Letter to the Editor

# Cluster Sampling for Disease Incidence Data

G. Hughes, L. V. Madden, and G. P. Munkvold

First author: Institute of Ecology and Resource Management, University of Edinburgh, West Mains Road, Edinburgh EH9 3JG, Scotland, UK; second author: Department of Plant Pathology, Ohio State University, Ohio Agricultural Research and Development Center, Wooster 44691; and third author: Department of Plant Pathology, Iowa State University, Ames 50011.

Accepted for publication 6 October 1995.

The existence of epidemiological data presupposes some form of sampling. Guidelines for sampling procedures taken from the entomological literature (e.g., 13,24) have often been put to use by plant pathologists. Such sources make it clear that different procedures are appropriate in different circumstances. In this letter, we discuss the problem of sampling for disease incidence data (collected on the basis of scoring plants, or plant units, as either 'healthy' or 'diseased'), when the objective is to estimate mean disease incidence (the proportion of plants, or plant units, diseased) with a prespecified degree of reliability. We show that some well-known methods based on unrestricted random sampling may not be appropriate without modification. An important aspect of this problem is the spatial pattern of diseased plants. Methods of characterizing pattern that can be incorporated into formulae for sample size determination are discussed.

**Unrestricted random sampling.** Consider, first, data that comprise 'counts.' Campbell and Madden (1) give a (hypothetical) example in their Table 11.2, where the 36 observations are referred to as 'counts per quadrat' (the quadrat is the sampling unit). For data of this type, the lower limit is, obviously, 0 counts per quadrat, but there is no theoretical upper limit to the number of counts. Phytopathological data such as 'number of lesions per leaf' or entomological data such as 'number of larvae per plant' are examples of counts. Such data may be obtained from an unrestricted random sample, in which every sampling unit (quadrat, leaf, or plant) in the population being studied had an equal chance of being assessed. For the data in Campbell and Madden's (1) Table 11.2, the estimated mean and variance of the sample are, respectively, 4.5 and 10.66.

The fact that the variance of this sample exceeds the mean is informative. Random counts are often described by a Poisson distribution, for which the variance is equal to the mean. Since, in this case, the observed variance is larger than the mean, a Poisson distribution is unlikely to provide a good description of the observed frequency distribution of counts per quadrat, and aggregation (spatial heterogeneity) is indicated. Thus, for count data, there is information about aggregation in an unrestricted random sample. To quantify this information, one could, for example, fit a discrete, two-parameter statistical probability distribution (such as the negative binomial distribution) to the data. Campbell and Madden (1) show that the negative binomial distribution with $k = 2.72$ is a better fit to their data than the Poisson distribution with the same mean. Indeed, this was the reason for the example in the textbook. If several such data sets were available, covering a range of mean values, an alternative to quantifying information about aggregation by fitting a separate negative binomial distribution to each data set would be, instead, to characterize the coefficients of Taylor's (26) power-law variance-mean relationship. Formulae for calculation of optimum sample size based on the Poisson distribution, the negative binomial distribution, and Taylor's power-law, with unrestricted random sampling, have previously been published (1,12,13,24,28).

Now consider disease incidence data. Entomologists might collect similar data by scoring the proportion of plants, or plant units, damaged (e.g., 6). Suppose that of $n = 360$ individual plants scored as either healthy or diseased in an unrestricted random sample, 261 fell in the former category and 99 in the latter. Whether a plant is diseased is described in terms of a random variable, $X$, that may take one of two values, corresponding to 'healthy' or 'diseased.' For convenience, the two values that $X$ may take are denoted $X = 0$ for healthy and $X = 1$ for diseased. The probability distribution of $X$ is then $P(X = x) = p^x(1 - p)^{1-x}$; $x = 0,1$. This is the Bernoulli distribution, and $p$ is the probability of a plant being diseased (4). For the sample in question, $p$ is estimated by the mean incidence, $\Sigma X_i/n = 0.275$ ($X_i = 0$ for a healthy plant, 1 for a diseased plant, $i = 1,2,\ldots,n$; with $n = 360$ here). The theoretical Bernoulli variance is $p(1 - p)$, which for the sample in question is equal to $0.275 \cdot 0.725$ (=0.20 correct to two decimal places). The observed variance of the sample is estimated by $\Sigma(X_i - p)^2/(n - 1)$, which is also equal to 0.20 correct to two decimal places. The two variances are identical if $n$ is used as the denominator in the calculation of the observed variance. This simple illustration shows that, for incidence data, there is no information about aggregation in an unrestricted random sample. Regardless of the actual spatial pattern, the observed variance is equal to the theoretical Bernoulli variance. In order to characterize aggregation in disease incidence data, an alternative approach is required. This will then allow aggregation to be taken into account in the determination of sample size.

