TESTS FOR AGGREGATION AND SIZE-BASED SAMPLE-UNIT SELECTION
WHEN SAMPLE UNITS VARY IN SIZE

EDWARD F. CONNOR,1 ELIZABETH HOSFIELD,1 DUANE A. MEETER,2 AND XUFE NG NIU2

1Department of Environmental Sciences, University of Virginia, Charlottesville, Virginia 22903 USA and
Blandy Experimental Farm, P.O. Box 175, Boyce, Virginia 22620 USA
2Department of Statistics, Florida State University, Tallahassee, Florida 32306 USA

Abstract. Previous tests for aggregation of organisms among sampling units that vary in size and tests for size-based sample-unit selection fail to account for variation in the size of the sample unit. When sample units vary in size, tests that assume equal-sized sample units overestimate the degree of aggregation and have a tendency to find selection for large-sized sample units. Previous tests for selection of large or small leaves by phytophagous insects have been biased toward detecting selection for large leaves when no leaf selection or selection for small leaves may actually be present, and toward detecting aggregation even when populations are Poisson random or repulsed. We derive explicitly the population mean and variance of the number of organisms per sample unit when the size of the sample unit is random, and we outline procedures to estimate the degree of aggregation and test for size-based sample-unit selection using a generalized linear model based on a Poisson null hypothesis of independent random placement.

Key words: aggregation; gall formers; generalized linear models; leaf miners; leaf-size selection; Poisson random; sample-unit size.

INTRODUCTION

Using the Poisson distribution as a null model for the distribution of counts of organisms from a series of samples has become so common in ecology as to be taught in introductory college classes (Krebs 1989, Ricklefs 1990). It even has become so commonplace as to appear in popular novels such as Thomas Pynchon’s Gravity’s Rainbow (Pynchon 1973). The Poisson distribution used as a null model for the probability that the observed number of objects in a sample, \( Y \), is exactly \( k \), is usually represented as:

\[
P(Y = k) = \frac{e^{-\mu} \mu^k}{k!},
\]

where \( \mu \) is the mean number of objects per sample, \( k = 0, 1, 2, 3, \ldots \), and sample units are assumed to be mutually independent and of constant size \( A \) such that for sufficiently small \( A \), the probability of finding more than one object in a sample is small compared to \( A \) (Feller 1968, Poole 1974, Pielou 1977, Southwood 1978, Ludwig and Reynolds 1988). However, many biological applications of Eq. 1.1 violate the assumption of a constant size sample unit. We use the phrase “size of the sample unit” to refer to the spatial or temporal extent of each individual sample, not to the “sample size,” which connotes the number of sample units. For example, in studies of the distribution of folivorous insects among the leaves of plants, leaves vary considerably in size, yet Eq. 1.1 still has been applied. If the variance in \( A \) is nonnegligible, Eq. 1.1 is inappropriate, and conclusions based on its application may be erroneous. A more appropriate null model is that organisms occur independently in samples at a rate, \( \lambda \), per unit sample size (area, volume, etc.) so that the expected number of organisms in a sample unit of size \( A \) is:

\[
E(Y) = \mu = \lambda A.
\]

where \( \lambda \) is the density of organisms (number per unit sample size), and \( A \) is the size of the sample unit (Feller 1968: section VI.6). Hence, if the shape of the sample unit is unimportant, the Poisson model for data in which the sample unit varies in size would be:

\[
P(Y = k) = \frac{e^{-\lambda A} (\lambda A)^k}{k!},
\]

with the remaining assumptions as in Eq. 1.1.

