

This article is from the  
December 2005 issue of

# *Phytopathology*

published by  
The American Phytopathological Society

For more information on this and other topics  
related to plant pathology,  
we invite you to visit *APSnet* at  
**[www.apsnet.org](http://www.apsnet.org)**



*Healthy Plants • Healthy World*

# Sampling for the Incidence of Aphid-Transmitted Viruses in Snap Bean

Denis A. Shah, Helene R. Dillard, and Brian A. Nault

First and second authors: Departments of Plant Pathology; and third author: Entomology, New York State Agricultural Experiment Station, 630 W. North St., Geneva 14456.

Accepted for publication 29 July 2005.

## ABSTRACT

Shah, D. A., Dillard, H. R., and Nault, B. A. 2005. Sampling for the incidence of aphid-transmitted viruses in snap bean. *Phytopathology* 95: 1405-1411.

Data collected in 2002 and 2003 on *Alfalfa mosaic virus* and *Cucumber mosaic virus* incidences of infection in commercial snap bean fields in New York State were used to develop relationships between disease incidence ( $p_{low}$ ) and sample size while accounting for the inherent spatial aggregation of infected plants observed with these two viruses. For a plan consisting of 300 sampled plants ( $N = 60$  quadrats,  $n = 5$  plants per quadrat), estimating  $p_{low}$  from the incidence of positive groups ( $p_{high}$ ; testing of  $N = 60$  grouped samples) provides the same precision in  $p_{low}$  as

testing 200 plants individually, up to about  $p_{low} = 0.2$ . Above that, the confidence interval width for  $p_{low}$  obtained via group testing becomes markedly larger than the width obtained by testing individual plants. Our results suggest using group testing until  $p_{high}$  is in the range [0.35, 0.59], which corresponds to  $p_{low}$  in [0.1, 0.2]. Results indicate that group testing can be more economical than the testing of individual plants without loss of precision, at lower incidences of infection. The approach presented provides a general framework for sampling and the estimation of incidence of other aphid-transmitted viruses in snap bean.

*Additional keywords:* cluster sampling, hierarchical sampling, virus epidemiology.

Sampling is a fundamental component of disease assessment, but how to sample is no trivial matter. Even with simple random sampling, there is still the question of how many samples should be taken. In developing a sampling plan, one has to balance several factors, which can be broadly categorized into cost, precision, and goal. Cost includes all factors related to the site selection (e.g., fields), sample selection within sites, time (travel to sites, sample procurement, and processing), and labor involved. Added costs include materials associated with serological or molecular diagnostic tools in situations where disease assessments are not made visually. Precision of disease assessment estimates is related to the goal of sampling. For example, researchers may require fairly precise estimates of disease in designed experiments comparing the effects of treatments. Precision requirements may be lower if the goal of sampling is to determine whether disease is below or above some decision-triggering threshold, or could be much higher in the immediate region of the actual threshold (10).

The suitability of a sampling scheme is impacted by the characteristics of the particular crop and disease under study. Snap bean (*Phaseolus vulgaris* L.) crops are sown in relatively large field blocks (>2 ha) at high plant densities (about 43 plants per m<sup>2</sup>). Here, sparse sampling for disease assessment is a necessity, in that incidence estimates inferred to the whole field will be made from a relatively small sample of the plant population within that field. With virus-induced plant diseases, incidence is in many situations much easier to assess than severity. Even so, plants may be infected but asymptomatic (literature citation 22 is an example). Therefore, disease incidence assessed visually may drastically underestimate the true incidence of virus-infected plants (22). In such instances, serological or molecular-based virus diagnostic techniques, though obviously more costly, are far more reliable

than visual assessments. Cost then becomes a significant concern and limiting factor, and one then has to balance the cost of sampling against the quality of the information derived. If serological or molecular-based diagnostic techniques are the methods of choice for assessing disease incidence, then the most practical approach is to collect all samples from a given field (or plot) at the same time, because it is most efficient to process several samples simultaneously back in the laboratory. Sequential sampling schemes (10) are therefore not attractive from a logistical standpoint.

There is growing evidence that many plant diseases are not distributed randomly in the spatial sense, but occur as aggregates of infected individuals (3,17,26 provide recent examples), and this is also true of aphid-transmitted viruses (2,4,11). A sampling scheme for assessing the incidence of virus-infected plants therefore should account for the inherent spatial properties of the disease. Recently, there have been seminal advancements in the theory of sampling and estimation of plant disease incidence (6,9,10,12), and the results are finding their way into practical sampling plans (4,23,24). In this paper, we draw on this recent theoretical work to demonstrate how spatial pattern information is incorporated into a sampling framework for ascertaining the incidence of infection by the aphid-transmitted viruses *Alfalfa mosaic virus* (AIMV) and *Cucumber mosaic virus* (CMV) in processing snap bean.

## MATERIALS AND METHODS

Six early-planted and six late-planted snap bean fields in western New York were sampled in 2002 and 2003. Early plantings occurred from late May through the end of the first week of June, whereas late plantings occurred during the first half of July. Sampled fields are described in more detail elsewhere (14,19). Sampling was done when the fields were at the bloom or early pin stage. In each field, 16 quadrats were indiscriminately sampled, each quadrat consisting of five adjacent plants. This is a form of

Corresponding author: D. A. Shah; E-mail address: das28@cornell.edu

DOI: 10.1094/PHYTO-95-1405

© 2005 The American Phytopathological Society

cluster sampling. Plants were tested for the presence of AIMV and CMV as previously described (19).

**Binary power law.** Spatial pattern analysis had indicated that plants infected by AIMV or CMV could be aggregated within fields (19), and therefore the level of aggregation should be accounted for in a sampling plan for determining virus incidence. The binary power law (5) is useful for quantifying the relationship between aggregation and incidence, because the extent of aggregation could change with the incidence of infected plants (2,4,17,18,23).

The proportion of virus-infected plants ( $x$ ) in a sampled quadrat is given by  $x = X/n$ , where  $X$  is the number of infected plants out of  $n$  sampled plants in the quadrat. If  $c$  is the observed variance in  $x$  in a given field, and  $p_{low}$  is the proportion of infected plants in the field, then the binary power law

$$\ln(c) = \ln(A) + b \ln[p_{low}(1 - p_{low})/n] \quad (1)$$

specifies a linear relationship between  $c$  and  $p_{low}$ , where  $\ln(\Psi)$  is the natural logarithm of  $\Psi$ . The parameters  $A$  and  $b$  can be estimated by linear regression, and their interpretation is discussed in detail elsewhere (10). There were 24 data points available for each virus. We fit the binary power law to the data for each virus (AIMV and CMV) separately and also to the combined data set, using SAS Proc REG (SAS Institute, Cary, NC). Unless otherwise stated, all other modeling was done with Mathematica version 4.1 (Wolfram Research Inc., Champaign, IL).

