUN, I. A. (1972). *Handbook of Mathematical Functions*. New York: Dover.

ANDERSON, R. M. (1982). The population dynamics and control of roundworm and hookworm infections. In *Population Dynamics of Infectious Diseases: Theory and Applications* (ed. Anderson, R. M.), pp. 67–106. London: Chapman & Hall.

ANDERSON, R. M. & GORDON, D. M. (1982). The regulation of host population growth by parasite species. *Parasitology* **76**, 119–57.

ANDERSON, R. M. & MAY, R. M. (1985). Helminth infections of humans: mathematical models, population dynamics and control. *Advances in Parasitology* **24**, 1–101.

DAS, P. K., MANOHARAN, A., SRIVIDYA, A., GRENFELL, B. T., BUNDY, D. A. P. & VANAMAIL, P. (1990). Frequency distribution of *Wuchereria bancrofti* microfilariae in human populations and its relationships with age and sex. *Parasitology* **101**, 429–34.

DIETZ, K. (1982*a*). The population dynamics of onchocerciasis. In *Population Dynamics of Infectious Diseases: Theory and Applications* (ed. Anderson, R. M.), pp. 209–241. London: Chapman & Hall.

DIETZ, K. (1982*b*). Overall population patterns in the transmission cycle of infectious disease agents. In *Population Biology of Infectious Diseases* (ed. Anderson, R. M. & May, R. M.), pp. 87–102. Berlin: Springer Verlag.

DENHAM, D. A., DENNIS, D. T., PONNUDURAI, T., NELSON, G. S. & GUY, F. (1971). Comparison of a counting chamber and thick smear methods of counting microfilariae. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **65**, 521–6.

FORSYTHE, K. P., GRENFELL, B. T., SPARK, R., KAZURA, J. W. & ALPERS, M. P. (1989). Age-specific patterns of change in the dynamics of *Wuchereria bancrofti* infection in Papua New Guinea. *American Journal of Tropical Medicine and Hygiene* (In the Press).

HAIRSTON, N. A. & DE MEILLION, B. (1968). On the inefficiency of transmission of *Wuchereria bancrofti* from mosquito to human host. *WHO Bulletin* **38**, 308–12.

HAIRSTON, N. A. & JACHOWSKI, L. A. (1968). Analysis of the *W. bancrofti* population in people of Western Samoa. *WHO Bulletin* **38**, 29–59.

JOHNSON, N. L. & KOTZ, S. (1969). *Discrete distributions*. New York: John Wiley & Sons.

MACDONALD, G. (1965). The dynamics of helminth infections, with special reference to schistosomes. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **59**, 489–506.

MAY, R. M. (1977). Togetherness among schistosomes: its effects on the dynamics of the infection. *Mathematical Biosciences* **35**, 301–43.

NELDER, J. A. & MEAD, R. (1965). A simplex method of function minimization. *Computer Journal* **7**, 308–12.

PACALA, S. W. & DOBSON, A. P. (1988). The relation between the number of parasites/host and host age: population dynamic causes and maximum likelihood estimation. *Parasitology* **96**, 197–210.

PARK, C. B. (1988). Microfilarial density distribution in the human population and its infertility index for the mosquito population. *Parasitology* **96**, 265–71.

PICHON, G., MERLIN, M., FAGNEAUX, G., RIVIERE, F. & LAIGRET, J. (1980). Etude de la distribution des numerations microfilariennes dans les foyers de filariose lymphatique. *Tropenmedizin und Parasitologie* **31**, 165–80.

RAJAGOPALAN, P. K. & DAS, P. K. (1987). *The Pondicherry Project on Integrated Disease Vector Control*. Vector Control Research Centre, Pondicherry.

SASA, M. (1976). *Human Filariasis : a Global Survey of Epidemiology and Control*. Tokyo: University of Tokyo Press.

SCHENZLE, D. (1979). Fitting the truncated negative binomial distribution without the second sample moment. *Biometrics* **35**, 637–9.

SICHEL, H. S. (1971). On a family of discrete distributions particularly suited to represent long-tailed frequency data. In *Proceedings of the Third Symposium on Mathematical Statistics* (ed. Laubscher, N. F.), pp. 51–97. Pretoria, Council for Scientific and Industrial Research.

SOUTHGATE, B. A. (1974). Problems of clinical and biological measurements in the epidemiology and control of filarial infections. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **68**, 177–86.

STEIN, G. Z., ZUCCHINI, W. & JURITZ, J. M. (1987). Parameter estimation for the Sichel distribution and its multivariate extension. *Journal of the American Statistical Association* **82**, 938–44.

TALLIS, G. M. & DONALD, A. D. (1964). Models for the distribution on pasture of infective larvae of the gastrointestinal nematode parasites of sheep. *Australian Journal of Biological Sciences* **17**, 504–13.

TALLIS, G. M. & LEYTON, M. (1966). A stochastic approach to the study of parasite populations. *Journal of Theoretical Biology* **13**, 251–60.

WENK, P. (1986). The function of non-circulating microfilariae: *Litmosoides carinii* (Nematoda: Filarioidea). *Deutsche tierärztliche Wochenschrift* **93**, 414–18.