Journal of Helminthology (2004) **78**, 57–61 DOI: 10.1079/JOH2003214

Parasite aggregations in host populations using a reformulated negative binomial model

P. Pal* and J.W. Lewis

School of Biological Sciences, Royal Holloway, University of London,
Egham, Surrey, TW20 0EX, UK

Abstract

The negative binomial distribution model is reformulated and used to demarcate a host population at a specific level of infection by defining an attribute spanning a range of parasite aggregations. The upper limit of the range specifies the boundary for the classification of the host population and provides a technique to determine the cumulative probability at any level of parasite infection to a high degree of accuracy. This approach also leads to the evaluation of the *k* parameter, i.e. an inverse measure of dispersion of parasite aggregation, for each fraction of the host population with a discrete level of infection. The basic mathematical premise of the negative binomial function is unaltered in developing this reformulation which was applied to data on the distribution of the trichostrongylid nematode *Heligmosomoides polygyrus* in populations of the field mouse, *Apodemus sylvaticus*.

Introduction

Empirical distributions have been extensively represented by the negative binomial probability function for the past 60 years or so. However, in a critical report, Smith *et al.* (1991) pointed out that the aggregated nature of parasite frequency distributions have been dealt with in an approximate or superficial way. The present study attempts to address this issue through a reformulation of the negative binomial probability function.

Crofton (1971) raised the question of a probabilistic distribution model which was subsequently addressed by May (1977) in the development of deterministic models involving the dynamics of host–parasite interactions. In May's model the dynamics are formulated as a set of differential equations involving birth and death processes of host–parasite systems. In the process of solving these equations, approximations to the time independent negative binomial distribution (NBD) were introduced (May, 1977; Anderson & May, 1978).

Phenomenologically parasite distribution is characterized by the aggregation of parasites of variable numbers among hosts. Analyses on observed data from field/laboratory studies are performed using four single valued summary indices including the mean intensity of infection *m*, variance *v*, prevalence *p*. The fourth index *k* is used as an inverse measure of parasite aggregation. Despite the inherent variability of parasite aggregation, the choice of this single valued measure is based on the fact that *k* varies slowly with the mean intensity for a sample of size *N* of the host population and is sensitive to the actual frequency distribution (Pielou, 1977).

The primary aim of the present paper is to reformulate the NBD function without altering its basic mathematical tenet. The reformulated model provides techniques to evaluate the *k* index of parasite dispersion in host populations at discrete levels of aggregation.

Materials and methods

Reformulation of the negative binomial

In general terms, the binomial distribution determines the probabilities of fractions of a random population that possesses a specific attribute. In parasite frequency distribution studies, the application of the NBD is focused on evaluating the probability of an uninfected

*Fax: +44 1784 434348
E-mail: p.pal@rhul.ac.uk

https://doi.org/10.1079/JOH2003214 Published online by Cambridge University Press

fraction of a host population of size N. In the process, the population is demarcated into two classes by the attribute of infection, i.e. hosts with zero and non-zero infections.

The probability of the zero infected class is readily calculated through the well known zero order NBD equation whereas subsequent probabilities with respect to non-zero parasites at successive levels of infection are calculated in sequence. These step by step calculations are subject to a propagation of errors at each step.

In reformulating the NBD probability, an attribute α is set to span a defined range $0 \le \alpha \le i^*$, where the number of parasites is limited between the lower and upper limits of 0 and i^* respectively. The upper limit of the range sets the reference line for the level of infectivity and divides the host population into two sub-populations: (i) hosts infected with any number of parasites between 0 and i^*, and (ii) hosts infected with parasites greater than i^*.

The NBD function is expressed in terms of the attribute α in a form similar to the conventional representation as follows:

$$\left[\frac{N_{0\le\alpha\le i^*}}{N}\right] = [1 - p_\alpha] = \left[1 + \frac{m}{k_\alpha}\right]^{-k_\alpha} \tag{1}$$

In equation 1, k_α represents the inverse of parasite dispersion of greater than i^* and p_α is the prevalence of infection in the host population.