**Cluster sampling.** In fact, the above example, based on data from Snedecor and Cochran (25, Table 21.5.1), provides data from a cluster sampling procedure. In cluster sampling, the sampling unit is not the individual (plant, in this case), but a group ('cluster') of individuals. With disease incidence data, each individual in a cluster is classified as healthy or diseased. For data of this type, the lower limit is 0 (diseased individuals) per cluster, but unlike count data, there is an upper limit, when every individual in a cluster is diseased. From the number of individuals in a cluster, and the number diseased, the proportion of individuals diseased can be calculated.

In Snedecor and Cochran's example, the 360 plants came from $N = 40$ clusters (the sampling units were, in this case, quadrats), each of $n = 9$ plants. The product $Nn$ is the total number of individuals (in this case, plants) sampled. The numbers of diseased plants out of 9 were 2, 5, 1, 1, 1, 7, 0, 0, 3, 2, 3, 0, 0, 0, 7, 0, 4, 1, 2, 6, 0, 0, 1, 5, 4, 0, 1, 4, 2, 6, 0, 2, 4, 1, 7, 3, 5, 0, 3, and 6. Now, the sample mean incidence is estimated by $p = \Sigma p_i/N = 0.275$ (as before), and the observed sample variance of the proportions is estimated by $v_{obs} = \Sigma(p_i - p)^2/(N - 1) = 0.067$ (where the $p_i$ are the

Corresponding author: G. Hughes; E-mail address: eagr11@castle.ed.ac.uk

Publication no. P-1995-1212-02O

proportions of plants diseased in each quadrat, $i = 1,2,\ldots,N$; with $N = 40$ here). These estimates of $p$ and $v_{obs}$ can also be obtained from the numbers of diseased plants per quadrat. In this case, sample mean incidence is estimated by $p = \Sigma X_i/Nn$ (=0.275), and the observed sample variance of the proportions is estimated by $v_{obs} = \Sigma(X_i - np)^2/n^2(N - 1)$ (=0.067) (where the $X_i$ are the numbers of diseased plants per quadrat, $i = 1,2,\ldots,N$; with $N = 40$ here). Note that with unrestricted random sampling for disease incidence, $X_i$ could only take one of two values, but with cluster sampling for disease incidence, $X_i$ can take any of $n + 1$ integer values, from 0 to $n$, where $n$ is the size of the sampling unit.

For grouped data, random proportions are often described by a binomial distribution, for which the theoretical variance is, in practice, estimated by $v_{bin} = p(1 - p)/n$, in which $n$ is the number of individuals in a sampling unit. In fact, this formula provides an estimate that is slightly biased and ignores the finite population correction (3). Note that the formula for $v_{bin}$ does not involve $N$. The binomial variance for the sample in question is equal to $(0.275 \cdot 0.725)/9$ (=0.022). Since the observed variance is in excess of the theoretical binomial (random) variance (i.e., $0.067 > 0.022$), aggregation of diseased plants is indicated. These variances can be used in significance tests of departure from a random pattern of diseased plants (2,17).

It is important to realize that count data and incidence data have different statistical properties and that these properties lead to distinct methods for spatial pattern assessment. Under most circumstances (and especially when $p$ is larger than about 0.1), assessments of pattern based on indices such as the variance-to-mean ratio, the $k$ parameter of the negative binomial distribution, and Lloyd's indices of mean crowding and mean patchiness (5) will be erroneous when applied to disease incidence data (discussed in 17).

Aggregation can be thought of as the tendency for plants that are in the same sampling unit to have the same disease status. Mak (20) showed that the probability that any two members of the same sampling unit have the same status ($p_s$) is given by $p_s = 1 - 2p(1 - p)(1 - \rho)$, where $\rho$ is the intracluster correlation coefficient. For any given value of mean disease incidence ($p$), $p_s$ increases with $\rho$. The tendency for plants that are in the same sampling unit to have the same disease status can, therefore, be measured directly by estimating $\rho$.

The calculation of $\rho$ (4, section 6.3) can be rather laborious, since for data comprising $N$ clusters, each of size $n$, $\rho$ is a correlation among $Nn(n - 1)$ pairs (note, however, it is not a requirement for all clusters to be the same size). For Snedecor and Cochran's data, $\rho = 0.24$, based on 2,880 pairs (this calculation was carried out using a set of MINITAB [22] command files for the calculation of $\rho$, available from the authors on request). When $\rho = 0$, this indicates that the disease status of one plant in a cluster does not influence the disease status of other plants in the same cluster. Values of $\rho$ greater than 0 characterize aggregation (the upper limit of $\rho$ is 1). Three methods of characterizing variance inflation due to aggregation, by incorporating $\rho$, are outlined below.

First, Kish (14) defined the *deff* (design effect) as the ratio of the actual variance of a sample to the variance of an unrestricted random sample of the same size and showed that, to a good approximation, $deff = 1 + \rho(n - 1)$. For Snedecor and Cochran's data, the *deff* is calculated by $v_{obs}/v_{bin}$, using, for both variance estimates, the formulae that reflect the cluster sampling procedure by which the data were collected. Thus, $deff = 0.067/0.022 = 3.05$, which corresponds to $\rho = 0.26$.