The null hypothesis tested when applying the Poisson distribution to count data is that organisms are distributed independently and at random among the observed set of sample units. Rejection of this null hypothesis using a \( \chi^2 \) test to compare the observed and expected frequencies of sample units with \( k = 0, 1, 2, 3, \ldots \) organisms is taken as an indication that the observed distribution is either aggregated or repulsed. Since the variance of the Poisson distribution equals its mean, the ratio of the variance of counts of organisms among sample units to the average number in a sample unit is computed as an index of dispersion, values of the variance/mean ratio >1 indicating aggregation and values <1 indicating repulsion. We show
below that when sample units vary in size (area or volume), under an hypothesis of independent random placement the variance/mean ratio of counts of organisms is $>1$, so application of Eq. 1.1 under these circumstances will tend to lead to conclusions that organisms are aggregated when in fact they are not. The Poisson distribution could also be used to test the hypothesis that the number of organisms in a sample unit is independent of the size of the sample unit. Rejection of such a hypothesis has been interpreted as implying that organisms are aggregated in small or large sample units. For example, tests for "leaf-size selection" in leaf-consuming insects, which are essentially tests for a particular form of aggregation, are commonly performed (Whitham 1978, 1980, Tuomi et al. 1981, Mopper et al. 1984, Bultman and Faeth 1986, Faeth 1985, 1991a, b, Simberloff and Stiling 1987, Auerbach and Simberloff 1989, Clancy et al. 1993, Horne 1995). However, these tests for leaf-size selection are biased toward detecting selection for large leaves because they do not account for the fact that under Poisson random placement large leaves will contain more insects.

To use the Poisson model to test hypotheses concerning size selectivity for leaves or other sample units a reasonable alternative model for the expected number of organisms in a sample unit is:

$$E(Y) = \mu = \lambda A^a.$$  (1.4)

This model reduces to Eq. 1.2 when the exponent $b_1 = 1$, but $b_1 > 1$ implies that on average organisms are found in larger sample units at a higher rate than expected, assuming constant density, suggesting selection of larger sample units. Conversely, concluding that $b_1 < 1$ implies that organisms occur at a higher rate than expected in small sample units. Thus, a Poisson model for data in which the sample unit varies in size and in which size selectivity may be present is:

$$P(Y = k) = \frac{e^{-\lambda A} (\lambda A^a)^k}{k!},$$  (1.5)

with the same assumptions as for Eq. 1.3.

We derive explicitly the expected population mean and variance of the number of organisms per sample unit when the size of the sample unit is a random variable, and show that under the Poisson null hypothesis of independent random placement the variance/mean ratio always exceeds one. We also outline an approach based on the generalized linear model (McCullagh and Nelder 1989) to fit Eqs. 1.3 and 1.5 to test for aggregation and size-based sample-unit selection. We use data on the distribution of the leaf-mining moth *Cameraria hamadryadella* among leaves of its host trees, *Quercus alba* and *Quercus macrocarpa*, to illustrate these procedures. We also provide general guidelines concerning when Eq. 1.1, the traditional Poisson model, may still be appropriate even when the size of the sampling unit is variable.

**Methods**

**Field methods**

As part of a long-term study of the population dynamics of the leaf-mining moth *Cameraria hamadryadella* (Clemens) (Lepidoptera: Gracillariidae), samples of leaves were removed from the lower crown of several individual trees of either *Quercus alba* L. or *Q. macrocarpa* Michaux from the Orland E. White Arboretum and in a natural woodland located at Blandy Experimental Farm, Boyce, Virginia, USA, in 1984, 1989, 1991, and 1994. At least three twigs, each consisting of 25–50 leaves, were removed from at least three trees at each site. Each leaf was visually inspected and the number of leaf mines of *C. hamadryadella* on individual leaves was recorded. The area of each leaf was determined using a LICOR LI-3000 leaf area meter. These samples were collected with the primary goal of estimating the density of *C. hamadryadella* on trees, and the sample tree was treated as the independent sampling unit.

For the purpose of illustrating the models in Eqs. 1.3 and 1.5 we treat leaves as the independent sampling unit and pool leaves from different twigs and trees within tree species to test for leaf-size selection and aggregation for each host tree species in each year and at each site. We also fit the models in Eqs. 1.3 and 1.5 with host tree species and year as covariates (model in Eq. 2.13) to illustrate the ability of the generalized linear model to incorporate more complicated and realistic sampling designs.

In addition, because *C. hamadryadella* is an outbreak species, which over the course of this study varied in mean density between 0.012 and 34.04 individuals/leaf, these data allow the examination of the effect of mean density on our tests for aggregation and leaf-size selection (Connor and Beck 1993, Connor et al. 1994, Auerbach et al. 1995).