The left hand side of equation 1 is a cumulative fraction of α attributed hosts (the first sub-population) infected with parasite numbers in the range between 0 and i^* as follows:

$$\left[\frac{N_{0\le\alpha\le i^*}}{N}\right] = \frac{N_0 + N_1 + \cdots + N_{i^*}}{N} \tag{2}$$

Equation 1 is a general expression which reduces to the zero order negative binomial expression at $\alpha = 0$, as follows:

$$\left[\frac{N_0}{N}\right] = [1 - p] = \left[1 + \frac{m}{k}\right]^{-k} \tag{3}$$

In this case the p and k parameters are left unsubscripted. At $\alpha = 0$, the population of N hosts is divided in two classes: (i) an uninfected class with 0 parasites, comprising N_0 hosts and (ii) an infected class with non-zero parasites, comprising $N - N_0$ hosts and this defines the prevalence of infection.

The left hand side of equation 3 represents the fraction N_0/N of the host population. The right hand side is indexed by k operating on the mean intensity of infection and provides a measure of the distribution of parasite (non-zero) aggregation in the host population. Small values of k imply a high level of aggregation among few hosts (overdispersion) whereas large k values indicate a low level of aggregation (underdispersion). As k increases, the dispersion tends to form a random pattern (Poisson).

The reformulated NBD equation 1 within a specified range of the attribute can be applied to demarcate a host population at any level of infection. In the range $0 \le \alpha \le 1$, the host population is divided into two classes comprising (i) hosts infected with 0 or 1 parasite and (ii) the remainder with 2 or more parasites. Therefore the cumulative NBD probabilities of the orders 0 and 1 are lumped into one sub-population and the remaining probabilities into the other sub-population.

Furthermore, using the reformulated NBD equation, it is possible to systematically determine the aggregation index which in turn provides a measure of dispersion of parasites above the i^* threshold of the α range within the population. The process can be sequentially extended by increasing the i^* value and the host population fractionated at successive levels of infection. Thus the parameter k_α corresponding to each level of infection above the α range is determined. This leads to a complete decomposition of the k parameter. Thus the distribution is quantified by a set of k_α components corresponding to discrete levels of parasite aggregation in a host population.

Evaluation of k

By taking the natural log of both sides, equation 1 is transformed into the following quasi-linear form:

$$\ln\left[\frac{N_{0\le\alpha\le i}}{N}\right] = -k_\alpha \ln\left[1 + \frac{m}{k_\alpha}\right] \tag{4}$$

The left hand side of equation 4 is determined by inserting the ratio $[N_{0\le\alpha\le i^*}/N]$ corresponding to the α range. With the observed mean m on the right hand side, the value of k_α is systematically varied until the two sides are equal.

Using standard numerical techniques, k_α can be calculated for any desired fraction. The resulting k_α value is a measure of parasite aggregation above the upper limit of the α range.

The conventional approach calculates an estimate of k using the mean m and variance v of the observed data set with the following relationship:

$$k = \frac{m^2}{v - m} \tag{4a}$$

A comparison of k values calculated from equations 4a and 3 respectively shows a marked divergence. A re-calculation of the prevalence value using the k estimate from equation 4a gives a discrepancy in the prevalence value relative to the observed infection in a host population. A further correction can be made (Gregory & Woolhouse, 1993) by subtracting the standard error of the mean from the numerator of equation 4a as follows:

$$k = \frac{m^2 - v/N}{v - m} \tag{4b}$$

Traditionally, the k value calculated from equation 4a is modified by employing a maximum likelihood procedure (Bliss & Fisher, 1953) and a significance analysis is performed through the chi-square test between each observed frequency and the associated NBD probability. Even after refining the k index either by equation 4b or by the maximum likelihood procedure, the discrepancy in the prevalence is slightly improved. Using conventional methods, the computation of the cumulative NBD probability at successive non-zero levels of infection is cumbersome. In addition, any numerical error introduced at the zero level of infection is propagated in these subsequent calculations.