**Sampling for precision.** If the level of aggregation varies with incidence according to the binary power law, then the number of quadrats ( $N$ ) that ought to be sampled to estimate the mean incidence of infection ( $\hat{p}_{low}$ ) with a desired level of precision ( $C$ ) is given by

$$N = An^{-b} \hat{p}_{low}^{b-2} (1 - \hat{p}_{low})^b / C^2 \quad (2)$$

where  $C$  is the ratio of the standard error of  $\hat{p}_{low}$  (i.e.,  $se[\hat{p}_{low}]$ ) to  $\hat{p}_{low}$  ( $C = se[\hat{p}_{low}] / \hat{p}_{low}$ ). The parameters of the binary power law fitted to the combined data set were used in equation 2 to estimate the number of quadrats required to achieve specified levels of precision in  $\hat{p}_{low}$  for quadrats of size  $n = 5$ .

**Group testing for estimating incidence.** Testing samples individually may not be as efficient as testing groups of plants and then estimating the incidence of infection from the results on the grouped samples. This method, termed group testing, has found application in plant pathology (16,20). So-called hierarchical sampling (6) is a form of group testing conditioned on sampling within a spatial hierarchy. Assume that all samples from a given quadrat are bulked into one sample for that quadrat and then tested by serology. Let  $p_{high}$  represent the incidence of quadrats that were positive for virus infection in the same field. Also, assume the number of virus-infected plants per quadrat is described by the beta-binomial probability distribution, which has now been shown to fit disease incidence data very well (references cited within literature citation 9), and which is a way of accounting for the aggregation of infected plants. The relationship between  $p_{low}$  and  $p_{high}$  is then described by

$$p_{high,E} = 1 - \prod_{i=0}^{n-1} \frac{1 - p_{low} + i\theta}{1 + i\theta} \quad (3)$$

where  $\Pi$  is the product function,  $i$  indexes integers from 0 to  $n - 1$ , and  $\theta$  is the aggregation parameter of the beta-binomial distribution. The subscript  $E$  on  $p_{high,E}$  signifies that this is an exact estimate of  $p_{high}$ , assuming that the beta-binomial distribution fully describes the data. If the binary power law applies, then

$$\theta = \frac{An^{-b} - [p_{low}(1 - p_{low})]^{1-b} / n}{[p_{low}(1 - p_{low})]^{1-b} - An^{-b}} \quad (4)$$

reflecting the variation of  $\theta$  with  $p_{low}$  (9).

Equation 3 is not amenable to rearrangement to allow the prediction of  $p_{low}$  from  $p_{high}$ , except for  $n = 2$  or  $n = 4$  (6). An alternative expression for the relationship between  $p_{low}$  and  $p_{high}$  is

$$p_{high,v} = 1 - (1 - p_{low})^v \quad (5)$$

in which  $v$  is a parameter representing the so-called effective sample size (9). The subscript  $v$  on  $p_{high,v}$  is used to indicate that estimation of  $p_{high}$  is based on the parameter  $v$ . Equation 5 is of the same format as the equation relating  $p_{high}$  to  $p_{low}$  if incidence data are described by the binomial distribution (i.e., there is no aggregation of infected plants), in which case  $v = n$ . In equation 5, aggregation forces the restriction that  $v < n$ . Rearrangement gives  $p_{low}$  as a function of  $p_{high}$ :

$$p_{low,v} = 1 - (1 - p_{high})^{1/v} \quad (6)$$

The subscript  $v$  on  $p_{low,v}$  indicates that this estimate of  $p_{low}$  depends on the parameter  $v$ . It is clear from equations 5 and 6 that an exact expression for  $v$  can be derived, but it will be a function of  $p_{low}$ , which is not a desirable property for predicting  $p_{low}$  from  $p_{high}$  in equation 6. However, a reasonable estimate of  $v$  can be obtained empirically. We used a two-step approach in doing so.

**Step 1.** CLL( $p$ ) is called the complementary log-log transformation of  $p$ , and is given by  $CLL(p) = \ln[-\ln(1 - p)]$ . The relationship between  $p_{low}$  and  $p_{high}$  can then be described by

$$CLL(p_{low}) = \ln(1/v_{CLL}) + CLL(p_{high}) \quad (7)$$

where the subscript on  $v$  indicates that the parameter  $v$  is estimated from the CLL relationship between  $p_{low}$  and  $p_{high}$ . Equation 7 was used to obtain a first estimate and possible range (the 95% profile likelihood function-derived confidence interval) for  $v_{CLL}$ . Equation 7 specifies an intercept of  $\ln(1/v_{CLL})$  and slope = 1. The intercept was estimated using SAS Proc GENMOD with link=identity, dist=normal, and offset=CLL( $p_{high}$ ) options in the model statement. The estimate of  $v_{CLL}$  obtained this way serves only as a general starting point; the CLL relationship based on the beta-binomial distribution is not strictly a straight line, but exhibits curvature at higher values of  $p_{low}$  (9). However, the curvature at higher values of  $p_{low}$  is obvious only at relatively large values of  $\theta$ . We did not discern any curvature in a CLL( $p_{low}$ ):CLL( $p_{high}$ ) plot of our data (data not shown).

**Step 2.** The absolute difference between  $p_{high,E}$  and  $p_{high,v}$  (i.e.,  $|p_{high,E} - p_{high,v}|$ ) is a function of both  $p_{low}$  and  $v$ . To further refine the primary estimate  $v_{CLL}$ , we constructed a three-dimensional plot of  $|p_{high,E} - p_{high,v}|$  versus  $p_{low}$  and  $v$ , with the 95% confidence interval for  $v_{CLL}$  as a starting point. This plot was used to obtain a narrower parameter space for  $v$  (i.e.,  $[v_1, v_2]$ ) which appeared to minimize  $|p_{high,E} - p_{high,v}|$ . We then calculated (by numerical integration) and plotted  $\int_0^{0.6} |p_{high,E} - p_{high,v}| dp_{low}$  over  $[v_1, v_2]$ . The integration was done over the range  $[0, 0.6]$  for  $p_{low}$  because there was not much further change in  $p_{high,E}$  for  $p_{low} > 0.6$  (for which  $p_{high,E} = 0.964$ ). The optimal value for  $v$  ( $v_{opt}$ ) was that value of  $v$  which minimized the integrand.