Second, just as the negative binomial distribution may be used to characterize aggregation in count data, the beta-binomial distribution may be used to characterize aggregation in incidence data (10). For Snedecor and Cochran's data, the maximum likelihood estimate of the beta-binomial aggregation parameter $\theta$ is 0.334 (16), and $\rho = \theta/(1 + \theta) = 0.25$. Note that the theoretical beta-binomial variance, written in terms of $\rho$, is equal to $p(1 - p)(1 + \rho[n - 1])/n$: equivalent to the binomial variance ($p[1 - p]/n$) multiplied by a 'heterogeneity factor' $(1 + \rho[n - 1])$ that is, in effect, the *deff*. There is a discrepancy between the empirically calculated estimate of $\rho$ and the estimate of $\rho$ based on the assumption of a beta-binomial distribution of number of diseased plants per quadrat because the observed data are not perfectly fit by the beta-binomial distribution.

Third, just as Taylor's power-law provides a method of summarizing aggregation in count data from several data sets covering a range of mean values, an analogous relationship may be used to characterize aggregation in incidence data from several data sets (9). If $v_{obs} = Av_{bin}{}^b$ (in which $A$ and $b$ are parameters to be estimated), it can be shown that:

$$\rho = \frac{n}{n-1}\left(\frac{a}{f(p)} - \frac{1}{n}\right) \tag{1}$$

in which $a = An^{-b}$ and $f(p) = (p[1 - p])^{1-b}$. When both $A$ and $b = 1$, $\rho = 0$ (and the *deff* is equal to 1). When $A > 1$ and $b = 1$, $\rho > 0$, but aggregation, as characterized by $\rho$, does not vary with mean incidence. That is to say, both $\rho$ and the *deff* are constant when $b = 1$. When both $A$ and $b$ are greater than 1, this indicates that aggregation varies with mean incidence in such a way that $\rho$ is smallest when $p$ is close to 0 or 1 and largest around $p = 0.5$. Since $\rho$ can be written as a function of $A$, $b$, $p$, and $n$, the relationship $v_{obs} = Av_{bin}{}^b$ effectively characterizes aggregation by the systematically changing shape of an underlying set of frequency distributions that are sampled when mean incidence is estimated.

**Sample size determination.** In a deservedly well-known paper, Karandinos (13) gave formulae for determination of sample size with a prespecified degree of reliability. Reliability was defined either by the coefficient of variation or by setting one-half the length of the required confidence interval of the estimated mean equal either to a fixed proportion of the mean or to a fixed positive number. Various versions of the formulae were presented, appropriate for different distributional assumptions about the data. Subsequently, Wilson and Room (28) discussed the appropriate formulae if Taylor's power-law described the data. The formulae given in Table 1 of Karandinos (13) and later in Tables 3.2 and 3.3 of Ruesink (24), Table 2 of Ives and Moon (12), and sections 13.5.4.1 and 13.4.5.2 of Campbell and Madden (1) are primarily intended for use with count data and are appropriate for unrestricted random sampling. Here, the equivalent formulae appropriate for cluster sampling disease incidence data are derived. The formulae presented are appropriate for randomly arranged clusters but may be modified for other arrangements.

First, it is necessary to review briefly the ground covered by Karandinos (13).

1) If reliability is defined by the coefficient of variation ($C$), the relationship:

$$C = se(p)/p \tag{2}$$

in which $p$ is mean disease incidence and $se(p)$ is its standard error, provides a basis for sample size determination.

2) If reliability is to be defined by a formal probabilistic statement, then, by appeal to the central limit theorem, a confidence interval for $p$ can be written as:

$$p \pm z_{\alpha/2} \cdot se(p)$$

in which $z_{\alpha/2}$ is the upper $\alpha/2$ point of the standard normal distribution. For a 95% confidence interval ($1 - \alpha = 0.95$), $z_{\alpha/2} = 1.96$. Karandinos gave two ways of using this as a basis for sample size determination.

a) If half the length of the required confidence interval is set equal to a fixed proportion ($H$) of the mean ($p$):

$$z_{\alpha/2} \cdot se(p) = Hp \tag{3}$$

b) If half the length of the required confidence interval is set equal to a fixed positive number ($h$):

$$z_{\alpha/2} \cdot se(p) = h \qquad (4)$$

Now, in order to use these relationships, the appropriate formulae for $se(p)$ under different assumptions about the sampling distribution of $p$ are required. In the formulae that follow (equations 5, 6, and 7), it is assumed that $p$ and $se(p)$ are based on $N$ clusters each of size $n$. If the binomial distribution is appropriate:

$$se(p) = \sqrt{[p(1-p)]/nN} \qquad (5)$$

(11). If the beta-binomial distribution is appropriate:

$$se(p) = \sqrt{p(1-p)[1+\rho(n-1)]/nN} \qquad (6)$$

(11), in which ρ is the intracluster correlation coefficient. Equation 6 reduces to equation 5 if $\rho = 0$. If the relationship $v_{obs} = Av_{bin}^{\,b}$ is appropriate:

$$se(p) = \sqrt{a[p(1-p)]^b/N} \qquad (7)$$

in which $a = An^{-b}$. Equation 7 reduces to equation 5 when $A = 1$ and $b = 1$. Obviously, equation 7 may also be derived by substituting equation 1 into equation 6. In each of equations 5, 6, and 7, $se(p)$ is the square root of the appropriate theoretical variance divided by the square root of $N$.