**Statistical Methods**

**The variance/mean ratio and a test for aggregation**

For sample units of fixed size $A$ (where $A$ may be in units of area, volume, time, or distance), $Y$, the number of objects in the sample, has a conditional Poisson distribution with mean and variance $\lambda A$:

$$E(Y | A) = \lambda A \quad V(Y | A) = \lambda A.$$  (2.1)

If $A$ is a random variable with mean $= \mu_A$ and variance $= \sigma^2_A$, then the unconditional mean of $Y$ is $E(Y) = E(\lambda A) = \lambda \mu_A$. For calculation of the unconditional variance of $Y$, we use the conditional variance formula (Bain and Engelhardt 1992):

$$V(Y) = E[V(Y | A)] + V[E(Y | A)],$$  (2.2)

which upon substituting Eq. 2.1 yields

$$V(Y) = E(\lambda A) + V(\lambda A) = \lambda \mu_A + \lambda^2 \sigma^2_A.$$  (2.3)

The ratio of the variance of $Y$ to the mean of $Y$ is:
\[ V(Y)/E(Y) = (\lambda \mu_A + \lambda^2 \sigma_A^2)/\lambda^2 \mu_A = 1 + \lambda (\sigma_A^2/\mu_A), \quad (2.4a) \]

This indicates that when sample units vary in size, the Poisson-based variance/mean ratio of the number of organisms counted in a sampling unit is increased by the product of the density (e.g., the number per unit area) and the variance/mean ratio of the sizes of the sampling units. Another way to express Eq. 2.4a is:

\[ V(Y)/E(Y) = 1 + \lambda \mu_A(\text{CV}(A))^2, \quad (2.4b) \]

where CV(A) is the coefficient of variation of the sizes of the sample units. Now, \( \lambda \mu_A \), the average number of organisms per sample unit is estimated as the total number of organisms summed among all sample units divided by the summed area of all sample units times the average size of a sample unit. For example, if 185 organisms were collected on 33 units whose mean sizes had a CV of 20\%, the variance/mean ratio under Poisson random occurrences would be estimated not as unity, but as \( 1 + (185/33)(0.2)^2 = 1.22 \).

The traditional \( \chi^2 \) statistic applied to test for aggregation when sample units are of constant size is \( (n - 1)s^2/\bar{y} \), which approximately follows a \( \chi^2 \) distribution with \( n - 1 \) degrees of freedom (where \( \bar{y} \) and \( s^2 \) are the sample estimates of the mean and variance of organisms per sample unit). However, when sample units vary in size, this ratio is biased upwards by a factor of \( 1 + \lambda (\sigma_A^2/\mu_A) \). We have not found it easy to adjust the test statistic \( (n - 1)s^2/\bar{y} \) in the obvious way by dividing it by an estimate of \( 1 + \lambda (\sigma_A^2/\mu_A) \), since it appears that the resulting statistic must have its degrees of freedom adjusted in a way that depends on \( \lambda \), \( \sigma_A^2 \), and \( \mu_A \). A simpler approach is to use the Pearson \( \chi^2 \) statistic

\[ \chi^2 = \sum_{i=1}^{n} (y_i - \bar{\mu}_i)^2/\bar{\mu}_i, \quad (2.5) \]

where \( \bar{\mu}_i \) is an estimate of the mean of \( y_i \) derived from a model such as in Eqs. 1.3 and 1.5, which takes into account its relationship to the size \( A_i \) of the sampling unit. We use Eq. 2.5 to test for Poisson randomness within a generalized linear model. If Eq. 2.5 is divided by its degrees of freedom it estimates the variance/mean ratio, denoted by \( \phi \) in the next section, without being inflated by variation in the sizes of the sampling units. We also provide a nomograph so that one can determine, for a specified population density and variance/mean ratio in the size of the sample units, the degree to which the observed variance/mean ratio in the number of organisms per sample unit is inflated by variation in the size of the sample units.