**Sampling effort and confidence interval widths.** Upper and lower confidence interval points for  $p_{low}$  were obtained by first calculating the respective limits for  $p_{high}$  and then using the monotonic relationship between  $p_{low}$  and  $p_{high}$  over  $[0, 1]$  for fixed  $v$  (21). The upper and lower confidence limits for  $p_{high}$  were calculated using the Wilson score interval which, when applied at the group level, generally provides coverage for  $p_{low}$  closest to the nominal level compared with other methods for determining confidence intervals (21). If  $t$  is the number of quadrats positive for virus infection out of the  $N$  quadrats sampled, then the Wilson interval is given by

$$\frac{t + \frac{1}{2} z_{1-\alpha/2}^2}{N + z_{1-\alpha/2}^2} \pm \frac{z_{1-\alpha/2} N^{1/2}}{N + z_{1-\alpha/2}^2} \sqrt{\frac{t}{N} \left(1 - \frac{t}{N}\right) + \frac{z_{1-\alpha/2}^2}{4N}} \quad (8)$$

where  $z_{1-\alpha/2}$  is the  $1 - \alpha/2$  quantile of the standard normal distribution.  $p_{low}$  was plotted as a function of  $p_{high}$  for  $v_{opt}$  and for when  $v = n$  (using equation 6), along with their respective upper and lower confidence limits (equation 8).

Equations 6 and 8 were used to estimate the width of the confidence interval for  $p_{low}$  when  $N = 16$  and  $n = 5$ , reflecting the sampling scheme we used in 2003 and 2004. One may question whether increasing the sampling effort, in terms of the number of quadrats sampled, pays off in terms of an appreciable increase in the confidence of the estimate of  $p_{low}$  (i.e., a smaller confidence interval width for  $p_{low}$ ) when group testing is used. To address this question, equations 6 and 8 were used to derive an expression for the confidence interval width of  $p_{low}$  ( $CI_{low}$ ) as a function of  $p_{low}$  and  $N$ . Suppressing the subscripts on  $z$  for simplicity,  $CI_{low}$  is given by

$$CI_{low} = \left( \frac{N(1-p_{low})^v + 0.5z^2 + \sqrt{Nz} \sqrt{[1-(1-p_{low})^v](1-p_{low})^v + \frac{z^2}{4N}}}{N+z^2} \right)^{1/v} - \left( \frac{N(1-p_{low})^v + 0.5z^2 + \sqrt{Nz} \sqrt{[1-(1-p_{low})^v](1-p_{low})^v + \frac{z^2}{4N}}}{N+z^2} \right)^{1/v} \quad (9)$$

Equation 9 was used to plot  $CI_{low}$  as a function of  $p_{low}$  for fixed values of  $N$  and  $v = v_{opt}$ .

**Rate of change in the confidence interval width.** Although increasing the number of quadrats sampled leads to more precise estimates of  $p_{low}$ , there may be a point at which further increases in  $N$  lead to only marginal reductions in  $CI_{low}$ , which may not justify the sampling effort and cost. We therefore calculated and plotted the second partial derivative of  $CI_{low}$  with respect to  $N$  (i.e.,  $\partial^2 CI_{low} / \partial N^2$ ) against  $N$  for different fixed values of  $p_{low}$ . The expression for  $\partial^2 CI_{low} / \partial N^2$  is rather long and is not given here. This plot represents how much the rate of change of  $CI_{low}$  depends on  $N$  at different  $p_{low}$ .

**Group versus individual testing.** An important decision is whether to individually test all  $n \cdot N$  samples, or use group testing and then estimate  $p_{low}$  from  $p_{high}$ . Although group testing is less costly in terms of labor and reagents, it can be more costly if the confidence intervals for  $p_{low}$  are wider than one is willing to accept. One way of addressing the question is to determine the minimum number of quadrats to sample so that the confidence interval width for  $p_{low}$  is at least equal to that obtained by testing each of the  $n \cdot N$  samples individually. This was done by solving  $CI_{diff} = 0$  for  $N$ , where  $CI_{diff}$  is the difference in the confidence interval widths for the estimate of incidence obtained by individual and group testing. Aggregation of virus-infected plants within quadrats was accounted for by using  $v = v_{opt}$  in equation 9 when calculating  $CI_{low}$ .

The Wald-based confidence interval ( $\hat{p}_{low} \pm z_{1-\alpha/2} \cdot se[\hat{p}_{low}]$ ) is the simplest method for calculating an associated confidence interval for the estimate of the incidence of virus-infected plants (10). Although the Wald interval is widely used, there is now

sufficient evidence showing that its coverage is chaotic even for large sample sizes (1), and not only when  $p_{low}$  is close to its extreme values ( $p_{low} = 0$  or 1). An alternative is the Jeffreys interval, which has been recommended as a replacement for the Wald interval (1). The Jeffreys interval is given by [ $L_J(p_{low})$ ,  $U_J(p_{low})$ ]:

$$L_{J(p_{low})} = beta \left[ \frac{\alpha}{2}, np_{low} + \frac{1}{2}, n - np_{low} + \frac{1}{2} \right] \quad (10a)$$

$$U_{J(p_{low})} = beta \left[ 1 - \frac{\alpha}{2}, np_{low} + \frac{1}{2}, n - np_{low} + \frac{1}{2} \right] \quad (10b)$$

where  $beta[l, m_1, m_2]$  is the  $l$ th quantile of the  $beta[m_1, m_2]$  distribution. For example,  $\alpha = 0.05$  for a 95% confidence interval.

We used the Jeffreys interval to estimate the confidence interval width for  $\hat{p}_{low}$  (i.e.,  $CI_{J(p_{low})}$ ) when testing was done on plants individually. This is an approximation, as the Jeffreys interval does not account for the aggregation of virus-infected plants, which, as we described above, varies with  $p_{low}$ . It is nevertheless in this case a fair approximation, because  $\theta$  was not very high (described in Results section) and so the appropriate standard error based on the binary power law

$$se_a(\hat{p}_{low}) = \sqrt{\frac{An^{-b} [\hat{p}_{low}(1-\hat{p}_{low})]^b}{N}}$$

was not much different from the standard error obtained assuming no aggregation

$$\left( se_r(\hat{p}_{low}) = \sqrt{\frac{\hat{p}_{low}(1-\hat{p}_{low})}{nN}} \right)$$

For our data, the maximum difference between  $se_a(\hat{p}_{low})$  and  $se_r(\hat{p}_{low})$ , which occurs at  $p_{low} = 0.5$ , was 0.0079. The Jeffreys interval approximation as used here is also computationally more feasible than exact confidence interval methods based on the beta-binomial distribution (25).