Equations 5, 6, and 7 can now be combined, in turn, with equations 2, 3, and 4 to produce formulae for optimum sample size determination under different definitions of reliability and different assumptions about clustering. These formulae (equations 8, 9, 10, 11, 12, 13, 14, 15, and 16) are shown in Table 1. The formulae in Table 1 are written so that $N$ (the number of clusters [sampling units] required to estimate $p$ with a prespecified degree of reliability) rather than $nN$ (the total number of individuals to be sampled) is calculated. This is because the intracluster correlation coefficient (ρ) will usually vary with cluster size ($n$). Hence, it is preferable to retain the same $n$ as was used to make a preliminary estimate of the variance of disease incidence. If it is not possible to control cluster size so that it is constant, the mean cluster size can instead be used in the calculation.

**Examples.** Two examples are presented, based on previous studies of spatial pattern for fungal diseases of grape (*Vitis* spp.) (18,23). The first example concerns Eutypa dieback (caused by *Eutypa lata*). A total of 22 assessments of Eutypa dieback incidence were made in 8 vineyards over a 3-year period (23). Each assessment provided a map of disease incidence, based on whether each individual vine exhibited Eutypa dieback symptoms. Thus, vine disease incidence was assessed. For the purpose of the present study, the map for each disease assessment was divided into $N$ quadrats ($128 \le N \le 336$), each of $n = 9$ vines. The frequency distribution of diseased vines per quadrat was compiled for each disease assessment, and the observed and binomial variances of disease incidence for each frequency distribution then were calculated. Figure 1 shows the relationship between the observed variance ($v_{obs}$) and the binomial variance ($v_{bin}$). The position of the data above the line for the binomial distribution indicates that diseased vines had an aggregated pattern. The fact that the regression line describing the data is more-or-less parallel to the line for the binomial distribution indicates that, in this case, aggregation, as characterized by the intracluster correlation coefficient ρ, was effectively constant over the range of mean disease incidence. Aggregation in the data may, thus, be described equally well, either by the relationship $v_{obs} = 1.25 \cdot v_{bin}^{\,0.97}$, or by fitting beta-binomial distributions to the data with separate means for each disease assessment but with a common value of θ. The latter procedure resulted in the common value $\theta = 0.053$ (from which $\rho = 0.05$).

Figure 2 shows 'sampling curves' (relationships between $N$, the required number of sampling units [in this case quadrats], and $p$, mean disease incidence) based on the above analyses, when (for example) reliability is defined by setting one-half of the 95% confidence interval ($(1-\alpha) = 0.95$, $z_{\alpha/2} = 1.96$) for mean incidence ($p$) equal to a fixed proportion (for this example, $H = 0.2$) of $p$. Note that this is effectively the same as defining reliability by the coefficient of variation, with $C = 0.1$. For all practical purposes, the number of quadrats required here is the same whether the calculation is based on the relationship between the observed and binomial variances (equation 15) or on the beta-binomial distribution (equation 12). The disease incidence data provide evidence of an aggregated pattern of diseased vines, so the calculation of the required number of quadrats provides too low a value when based on the binomial distribution (equation 9; Fig. 2). Where there is evidence of aggregation, it would be misleading to claim that an estimate of mean disease incidence had a particular reliability status if the required number of quadrats had been calculated on the basis of the binomial distribution.

Suppose, now, that $p$ is expected to be about 0.2. Using Figure 2, equation 9 (the binomial case) gives $N = 43$ as the number of quadrats (each of 9 vines) to be assessed. Equations 12 and 15 both give $N = 60$, the larger value reflecting aggregation of dis-