**Estimating and testing for aggregation and size-based sample-unit selection using generalized linear models**

To estimate the variance/mean ratio of the number of leaf mines per leaf, and to test the hypothesis that *C. hamadryadella*’s pattern of oviposition is independent of leaf size, we fit generalized linear models based on the models in Eqs. 1.3 and 1.5 to the data for each of the sampled populations of *C. hamadryadella* on each host plant species.

Given a sample of \( n \) pairs \((y_i, A_i)\) of counts and areas, it is possible to obtain maximum likelihood estimates of \( \lambda \) and \( \beta \), by programming an iterative maximization of the likelihood function. The hypothesis that \( \beta = 1 \) then can be tested using a likelihood ratio test. However, it is possible to imbed the model in Eq. 1.5 within the class of generalized linear models (Nelder and Wedderburn 1972), which offers both more alternatives to Eq. 1.5 and existing computer software (Baker and Nelder 1978, McCullagh and Nelder 1989)

The common linear regression model (a form of the general linear model) specifies that the mean response \( \mu \) is identical to a linear function \( \eta \) of the predictor variables \( x_i \)

\[ E(Y) = \mu = \eta = \beta_0 + \sum_{j=1}^{p} \beta_j x_j, \quad (2.6) \]

and uses least squares as the criterion by which to estimate the unknown parameters \( \beta = (\beta_0, \beta_1, \ldots, \beta_p) \). When observations are independent and normally distributed with constant variance \( \sigma^2 \), least squares estimation of \( \beta \) and \( \sigma^2 \) is equivalent to maximum likelihood estimation.

Generalized linear models encompass the general linear model and enlarge the class of linear least squares models in two ways: the distribution of \( Y \) for fixed \( x \) is merely assumed to be from the exponential family of distributions, which includes important distributions such as the binomial, Poisson, exponential, and gamma distributions, in addition to the normal distribution. Also, the relationship between \( E(Y) = \mu \) and \( \eta \) is specified by a nonlinear link function \( \eta = g(\mu) \), which is only required to be monotonic and differentiable.

The canonical link function given by McCullagh and Nelder (1989) for the Poisson distribution is \( g(\mu) = \log(\mu) \). For our application, since we want \( E(Y) = \lambda A^\phi \), it is convenient that

\[ \log(\mu) = \beta_0 + \beta_1 \log(A) \quad (2.7) \]

is a linear model where \( \beta_0 = \log(\lambda) \). The dispersion parameter in the generalized linear model, denoted by \( \phi \), is equivalent to the variance/mean ratio in the Poisson version of the exponential family of distributions, so its estimate should be close to unity if the data are Poisson random (McCullagh and Nelder 1989).

**Estimation and testing.**—The parameters in a generalized linear model can be estimated by the maximum likelihood method. For a given probability distribution specified by \( f(y; \beta, \phi) \) and observations \( y = (y_1, y_2, \ldots, y_n)' \), the log-likelihood function for \( \beta \) and \( \phi \), expressed as a function of mean values \( \mu = (\mu_1, \ldots, \mu_n) \) of the responses \( \{Y_1, Y_2, \ldots, Y_n\} \), has the form...
\[
l(\mu; y) = \sum_{i=1}^{n} \log f(y_i; \beta, \phi).
\] (2.8)

The maximum likelihood estimates of the parameters \(\beta\) can be obtained by iterative reweighted least squares, which is the basis of the computer software GLIM (Baker and Nelder 1978). Detailed information about the iterative algorithm and asymptotic properties of the parameter estimates can be found in McCullagh and Nelder (1989). We fitted generalized linear models to data using a similar algorithm in the computer software S-PLUS (Statistical Sciences 1993).