## RESULTS

**Binary power law.** The binary power law was a good descriptor of the relationship between the observed and theoretical (binary) variance in the proportion of virus-infected plants per quadrat, for both AIMV and CMV (Table 1). Confidence intervals for the slope ( $b$ ) and intercept [ $\ln(A)$ ] parameters for each virus overlapped, so the data sets were combined and treated as one. For the combined data,  $b$  was  $>1$  (but only slightly) and  $\ln(A)$  was  $>0$  ( $A = 1.762$ ), which meant that the degree of aggregation varied with  $p_{low}$  systematically (10). For  $p_{low}$  in the range  $[0.01, 0.5]$ ,  $\hat{\theta}$  (estimated from equation 4) was between 0.069 and 0.142.

**Sampling for precision.** Figure 1 shows the number of quadrats that should be sampled to achieve specified precision levels in the estimates of  $p_{low}$ . It is clear that for  $C = 0.1$ ,  $N$  becomes prohibitively large (from a practical standpoint) for  $p_{low} < 0.2$ . One could relax the stringency on the required precision levels, and so reduce the number of quadrats to be sampled. However,  $N$  is still quite large for precision levels given by  $C = 0.1$  or 0.2 for estimating incidence levels below 0.1 (Fig. 1).

TABLE 1. Parameter estimates for the binary power law relationship between the observed and theoretical (binary) variances of the proportion of virus-infected plants per quadrat

Virus <sup>a</sup>	Intercept <sup>b</sup>		Slope		R <sup>2</sup>
	Estimate	95% CI	Estimate	95% CI	
AIMV	0.542	(0.169, 0.915)	1.048	(0.981, 1.115)	0.979
CMV	0.606	(0.277, 0.936)	1.065	(0.994, 1.137)	0.978
Combined	0.566	(0.335, 0.798)	1.054	(1.009, 1.100)	0.980

<sup>a</sup> AIMV = *Alfalfa mosaic virus*, CMV = *Cucumber mosaic virus*, and combined = the AIMV and CMV data sets combined.

<sup>b</sup> Estimates and 95% confidence intervals (CI) for the intercept and slope obtained from the linear regression between the observed and theoretical variances in the proportion of virus-infected plants per quadrat. The overall fit of the regression is given by R<sup>2</sup>.

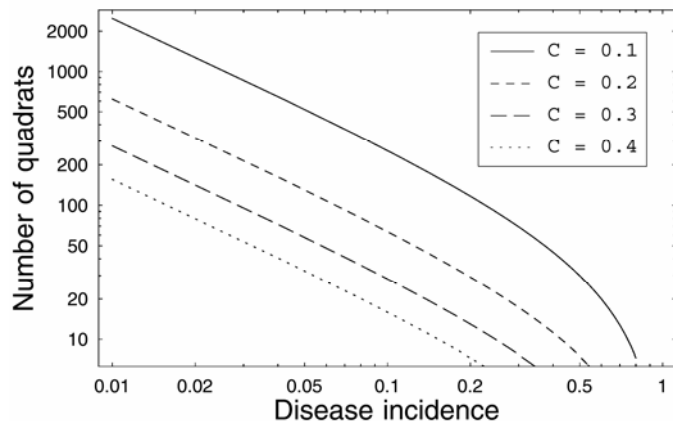
**Group testing for estimating incidence.** The estimate of  $\ln(1/v_{\text{CLL}})$  of equation 7 obtained from SAS Proc GENMOD was  $-1.4047$  (se = 0.0441) with a profile likelihood-based 95% confidence interval of  $[-1.4932, -1.3162]$ . The 95% confidence interval for  $v_{\text{CLL}}$  [3.73, 4.45] was then used in plotting  $|p_{\text{high},E} - p_{\text{high},v}|$  versus  $p_{\text{low}}$  and  $v_{\text{CLL}}$  (Fig. 2A). Using Figure 2A, we narrowed the parameter space  $[v_1, v_2]$  to [3.80, 4.00]. Over this range,  $p_{\text{high},v}$  at first underestimated  $p_{\text{high},E}$  and then overestimated  $p_{\text{high},E}$  as  $p_{\text{low}}$  increased further (Fig. 2B). It was apparent from the shape of the curves with respect to  $v$  in Figure 2B that a decision criterion would have to be established concerning the range of  $p_{\text{low}}$  for which  $p_{\text{high},E} - p_{\text{high},v}$  should be minimized. We chose to minimize  $|p_{\text{high},E} - p_{\text{high},v}|$  over  $p_{\text{low}} \in [0, 0.6]$  because there was not much further change in  $p_{\text{high},E}$  for  $p_{\text{low}} > 0.6$  (i.e.,  $p_{\text{high},E} \approx 1$ ). The value of  $v$  which minimized  $\int_0^{0.6} |p_{\text{high},E} - p_{\text{high},v}| dp_{\text{low}}$  over  $v \in [3.80, 4.00]$  was  $v_{\text{opt}} = 3.903$  (Fig. 2C), which gave a close match between  $p_{\text{high},v}$  and  $p_{\text{high},E}$  (Fig. 2D).

The parameter value  $v_{\text{opt}}$  was used in equation 6, which together with equation 8 was used to plot  $p_{\text{low}}$  as a function of  $p_{\text{high}}$  (Fig. 3). Explicitly accounting for aggregation of virus-infected plants was apparent in the  $p_{\text{low}}:p_{\text{high}}$  curve (the  $v = v_{\text{opt}}$  curve is above that for  $v = n$ ), becoming more so as  $p_{\text{high}}$  increased. Confidence interval limits were shifted as well, more noticeably for the upper confidence limit (Fig. 3).