TABLE 1. Formulae[a] for calculation of number of clusters ($N$), each of size $n$, required for estimation of mean disease incidence ($p$) with a prespecified degree of reliability

| Sampling distribution descriptor | Reliability defined by | | | | | |
|---|---|---|---|---|---|---|
| | | | Probabilistic statement: half length of confidence interval equal to | | | |
| | Coefficient of variation[b] | | Proportion of mean[c] | | Fixed positive number[d] | |
| Binomial distribution | $N = \dfrac{1-p}{npC^2}$ | (8) | $N = \dfrac{1-p}{np} \cdot \left(\dfrac{z_{\alpha/2}}{H}\right)^2$ | (9) | $N = \dfrac{p(1-p)}{n} \cdot \left(\dfrac{z_{\alpha/2}}{h}\right)^2$ | (10) |
| Beta-binomial distribution or empirical *deff*[e] | $N = \dfrac{(1-p)[1+\rho(n-1)]}{npC^2}$ | (11) | $N = \dfrac{(1-p)[1+\rho(n-1)]}{np} \cdot \left(\dfrac{z_{\alpha/2}}{H}\right)^2$ | (12) | $N = \dfrac{p(1-p)[1+\rho(n-1)]}{n} \cdot \left(\dfrac{z_{\alpha/2}}{h}\right)^2$ | (13) |
| $v_{obs} = Av_{bin}^{\,b}$ (with $a = An^{-b}$) | $N = \dfrac{a\,p^{\iota-2}(1-p)^b}{C^2}$ | (14) | $N = a\,p^{b-2}(1-p)^b \cdot \left(\dfrac{z_{\alpha/2}}{H}\right)^2$ | (15) | $N = a[p(1-p)]^b \cdot \left(\dfrac{z_{\alpha/2}}{h}\right)^2$ | (16) |

[a] The formulae omit the finite population correction. This should be included if the sampling fraction is more than 10% (3).
[b] To derive each of the formulae in this column (equations 8, 11, and 14), substitute each of equations 5, 6, and 7, in turn, into equation 2.
[c] To derive each of the formulae in this column (equations 9, 12, and 15), substitute each of equations 5, 6, and 7, in turn, into equation 3.
[d] To derive each of the formulae in this column (equations 10, 13, and 16), substitute each of equations 5, 6, and 7, in turn, into equation 4.
[e] Design effect.

eased vines as characterized by Figure 1. If the area to be sampled comprises a block of 2,500 vines, the population comprises, in this case, 277 quadrats. Since the sampling fraction ($f$ [=60/277 in this case]) is more than 10% of the population, the finite population correction (3) should be employed in the calculation of the required number of quadrats. This involves multiplying $se(p)$ by the factor $\sqrt{(1-f)}$, in whichever of equations 5, 6, or 7 is appropriate. The effect is to reduce the required $N$ to $(1-f)N$, in this case from 60 to 47. This is the number of (randomly selected) quadrats to be assessed from the population in this example, when $p$ is expected to be about 0.2 and reliability is defined as in Figure 2.

The second example concerns grape downy mildew (caused by *Plasmopara viticola*). A total of 108 assessments of downy mildew incidence were made in a vineyard, in plots subjected to different fungicide regimes, in three different years (18). Each assessment provided a record of the total number of leaves per shoot, and the number of leaves diseased, for each shoot observed. Thus, leaf disease incidence was assessed, with shoots as the sampling unit. For each assessment, $N = 15$ shoots were observed, and the disease status of each leaf was recorded. Since the number ($n$) of leaves per shoot (i.e., individuals per sampling unit) varied ($5 \leq n \leq 30$), a formula allowing for variable cluster size (3) was used in the calculation of the observed variance of disease incidence for each assessment. The binomial variance of disease incidence for each assessment was calculated using the mean value of $n$. The overall relationship between the observed variance ($v_{obs}$) and the binomial variance ($v_{bin}$) was $v_{obs} = 8.52 \cdot v_{bin}^{1.30}$ (18, Table 2). The binomial variance is raised to a power greater than 1, indicating that aggregation of diseased leaves is greatest at values of $p$ around 0.5 and smallest at values of $p$ close to 0 or 1. Figure 3 shows sampling curves based on this analysis (equation 15) and on the binomial distribution (equation 9). In this case, $N$ is the number of shoots required for the estimation of $p$ (mean disease incidence), using the same reliability criteria as for Figure 2 (one-half of the 95% confidence interval for $p$ is set equal to a fixed proportion, $H = 0.2$, of $p$). Madden et al. (18, Fig. 4) give a similar example. Calculations of the required number of shoots based on the binomial distribution provide misleadingly low values in the middle of the range of mean disease incidence but may be adequate at very low (and very high) levels of mean incidence.

Of the 108 assessments of downy mildew incidence, 75 provided data in which the number of diseased leaves per shoot could be described by a beta-binomial distribution (with $0.003 \leq \theta \leq$