In the reformulated NBD approach, any non-zero level of infection is reached by specifying the upper boundary of the α range and the corresponding k_α parameter can be calculated using a numerical search with minimum error. By adding increments to the demarcation threshold, i.e. the upper limit of the α range, this technique can be repeatedly applied to determine k_α components at any level of infection in the host population.

Application of the reformulated NBD

The reformulated negative binomial distribution approach has been applied systematically to a range of macroparasites in populations of the field mouse *Apodemus sylvaticus*, collected from a woodland site in Surrey, southern England (grid reference, 993693) during the months of September 1999 and 2001 (Behnke *et al.*, 1999). Nematodes, and in particular the trichostrongylid *Heligmosomoides polygyrus*, are dominant members of the intestinal helminth community of *A. sylvaticus* (Lewis, 1987).

The present reformulation was used to: (i) evaluate k in the overall distribution of *H. polygyrus* in the mouse population; and (ii) decompose k as a function of the mouse population fractions infected with *H. polygyrus*.

Results and Discussion

Comparison of k *and* p *values*

A direct comparison between conventional and reformulated methods was made by calculating k and p values using equations 4a,b and 3, respectively.

The number of *H. polygyrus* found in mouse populations examined in 1999 and 2001 are shown in tables 1 and 2, respectively.

With a 55% prevalence value of *H. polygyrus* in 1999 (table 1), the k value evaluated after reformulation is lower than that calculated conventionally. With a 100% prevalence in 2001 (table 2), the reformulated value is higher than that obtained by conventional methods, whereas in 1999 and 2001 conventional methods respectively produced discrepancies of 7% and 3% in prevalence values, relative to observed data.

Decomposition of the k *parameter*

From the 1999 and 2001 data sets, the k parameter was decomposed by systematically fractionating the mouse population at every level of infection (table 3).

The decomposed k values are plotted as a function of prevalence (see fig. 1) and in both 1999 and 2001, the data exhibit a non-linear relationship. The non-linearity becomes pronounced at higher prevalence levels of

Table 1. Frequency distribution of *Heligmosomoides polygyrus (Hp)* in a sample of *Apodemus sylvaticus* from a woodland site in Surrey, September 1999.

Mice	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Hp	0	0	0	1	14	0	1	31	0	2	0	3	5	18	9	0	0	0	14	61

No. of mice = 20 *Hp* = 159

Mean, $m = 7.95$ Variance, $v = 223.94$

Equation 4a		Equation 4b		Equation 3		Observed prevalence
k	p	k	p	k	p	
0.29	62%	0.24	57%	0.22	55%	55%

The comparison shows a discrepancy of 7% in the calculated prevalence using the estimated k value from equation (4a) compared with the observed prevalence of 55%. Similarly a discrepancy of 2% is found using the k value from equation (4b).

Table 2. Frequency distribution of *Heligmosomoides polygyrus (Hp)* in a sample of *Apodemus sylvaticus* from a woodland site in Surrey, September 2001.

Mice	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Hp	34	34	8	19	17	15	4	2	6	11	9	23	29	54	3	8	3	32	5	4

No. of mice = 20 *Hp* = 320

Mean, $m = 16.0$ Variance, $v = 201.16$

Equation 4a		Equation 4b		Equation 3		Observed prevalence
k	p	k	p	k	p	
1.38	97%	1.33	96.7%	4.61	100%	100%

The comparison shows a discrepancy in the calculated prevalence of 3% using the estimated k value from equation (4a) and a discrepancy of 3.3% using the k value from equation (4b), compared with an observed prevalence of 100%.

Table 3. Decomposed *k* components corresponding to prevalence levels of infection of *Apodemus sylvaticus* with *Heligmosomoides polygyrus* from a woodland site in Surrey, September 1999 and 2001.