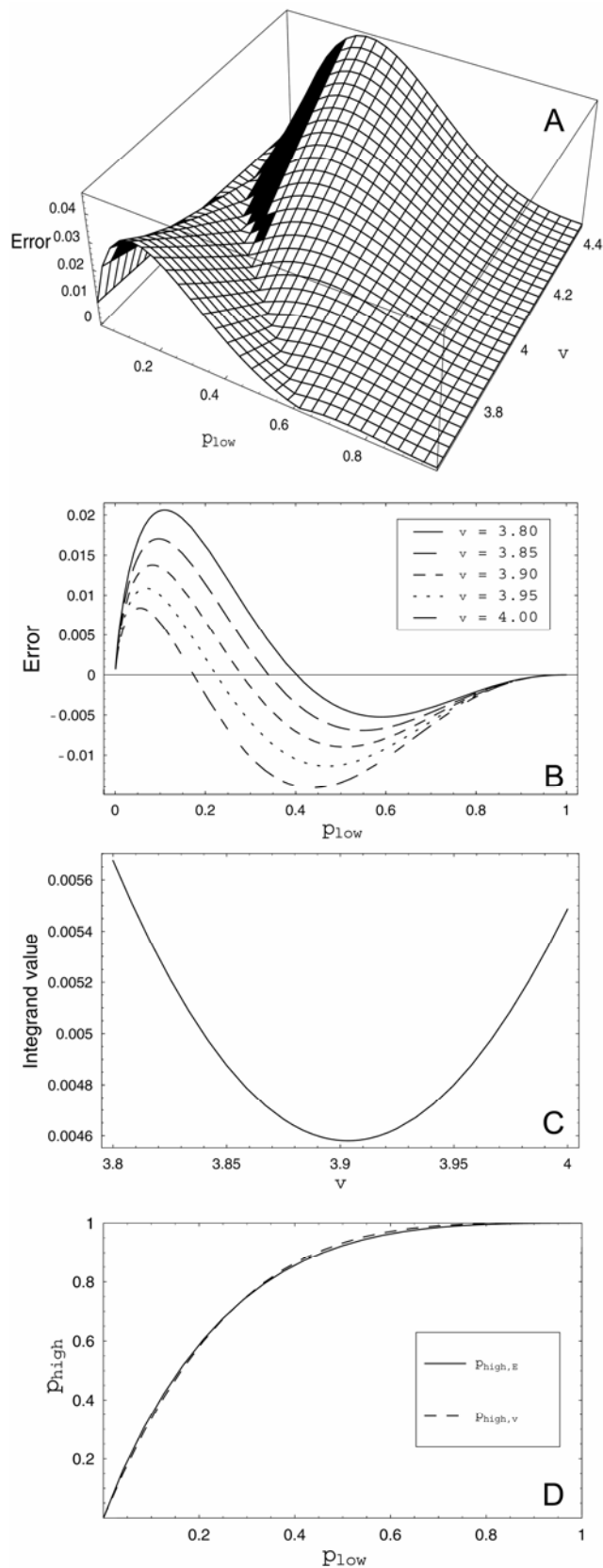
**Sampling effort and confidence interval widths.** It is apparent from Figure 3 that estimating  $p_{\text{low}}$  from  $p_{\text{high}}$  by group testing when  $N = 16$  results in a confidence interval width  $\text{CI}_{\text{low}}$  that is rather large when  $p_{\text{low}}$  is  $>0.1$ . Sampling more quadrats (i.e., increasing  $N$ ) may reduce  $\text{CI}_{\text{low}}$ . Increasing  $N$  had the biggest impact on  $\text{CI}_{\text{low}}$  at the higher levels of  $p_{\text{low}}$  (Fig. 4); for  $p_{\text{low}} < 0.1$ , the reductions in  $\text{CI}_{\text{low}}$  were marginal for  $N$  in the range [16, 60].

**Rate of change in the confidence interval width.** For different values of  $p_{\text{low}}$ ,  $\partial^2 \text{CI}_{\text{low}} / \partial N^2$  decreased in an exponential-like manner with increasing  $N$  (Fig. 5). The largest changes in  $\partial^2 \text{CI}_{\text{low}} / \partial N^2$  were observed up to about  $N = 40$  (Fig. 5). For  $N \in [40, 60]$ , the changes in  $\partial^2 \text{CI}_{\text{low}} / \partial N^2$  were smaller. This suggests that sampling  $N > 40$  quadrats may not provide enough gains in the confidence of the estimate of  $p_{\text{low}}$  to offset the added costs in sampling and sample processing. A sampling effort of  $N = 40$  may be a reasonable sample size for estimating  $p_{\text{low}}$  from  $p_{\text{high}}$ .

**Group versus individual testing.** The above result suggested that  $N = 40$  may be a reasonable upper limit on sampling effort for estimating  $p_{\text{low}}$  from  $p_{\text{high}}$  (corresponding to  $n \cdot N = 200$  plants). However, compared with testing 200 plants individually, it can be seen (Fig. 6A) that, even at low incidences of plant infection (i.e.,  $p_{\text{low}} < 0.1$ ), more than 40 groups would have to be tested to



**Fig. 1.** Number of quadrats ( $N$ ) required to achieve a desired level of precision, as specified by the coefficient of variation of the mean ( $C$ ), in relation to the incidence of virus-infected snap bean plants. Calculations are based on parameter estimates derived from the binary power law relationship (Table 1) and used in equation 2.

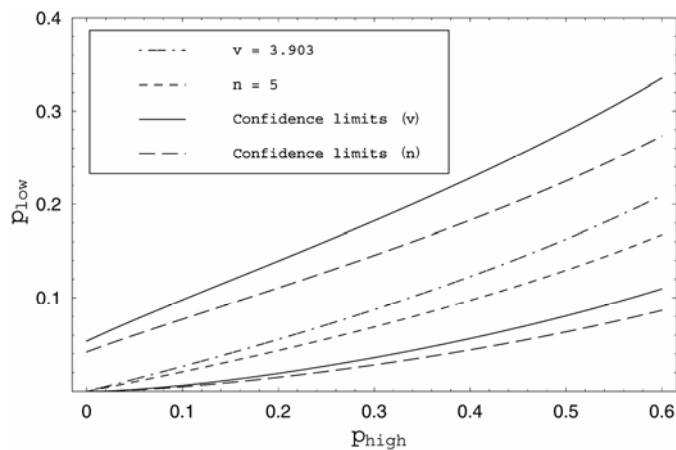


**Fig. 2.** **A**, Three dimensional plot of the absolute error ( $|p_{\text{high},E} - p_{\text{high},v}|$ ) in predicting  $p_{\text{high}}$  as a function of  $p_{\text{low}}$  and  $v$ , the effective sample size. The exact relationship ( $p_{\text{high},E}$ ) is given by equation 3 with  $n = 5$ , and  $\theta$  of the beta-binomial distribution given by equation 4 with  $A = 1.762$  and  $b = 1.054$ . The empirical relationship ( $p_{\text{high},v}$ ) is given by equation 5. **B**, Absolute error in predicting  $p_{\text{high}}$  from  $p_{\text{low}}$  for selected values of  $v$ . **C**, Plot of  $\int_0^{0.6} |p_{\text{high},E} - p_{\text{high},v}| dp_{\text{low}}$  versus  $v \in [3.80, 4.00]$ . **D**,  $p_{\text{high},E}$  and  $p_{\text{high},v}$  (with  $v_{\text{opt}} = 3.903$ ) as a function of  $p_{\text{low}}$ .