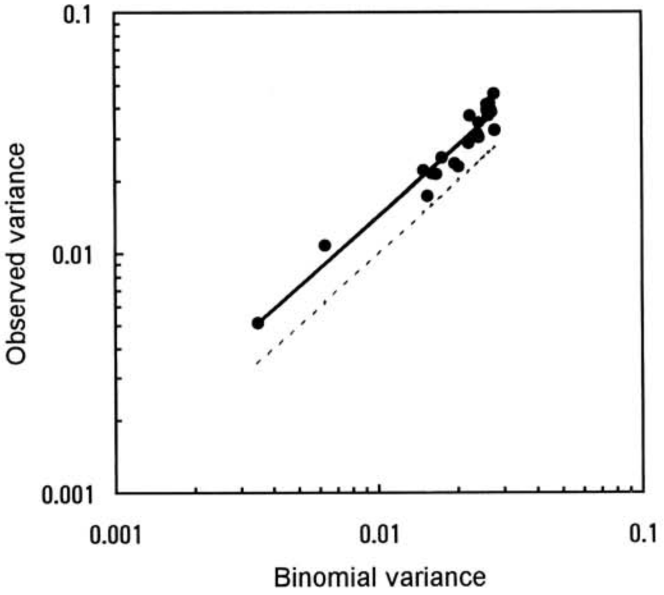

**Fig. 1.** Relationship between the observed variance and the theoretical variance for a random pattern (binomial distribution) for incidence of Eutypa dieback of grapevine, caused by *Eutypa lata* (note the use of logarithmic scales on both axes). Each point represents one disease assessment at a vineyard. The solid line represents the ordinary least squares regression line fitted to the points ($\log[v_{obs}] = 0.096 + 0.97 \cdot \log[v_{bin}]$), and the broken line represents the binomial line (i.e., observed variance = theoretical binomial variance). Munkvold et al. (23) contains information on disease assessments.

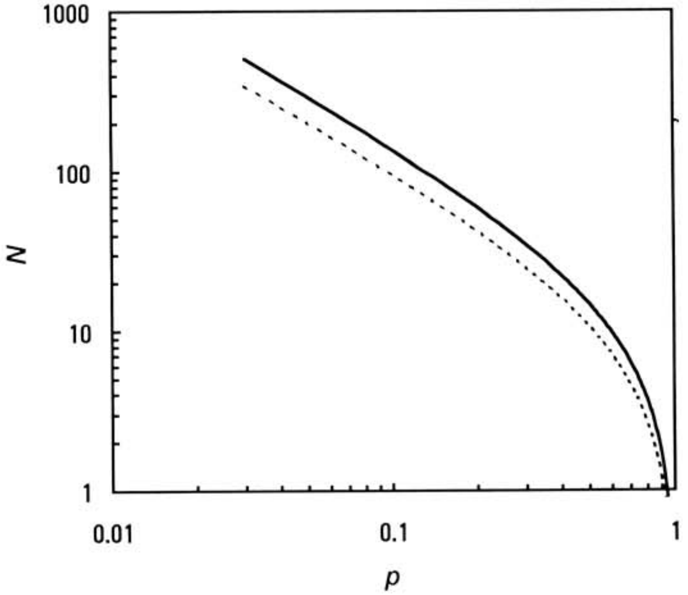

**Fig. 2.** Sampling curves for the number of vines required ($N$) for estimating mean incidence of Eutypa dieback ($p$), with reliability defined by one-half of the 95% confidence interval for $p$ being set equal to a fixed proportion ($H = 0.2$) of $p$. The solid line is based on equation 15 (Table 1), with $a$ and $b$ based on the coefficients of the regression line fitted to the points in Figure 1. In this case, the corresponding line based on equation 12 (Table 1) with $\rho = 0.05$ and $n = 9$ is indistinguishable from the line based on equation 15. The broken line is for the binomial case and is based on equation 9 (Table 1) with $n = 9$.

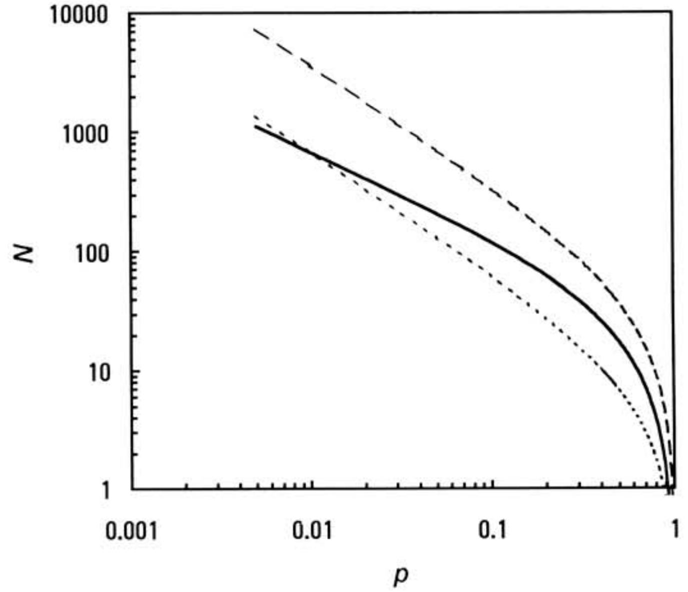

**Fig. 3.** Sampling curves for the number of grape shoots required ($N$) for estimating mean incidence of grape downy mildew ($p$), caused by *Plasmopara viticola*, with reliability defined by one-half of the 95% confidence interval for $p$ being set equal to a fixed proportion ($H = 0.2$) of $p$. The solid line is based on equation 15 (Table 1), with $a$ and $b$ based on the coefficients of the ordinary least squares regression line fitted to the relationship between the observed variance and the theoretical variance for a random pattern (binomial distribution) for 108 assessments of incidence of grape downy mildew (18, Table 2). The upper broken line (– – –) is based on equation 12 (Table 1) with $\rho = 0.34$ and $n = 14$. The lower broken line (---) is for the binomial case and is based on equation 9 (Table 1) with $n = 14$. Madden et al. (18) contains information on disease assessments.