1999			2001		
Prevalence fraction (%)	Parasite burden	k	Prevalence fraction (%)	Parasite burden	k
55	$k_{\alpha>0}$	0.221	99.9	$k_{\alpha>0}$	4.61
45	$k_{\alpha>1}$	0.119	95	$k_{\alpha>2}$	1.088
40	$k_{\alpha>2}$	0.097	85	$k_{\alpha>3}$	0.560
35	$k_{\alpha>3}$	0.076	75	$k_{\alpha>4}$	0.364
30	$k_{\alpha>5}$	0.055	70	$k_{\alpha>5}$	0.301
25	$k_{\alpha>9}$	0.045	65	$k_{\alpha>6}$	0.252
15	$k_{\alpha>14}$	0.022	50	$k_{\alpha>9}$	0.146
10	$k_{\alpha>18}$	0.013	30	$k_{\alpha>19}$	0.065
5	$k_{\alpha>31}$	0.0055	20	$k_{\alpha>32}$	0.037
			5	$k_{\alpha>34}$	0.0066

90–100%. The continuity of the trajectories in both years is interrupted where parasite data are missing.

The decomposition of the *k* parameter with respect to discrete fractions of a host population provides a new approach to distribution patterns of parasite aggregation. At a high level of prevalence, a wide range of parasite aggregation is expected among a large proportion of hosts and therefore the decomposed k_α component is relatively large (fig. 1). Conversely, as the host population narrows to a smaller fraction, a relatively reduced range of aggregated parasites prevail in the fewer hosts and the corresponding k_α components are small (fig. 1).

Therefore, although parasite aggregation is a consequence of host parasite interactions and observations of parasite aggregation at the host population level are essentially random events, an implicitly non-random pattern exists in the distribution of these random events.

Inverse or reciprocal values of the *k* components plotted against the host population fraction illustrate the dispersion of parasite aggregation in mice in both 1999 and 2001 (fig. 2). Continuity in the dispersion trajectories of both years is broken where observations are missing.

This plot is almost a mirror image of the k_α profile. Here the inverse of k_α components are plotted against prevalence of infection to show the dispersion profile, which represents an aggregated distribution of *H. polygyrus* in the mouse population under investigation.

The aggregation of parasites can be interpreted in terms of the attribute range of the reformulated model. The upper limit of the specified attribute range divides the host population into two classes. When this limit is low, the two classes include: (i) a small fraction of hosts with a relatively narrow range of parasite aggregation; and (ii) a large proportion with a wide variation of aggregation. This is characterized by the relative homogeneity within the host population. (underdispersion) so that the corresponding dispersion measure has a low value (fig. 2).

As the upper limit is increased, the proportion of hosts infected with a high degree of aggregation decreases, i.e. only a few hosts are infected with large numbers of parasites. Taking the entire host population into consideration, hosts with high aggregation levels are seen to be thinly distributed (heterogeneous) resulting in overdispersed values (fig. 2). Also aggregations of 61 and 54 parasites found in 5% of the mouse populations in 1999 and 2001 respectively reflect high dispersion values.

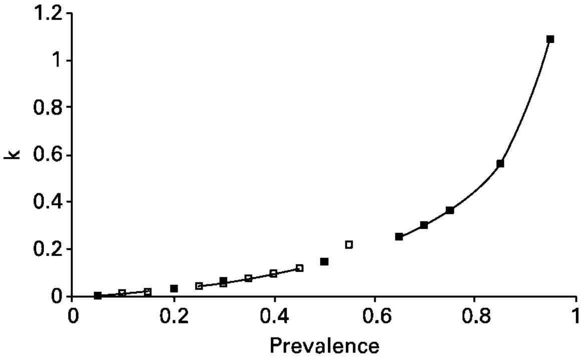

Fig. 1. The relationship between decomposed components of the k_α parameter and cumulative prevalences of infection of *Heligmosomoides polygyrus* in *Apodemus sylvaticus* from a woodland site in Surrey, September 1999 (□) and 2001 (■).