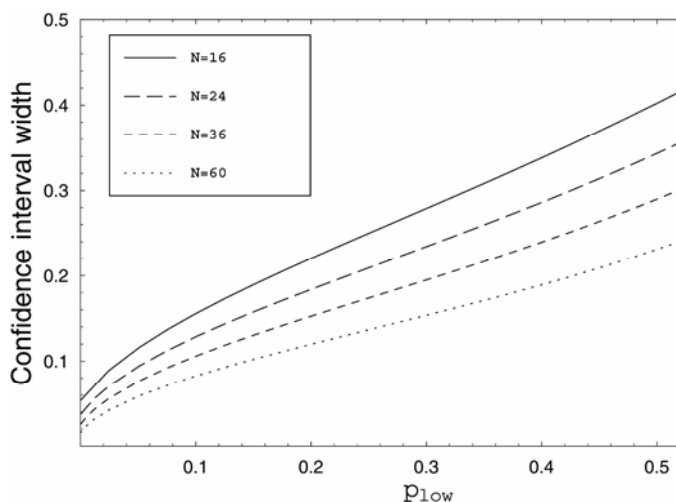
achieve  $CI_{low} \leq CI_{J(p_{low})}$ . In terms of the number of tests that need to be done, group testing will require less serological assays up to about  $p_{low} = 0.5$ . For higher incidences of plant infection, many more groups would need to be tested for attaining  $CI_{low} \leq CI_{J(p_{low})}$  (Fig. 6A). This result can also be shown for other choices of  $n \cdot N$ . Figure 6A suggested that for  $p_{low} < 0.5$  one could achieve  $CI_{low} \leq CI_{J(p_{low})}$  by group testing on  $Q_G < n \cdot N$  quadrats, where  $Q_G$  is the number of quadrats assayed by group testing. In Figure 6B, we plotted  $CI_{low}$  and  $CI_{J(p_{low})}$  when  $Q_G = 60$  (group testing;  $60 \cdot 5 = 300$  plants), and  $n \cdot N = 200$  (individual testing), which represents a 70% reduction in the amount of testing (albeit a 50% increase in sampling) compared with assaying all  $n \cdot N$  samples individually. For  $p_{low} \leq 0.1$ ,  $CI_{low}$  and  $CI_{J(p_{low})}$  were practically indistinguishable. As  $p_{low}$  increased further, so did the difference between  $CI_{low}$  and  $CI_{J(p_{low})}$  (Fig. 6B). However, the graph did suggest that for this particular sampling plan group testing when  $p_{low}$  was  $\leq 0.2$  provided as precise estimates as individual sampling, at a far lower investment in testing resources.

### DISCUSSION

Aphid-transmitted viruses and the diseases they induce are an emerging problem in processing snap bean in the Midwest and

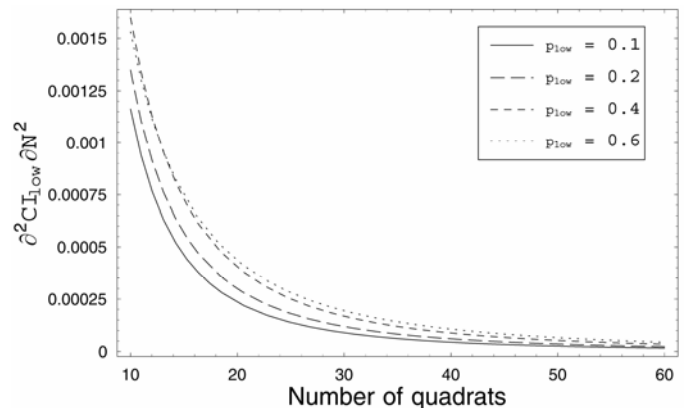


**Fig. 3.**  $p_{low}$  as a function of  $p_{high}$  (equation 6) in which aggregation of virus-infected plants is explicitly accounted for ( $v_{opt} = 3.903$ ) or ignored ( $n = 5$ ), for  $N = 16$  quadrats. The respective lower and upper 95% confidence limits (equation 8) are shown. Plot for  $N > 16$  were similar, except for narrower confidence intervals.

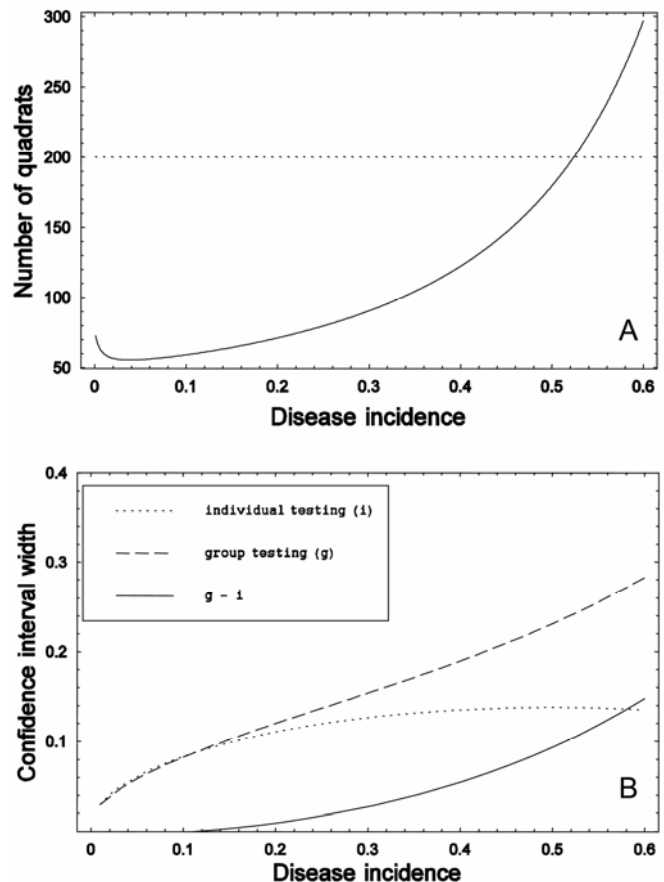


**Fig. 4.** Effect of sampling effort ( $N =$  number of quadrats sampled) on the 95% confidence interval width of the incidence of virus-infected plants ( $p_{low}$ ), when  $p_{low}$  is estimated by group testing (i.e., from  $p_{high}$ ) at the quadrat level.

Northeast regions of the United States (7). As researchers begin to understand which viruses are the major components of the disease complex, efforts will begin to shift from basic assessments of prevalence to more detailed disease assessment surveys, including also designed experiments on basic epidemiology and likely disease



**Fig. 5.** Relationship between the second derivative of the 95% confidence interval width for the incidence of virus-infected plants ( $\partial^2 CI_{low} / \partial N^2$ ) and the number of quadrats sampled ( $N$ ), when  $p_{low}$  is estimated by group testing at the quadrat level.  $\partial^2 CI_{low} / \partial N^2$  measures the rate of the rate of change in  $CI_{low}$  with respect to  $N$ . The relationship is plotted for four fixed values of  $p_{low}$ .