0.536), 25 provided data in which the number of diseased leaves per shoot could be described by a binomial distribution ($\theta = 0$), and in 8 assessments, disease incidence was 0. Thus, for the data set as a whole, $0 \leq \rho \leq 0.35$. If, however, beta-binomial distributions are fitted to these data, with separate means for each disease assessment but with a common value of $\theta$, as above with the Eutypa dieback data, the value $\theta = 0.509$ is obtained (from which $\rho = 0.34$). The latter procedure overestimates aggregation in the data and leads to the calculation of a required number of quadrats (using equation 12) far larger than actually needed, particularly at low levels of mean incidence (Fig. 3). If the intracluster correlation coefficient, $\rho$, is being used as a basis for the calculation of optimum sample size, caution should be exercised about the pooling of data from different disease assessments to obtain an estimate of $\rho$, especially if the assessments cover a wide range of values of mean incidence. In general, such pooling will only be valid if there is some evidence that $\rho$ does not vary with mean incidence, as with the Eutypa dieback data discussed previously.

**Concluding remarks.** Spatial heterogeneity is a function of scale (15). It is to be expected that estimated values of the beta-binomial aggregation parameter ($\theta$), the intracluster correlation coefficient ($\rho$), and the empirical *deff* will depend, to some extent, on the size of the sampling unit. The scale-dependence of $b$ in the relationship $v_{obs} = Av_{bin}{}^{b}$ has not yet been investigated, but Yamamura's work (29) on the scale-dependence of Taylor's power-law (for count data) suggests that the numerical value of this coefficient will probably also be influenced by the size of the sampling unit. Therefore, one question that may arise in connection with cluster sampling concerns the choice of an appropriate size of cluster (sampling unit).

In practice, the observational scale for disease assessment will depend on both the nature of the disease and on the objective of the investigator. In some cases, there will be a natural sampling unit: for example, the shoot is a natural unit for the assessment of leaves for grape downy mildew incidence. However, when whole plants are assessed as healthy or diseased (as with the Eutypa dieback data), the sampling unit (a quadrat) does not necessarily have an obvious, natural size. On purely statistical grounds, Cochran (2) suggested that examination of the pattern of diseased plants was facilitated by dividing an area into quadrats, each containing the same number of plants, between 6 and 12 per quadrat. From a biological point of view, methods for identification of natural spatial scales of epidemics, discussed by Campbell and Madden (1, section 11.4.2.4), may be useful, though care will be required to ensure that the statistical approach adopted is compatible with the analysis of disease incidence data. Examples of valid practices include the work of Gottwald et al. (7), who calculated the ratio $v_{obs}/v_{bin}$ at a range of quadrat sizes as part of a study of spatial pattern of sharka disease in stone-fruit orchards in eastern Spain, and Madden et al. (19), who calculated the beta-binomial aggregation parameter at a range of quadrat sizes as part of a study of spatial pattern of aster yellows in lettuce fields in Ohio.

Like other formulae for sample size determination (13), those in Table 1 require some preliminary parameter estimates. Specifically, some idea of the mean and variance of disease incidence is required. For the binomial distribution (i.e., when the pattern of disease incidence is indistinguishable from random), only a preliminary estimate of $p$ is required, since the variance can then be calculated for any specified value of $n$.

With aggregated patterns of disease incidence, preliminary estimates of both the mean and variance of disease incidence are required. The latter depends on the intracluster correlation coefficient, $\rho$. If the data are known to fit the beta-binomial distribution, $\rho$ can be estimated by $\theta/(1 + \theta)$. If no distributional assumption can be made about the data, a knowledge of the empirical *deff* fulfills the requirement for information about $\rho$. However, aggregation (whether assessed by $\rho$, $\theta$, or the *deff*) often varies with mean disease incidence. If the relationship $v_{obs} = Av_{bin}{}^{b}$ has been established, the resulting formulae for sample size determination have the advantage of being equally applicable over the whole range of mean disease incidence.

It is clear from the sampling curves shown in Figures 2 and 3 that the required $N$ may be large, especially when $p$ is small and disease is aggregated. In such circumstances (as noted by McArdle [21] in a related context), few investigators will accept the sample sizes required. One possibility would be to rearrange the formulae in Table 1 so that they provide an estimate of the degree of reliability that can be achieved by taking $N$ clusters, each of size $n$, where $N$ is determined by the budget available for sampling. Alternatively, methods based on sequential (27) or inverse (8) sampling may be appropriate. Note, however, these methods require that disease symptoms are assessed in the field during sampling and that this is not always the case (e.g., 7). The formulae in Table 1 are appropriate as sampling guidelines whether or not the investigator is able to determine the proportion or number diseased at the time that sampling is taking place.