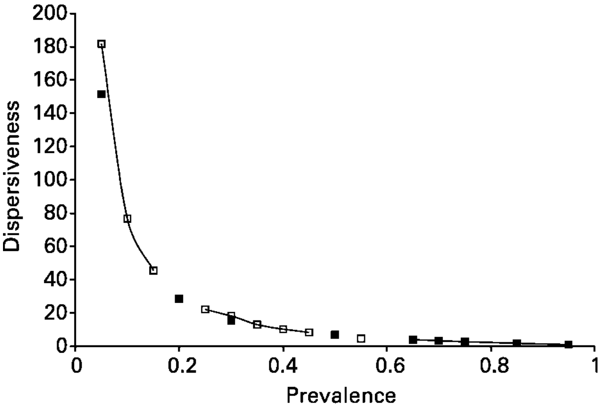

Fig. 2. The dispersion / aggregation of *Heligmosomoides polygyrus* as a function of prevalence of infection in *Apodemus sylvaticus* from a woodland site in Surrey, September 1999 (□) and 2001 (■).

https://doi.org/10.1079/JOH2003214 Published online by Cambridge University Press

In general, dispersion values tend to decrease sharply in 0.1–20% of the host population. In the 20–60% region, this decrease becomes gradual and the distribution pattern changes from overdispersion to underdispersion. Between the 60–100% prevalence levels, the cumulative parasite aggregation is underdispersed where a large proportion of hosts carry a wider range of parasite numbers. Inherent in this pattern lies the scope of assigning appropriate thresholds to demarcate a host population into risk and non-risk sub-sets in terms of infection.

In conclusion, the reformulated negative binomial approach provides an exposition of the one-to-one correspondence between fractions of the host population and parasite aggregation levels associated with them. Consequently the two populations (hosts and parasites) are broken down into multi-valued components. By using the attribute definition, the basic tenet of the negative binomial model is unaltered. Numerical calculations of distribution parameters consistent with fractions of the host population are highly accurate through this mathematically tractable procedure.

Practical implications of decomposing the k parameter lie in the fact that a set of equations are formed, which include population fractions together with their respective k indices for different mean intensities of infections. This leads to the development of theoretical predictor models. A comprehensive model is currently being developed to connect the micro-elements of the summary indices used in the study of parasite frequency distribution at the host population level. These micro-elements are the k_α components, which correspond to fractions of the host population at the level of mean intensity of infection and this can, in addition, apply to both macro and microparasites.

References

Anderson, R.M. & May, R.M. (1978) Regulation and stability of host–parasite population interactions. I. Regulatory processes. *Journal of Animal Ecology* **47**, 219–247.

Behnke, J.M., Lewis, J.W., Mohd Zain, S.N. & Gilbert, F.S. (1999) Helminth infections in *Apodemus sylvaticus* in southern England: interactive effects of host age, sex, and years on the prevalence and abundance of infections. *Journal of Helminthology* **73**, 31–44.

Bliss, C.A. & Fisher, R.A. (1953) Fitting the negative binomial to biological data and a note on the efficient fitting of the negative binomial. *Biometrics* **9**, 176–200.

Crofton, H.D. (1971) A model of host–parasite relationships. *Parasitology* **63**, 343–364.

Gregory, R.D. & Woolhouse, M.E.J. (1993) Quantification of parasite aggregation: a simulation study. *Acta Tropica* **54**, 131–139.

Lewis, J.W. (1987) Helminth parasites of British rodents and insectivores. *Mammal Review* **17**, 81–93.

May, R.M. (1977) Dynamical aspects of host–parasite associations: Crofton's model re-visited. *Parasitology* **75**, 259–276.

Pielou, E.C. (1977) *Mathematical ecology.* New York, J. Wiley.

Smith, G., Basanez, M.-G., Dietz, K., Gemmell, M.A., Grenfell, B.T., Gulland, F.M.D., Hudson, P.J., Kennedy, C.R., Lloyd, S., Medley, G., Nassel, I., Randolph, S.E., Roberts, M.G., Shaw, D.J. & Woolhouse, M.E. (1991) Macroscopic group report: problems in modelling the dynamics of macroscopic systems. pp. 209–229 *in* Grenfell, B.T. & Dobson, A.P. (*Eds*) *Ecology of infectious diseases in natural populations.* Cambridge, Cambridge University Press.

(Accepted 1 July 2003)

https://doi.org/10.1079/JOH2003214 Published online by Cambridge University Press