**Fig. 6. A,** Number of quadrats ( $Q_G$ ) to be assayed by group testing so that the derived 95% confidence interval width for the incidence of infection ( $CI_{low}$ ) is the same as the confidence interval width obtained when each plant is tested individually ( $CI_{J(p_{low})}$ ), for  $n \cdot N = 200$  (represented by the dotted line). Note that  $Q_G < n \cdot N$  for  $p_{low}$  less than about 0.5. **B,** Comparison of  $CI_{low}$  and  $CI_{J(p_{low})}$  for group testing when  $Q_G = 60$  ( $n = 5$ ) and individual testing when  $n \cdot N = 200$ . Confidence interval widths were based on equations 8 (with  $v_{opt} = 3.903$ ) and equations 10a and 10b. The solid line is the difference in the confidence interval widths obtained by individual and group testing.

mitigation strategies. Researchers will thus fairly soon require sampling guidelines for assessing virus intensity in snap beans.

Disease incidence is much easier and practical to assess than severity for virus-induced diseases in snap bean, especially in routine field assessments. In this paper, we demonstrate how a sampling plan for the incidence of aphid-transmitted viruses (AIMV and CMV) can be developed from an underlying understanding of the spatial pattern of virus-infected plants in commercial fields. Data on the incidence of snap bean infection by AIMV and CMV collected over two field seasons could be combined into one data set. The conformance in the observed spatial patterns suggests common spatial generative processes via viral transmission and spread in snap bean. Our sampling plan analysis may therefore be extended to provide a generalized basis for the sampling of other aphid-transmitted viruses in snap bean, but not without caution however. Several non-colonizing aphid species disperse throughout snap bean fields during the course of the season (14), and there may be differences in the viruses they vector as well as in transmission efficiencies. The influence of these factors could lead to different levels of aggregation of virus-infected plants, as observed with *Citrus tristeza virus* (4).

Estimating the incidence of virus infection at low incidences of infection by testing plants individually is prohibitive in terms of the number of samples and associated cost of the required serological tests (our experience has been that the average cost of assaying one sample in duplicate by serology is US\$1.30), for any reasonable specified precision ( $C = 0.1, 0.2$ ). Given the rather large sample sizes required at low infection incidences, it may be more practical and economical to instead use group testing (16,20). The incidence of plant infection ( $p_{low}$ ) is then estimated from incidence at the group level ( $p_{high}$ ). The relationship between  $p_{low}$  and  $p_{high}$  derived from the binomial distribution is a well-known function in group testing (equation 6 with  $v = n$ ), but applies only if infected plants are spatially random when sampling is done within a spatial hierarchy (e.g., snap bean plants within quadrats). However, as we have shown in this paper, AIMV- and CMV-infected snap bean plants are spatially aggregated, and the level of aggregation changes with mean incidence systematically. Therefore, in order to estimate  $p_{low}$  from  $p_{high}$ , one must use an expression relating  $p_{high}$  to  $p_{low}$  that accounts for the underlying spatial aggregation. Aggregation at the individual plant level can be described by the beta-binomial distribution (8), but this leads to a mathematically intractable functional relationship between  $p_{low}$  and  $p_{high}$  (6). The problem spurred research into finding approximate empirical functions for describing the  $p_{low}$ - $p_{high}$  relationship (9). We used this approach to derive an empirical relationship between  $p_{low}$  and  $p_{high}$  for virus-infected snap bean. Note that the spatial aggregation is nested (via the parameter  $v = v_{opt}$ ) within the statistical relationship between  $p_{high}$  and  $p_{low}$ , in that group testing has to be performed on groups comprising all  $n$  samples collected from within one quadrat. That is, the empirical  $p_{low}$ - $p_{high}$  relationship derived here is invalid if samples are randomly grouped for testing after collection. Sampling then necessarily involves more extensive sample tracking in the field and during processing in the lab, but in our experience does not unreasonably increase time and effort invested.

Our preference was for using confidence intervals (rather than standard errors) to convey information on the precision of disease incidence estimates. Confidence intervals not only carry information on precision, but simultaneously represent magnitude and shifts in coverage as well. For example, explicitly accounting for aggregation modifies the confidence interval endpoints for  $p_{low}$  when estimated from  $p_{high}$ , as seen in Figure 3. Confidence intervals for proportions (such as disease incidence) can be poor in their coverage probability (that is, not meet the nominal  $1-\alpha$  level) due to the inherent discreteness and skewness in the underlying probability distributions (binomial, beta-binomial, or other) (1,15). It was recently shown that the common Wald interval provides

erratic coverage even for large sample sizes and for  $p_{low}$  not close to 0 or 1 (1). In this study, we used confidence interval estimators that have already been investigated in terms of their properties (1,21). These intervals are approximate, but are an acceptable alternative to the more computationally intensive calculation of exact intervals. Additionally, their closed forms make for easier graphically based investigations. Though we justified approximating  $se_d(\hat{p}_{low})$  with  $se_r(\hat{p}_{low})$  in the calculation of confidence intervals for  $\hat{p}_{low}$ , this will not be reasonable in all situations, and there is the need for further work on appropriate confidence intervals for  $\hat{p}_{low}$  when disease is aggregated (13).

Although one can use the  $p_{low}$ - $p_{high}$  relationship to predict  $p_{low}$  from  $p_{high}$ , there is a trade-off in terms of wider confidence intervals for  $p_{low}$  when estimated from  $p_{high}$ , especially as  $p_{high}$  increases. The confidence interval width can increase to the point where it may be deemed unacceptable, depending on the project's goals. Confidence interval widths can be decreased by increasing the sampling effort, but one can question whether the increased costs of further sampling are offset by appreciable gains in confidence interval width reduction. We showed that increasing the number of quadrats sampled and tested (at the group level) can be a very effective means of reducing the confidence interval width for  $p_{low}$ , more so at higher incidences of plant infection. However, the gains in confidence interval width reduction are probably not worth the increased sampling costs after about  $N = 40$  in this particular study, because the rate of reduction may not keep pace with the increased cost of sampling.

Another approach is to determine the number of groups that ought to be tested so that the respective confidence interval widths derived from group and individual testing are the same. The question is whether the number of serological tests can be reduced by using group testing, without sacrificing precision in the estimates of incidence. For example, we showed that for a sampling plan consisting of  $n \cdot N = 200$  plants tested individually, the same precision can be obtained with group testing with  $N_G = 60$ , up to about  $p_{low} = 0.2$ . Above that, the confidence interval width for  $p_{low}$  obtained via group testing diverged markedly from the width obtained from the testing of individual plants. For the sampling scheme under consideration at this point ( $N = 60, n = 5$ ), our results suggest using group testing until  $p_{high}$  is in the range [0.35, 0.59], for which the corresponding range in  $p_{low}$  is [0.1, 0.2].  $\hat{p}_{low}$  is a biased estimator of  $p_{low}$  when derived from the  $p_{low}$ - $p_{high}$  relationship (20). We did not explicitly address the bias of  $\hat{p}_{low}$  in this paper, but it is relatively small in the present context. For the sampling plan consisting of  $N_G = 60, n = 5$ , we calculated the bias in  $\hat{p}_{low}$  to be 0.0008 when  $p_{low} = 0.1$ , and 0.002 when  $p_{low} = 0.2$ . If the incidence of virus-infected plants is expected to be higher subsequently, then testing at the individual plant level is more precise, from a confidence interval point of view.