## LITERATURE CITED

1. Campbell, C. L., and Madden, L. V. 1990. Introduction to Plant Disease Epidemiology. John Wiley & Sons, New York.
2. Cochran, W. G. 1936. The statistical analysis of field counts of diseased plants. Suppl. J. R. Stat. Soc. 3:49-67.
3. Cochran, W. G. 1977. Sampling Techniques. 3rd ed. John Wiley & Sons, New York.
4. Collett, D. 1991. Modelling Binary Data. Chapman & Hall, London.
5. Delp, B. R., Stowell, L. J., and Marois, J. J. 1986. Field runner: A disease incidence, severity, and spatial pattern assessment system. Plant Dis. 70:954-957.
6. de Ramos, M. B., Joshi, R. C., and Angcla, C. J. 1993. Sample size determination for the stalk-eyed fly *Diopsis longicornis* Macquart (Diptera: Diposidae) damage on rice under natural field conditions. Crop Prot. 12: 610-616.
7. Gottwald, T. R., Avinent, L., Llacer, G., Hermoso de Mendoza, A., and Cambra, M. 1995. Analysis of spatial spread of sharka (plum pox virus) in apricot and peach orchards in eastern Spain. Plant Dis. 79:266-278.
8. Haldane, J. B. S. 1945. On a method of estimating frequencies. Biometrika 33:222-225.
9. Hughes, G., and Madden, L. V. 1992. Aggregation and incidence of disease. Plant Pathol. 41:657-660.
10. Hughes, G., and Madden, L. V. 1993. Using the beta-binomial distribution to describe aggregated patterns of disease incidence. Phytopathology 83:759-763.
11. Hughes, G., and Madden, L. V. 1994. Aggregation and incidence of disease: Some implications for sampling. Aspects Appl. Biol. 37:25-31.
12. Ives, P. M., and Moon, R. D. 1987. Sampling theory and protocol for insects. Pages 49-75 in: Crop Loss Assessment and Pest Management. P. S. Teng, ed. The American Phytopathological Society, St. Paul, MN.
13. Karandinos, M. G. 1976. Optimum sample size and comments on some published formulae. Bull. Entomol. Soc. Am. 22:417-421.
14. Kish, L. 1965. Survey Sampling. John Wiley & Sons, New York.
15. Li, H., and Reynolds, J. F. 1995. On definition and quantification of heterogeneity. Oikos 73:280-284.
16. Madden, L. V., and Hughes, G. 1994. BBD—Computer software for fitting the beta-binomial distribution to disease incidence data. Plant Dis. 78:536-540.
17. Madden, L. V., and Hughes, G. 1995. Plant disease incidence: Distributions, heterogeneity and temporal analysis. Annu. Rev. Phytopathol. 33: 529-564.
18. Madden, L. V., Hughes, G., and Ellis, M. A. 1995. Spatial heterogeneity of the incidence of grape downy mildew. Phytopathology 85:269-275.
19. Madden, L. V., Nault, L. R., Murrall, D. J., and Apelt, M. R. 1995. Spatial pattern of aster yellows in Ohio lettuce fields (Abstr.) Phytopathology 85:1122.
20. Mak, T. K. 1988. Analysing intraclass variation for dichotomous variables. Appl. Stat. 37:344-352.
21. McArdle, B. H. 1990. When are rare species not there? Oikos 57:276-277.
22. Minitab Incorporated. 1991. Minitab Statistical Software: MINITAB Reference Manual. Release 8. Minitab, Inc., State College, PA.
23. Munkvold, G. P., Duthie, J. A., and Marois, J. J. 1993. Spatial patterns of grapevines with Eutypa dieback in vineyards with or without perithecia. Phytopathology 83:1440-1448.
24. Ruesink, W. G. 1980. Introduction to sampling theory. Pages 61-78 in:

Sampling Methods in Soybean Entomology. M. Kogan and D. C. Herzog, eds. Springer-Verlag, New York.

25. Snedecor, G. W., and Cochran, W. G. 1989. Statistical Methods. 8th ed. Iowa State University Press, Ames.
26. Taylor, L. R. 1961. Aggregation, variance and the mean. Nature (Lond.) 189:732-735.
27. Wald, A. 1947. Sequential Analysis. John Wiley & Sons, New York.
28. Wilson, L. T., and Room, P. M. 1982. The relative efficiency and reliability of three methods for sampling arthropods in Australian cotton fields. J. Aust. Entomol. Soc. 21:175-181.
29. Yamamura, K. 1990. Sampling scale dependence of Taylor's power law. Oikos 59:121-125.