The analysis we present in this paper shows that there are upper practical limits for sample sizes in assessing the incidence of virus-infected plants in snap bean fields, and that the return on additional sampling (increased precision) may not be justified by the higher labor and material costs involved. We hope that the approach presented here will be useful to other groups studying aphid-transmitted viruses, in snap bean or in other vegetable crops. We are continuing our studies on the aphid-snap bean-virus complex by examining the temporal progression of virus incidence, and for this a combination of group testing (at the early stages of an epidemic) followed by individual testing at the later stages of the epidemic will be used, as suggested by the analyses presented in this paper.

## ACKNOWLEDGMENTS

This research was supported by grants from the New York Vegetable Research Council/Association to B. A. Nault.

## LITERATURE CITED

1. Brown, L. D., Cai, T. T., and DasGupta, A. 2001. Interval estimation for a binomial proportion. *Stat. Sci.* 16:101-133.
2. Dallot, S., Gottwald, T., Labonne, G., and Quiot, J.-B. 2003. Spatial pattern analysis of Sharka disease (*Plum pox virus* strain M) in peach orchards of southern France. *Phytopathology* 93:1543-1552.
3. Evans, N., Baierl, A., Brain, P., Welham, S. J., and Fitt, B. D. L. 2003. Spatial aspects of light leaf spot (*Pyrenopeziza brassicae*) epidemic development on winter oilseed rape (*Brassica napus*) in the United Kingdom. *Phytopathology* 93:657-665.
4. Hughes, G., and Gottwald, T. R. 1999. Survey methods for assessment of citrus tristeza virus incidence when *Toxoptera citricida* is the predominant vector. *Phytopathology* 89:487-494.
5. Hughes, G., and Madden, L. V. 1992. Aggregation and incidence of disease. *Plant Pathol.* 41:657-660.
6. Hughes, G., McRoberts, N., Madden, L. V., and Gottwald, T. R. 1997. Relationships between disease incidence at two levels in a spatial hierarchy. *Phytopathology* 87:542-550.
7. Larsen, R. C., Miklas, P. N., Eastwell, K. C., Grau, C. R., and Mondjana, A. 2002. A virus disease complex devastating late season snap bean production in the Midwest. *Annual Report—Bean Improvement Cooperative* 45:36-37.
8. Madden, L. V., and Hughes, G. 1995. Plant disease incidence: Distribution, heterogeneity, and temporal analysis. *Annu. Rev. Phytopathol.* 33:529-564.
9. Madden, L. V., and Hughes, G. 1999. An effective sample size for predicting plant disease incidence in a spatial hierarchy. *Phytopathology* 89:770-781.
10. Madden, L. V., and Hughes, G. 1999. Sampling for plant disease incidence. *Phytopathology* 89:1088-1103.
11. Madden, L. V., Pirone, T. P., and Raccach, B. 1987. Analysis of spatial patterns of virus-diseased tobacco plants. *Phytopathology* 77:1409-1417.
12. McRoberts, N., Hughes, G., and Madden, L. V. 2003. The theoretical basis and practical application of relationships between different disease intensity measurements in plants. *Ann. Appl. Biol.* 142:191-211.
13. Miao, W. W., and Gastwirth, J. L. 2004. The effect of dependence on confidence intervals for a population proportion. *Am. Stat.* 58:124-130.
14. Nault, B. A., Shah, D. A., Dillard, H. R., and McFaul, A. C. 2004. Seasonal and spatial dynamics of alate aphid dispersal in snap bean fields in proximity to alfalfa and implications for virus management. *Environ. Entomol.* 33:1593-1601.
15. Newcombe, R. G. 1998. Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Stat. Med.* 17:857-872.
16. Rodoni, B. C., Hepworth, G., Richardson, C., and Moran, J. R. 1994. The use of a sequential batch testing procedure and ELISA to determine the incidence of five viruses in Victorian cut-flower Sim carnations. *Aust. J. Agric. Res.* 45:223-230.
17. Roumagnac, P., Pruvost, O., Chiroleu, F., and Hughes, G. 2004. Spatial and temporal analyses of bacterial blight of onion caused by *Xanthomonas axonopodis* pv. *allii*. *Phytopathology* 94:138-146.
18. Shah, D. A., Bergstrom, G. C., and Ueng, P. P. 2001. Foci of Stagonospora nodorum blotch in winter wheat before canopy development. *Phytopathology* 91:642-647.
19. Shah, D. A., Dillard, H. R., Mazumdar-Leighton, S., Gonsalves, D., and Nault, B. A. Incidence, spatial patterns and associations among viruses in snap bean and alfalfa in New York. *Plant Dis.* (In press.)
20. Swallow, W. H. 1985. Group testing for estimating infection rates and probabilities of disease transmission. *Phytopathology* 75:882-889.
21. Tebbs, J. M., and Bilder, C. R. 2004. Confidence interval procedures for the probability of disease transmission in multiple-vector-transfer designs. *J. Agric. Biol. Environ. Stat.* 9:75-90.
22. Thackray, D. J., Jones, R. A. C., Bwye, A. M., and Coutts, B. A. 2000. Further studies on the effects of insecticides on aphid vector numbers and spread of cucumber mosaic virus in narrow-leaved lupins (*Lupinus angustifolius*). *Crop Prot.* 19:121-139.
23. Turechek, W. W., Ellis, M. A., and Madden, L. V. 2001. Sequential sampling for incidence of Phomopsis leaf blight of strawberry. *Phytopathology* 91:336-347.
24. Turechek, W. W., and Mahaffee, W. F. 2004. Spatial pattern analysis of hop powdery mildew in the Pacific Northwest: Implications for sampling. *Phytopathology* 94:1116-1128.
25. Wypij, D., and Santner, T. J. 1990. Interval estimation of the marginal probability of success for the beta-binomial distribution. *J. Stat. Comput. Simul.* 35:169-185.
26. Xu, X.-M., and Madden, L. V. 2002. Incidence and density relationships of powdery mildew on apple. *Phytopathology* 92:1005-1014.