Analogous to the residual sum of squares in linear regression, the goodness-of-fit of a generalized linear model can be measured by the scaled deviance
\[
D(y; \hat{\mu}) = 2l(y; y) - l(\hat{\mu}; y),
\] (2.9)

where \(l(y; y)\) is the maximum likelihood achievable for an exact fit in which the fitted values are equal to the observed values, and \(l(\hat{\mu}; y)\) is the log-likelihood function calculated at the estimated parameters \(\hat{\beta}\). For the Poisson distribution family, the deviance is
\[
D(y; \hat{\mu}) = 2 \sum_{i=1}^{n} \{y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\}. \quad (2.10)
\]

Another important measure of goodness-of-fit is the Pearson \(X^2\) statistic, which has the form
\[
X^2 = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2/V(\hat{\mu}_i), \quad (2.11)
\]

where \(V(\hat{\mu}_i)\) is the estimated variance of the response \(Y_i\). If \(Y_i\) is normally distributed both the deviance and the Pearson \(X^2\) statistic have exact \(\chi^2\) distributions (McCullagh and Nelder 1989). If \(Y_i\) has a Poisson distribution with sufficiently large cell mean \(\mu_i\), both of the statistics are approximately \(\chi^2\) distributed with degrees of freedom \(n - p - 1\), where \(p\) is the number of parameters estimated (Agresti 1990).

The deviance function is very useful for comparing two models when one model has parameters that are a subset of the second model. Unlike the Pearson \(X^2\) statistic, the deviance is additive for such nested models if maximum likelihood estimates are used (McCullagh and Nelder 1989). Consider two nested models with the second having some covariates omitted and denote the maximum likelihood estimates in the two models by \(\hat{\mu}_1\) and \(\hat{\mu}_2\), respectively. Then the deviance difference \(D(y; \hat{\mu}_1) - D(y; \hat{\mu}_2)\) is identical to the likelihood ratio statistic and has an approximate \(\chi^2\) distribution with degrees of freedom equal to the difference between the numbers of parameters in the two models. For probability distributions in the exponential family the \(\chi^2\) approximation is usually quite accurate for differences of deviance even though it may be inaccurate for the deviances themselves (McCullagh and Nelder 1989).

\textbf{Overdispersion or aggregation}

If the sampling variance of a response variable \(Y_i\) is significantly greater than that predicted by an expected probability distribution, \(Y_i\) is said to be overdispersed or aggregated. For the number \(Y_i\) of organisms in the \(i^{th}\) sample unit, we may assume that \(V(Y_i) = \phi E(Y_i)\), where \(\phi > 1\) representing overdispersion or aggregation and \(\phi < 1\) representing underdispersion or repulsion relative to a Poisson distribution. Overdispersion in a Poisson distribution may arise in several ways (Pielou 1977).

The dispersion parameter from the fit of the generalized linear model based on Eq. 2.7, which assumes that \(\phi = 1\), provides an alternative estimate of the degree of aggregation adjusting for the fact that sample units vary in size, since in Eq. 2.7 the number per sample unit is explicitly modeled as a function of sample unit size. Wedderburn (1974) suggested estimating the dispersion parameter by \(\hat{\phi} = X^2/(n(p + 1))\) where \(X^2\) is the Pearson statistic. For Poisson distributions, \((n - p - 1)\hat{\phi}\) has an approximate \(\chi^2\) distribution with degrees of freedom \(n - p - 1\). Therefore, the Pearson statistic can be used to test hypotheses concerning the dispersion parameter.

The covariance matrix of \(\hat{\beta}\) is estimated by \(\text{COV} (\hat{\beta}) = \phi (X'WX)^{-1}\), where \(X\) is the covariate matrix and \(W\) is a weight matrix used in the iterative algorithm. If overdispersion occurs in a Poisson-based model, ignoring it (i.e., setting \(\phi = 1\)) will result in underestimating the standard errors of the parameter estimates, which may lead to incorrect conclusions. McCullagh and Nelder (1989) suggest modeling mean and dispersion jointly as a way to take possible overdispersion into account. For the leaf mine counts, \(Y_i\), we could assume that \(E(Y_i) = \mu_i\) and \(V(Y_i) = \phi_i \mu_i\), where the dispersion parameter \(\phi_i\) may vary over sample units. The relationship between mean \(\mu_i\) and leaf size is specified by Eq. 2.7, and similarly the dispersion parameter may also be modeled as a function of some covariates. In practice, the squared Pearson residuals \(d_i = (Y_i - \hat{\mu}_i)^2/\hat{\mu}_i\) may be used as measures of dispersion of \(Y_i\).

The joint model can be built by an algorithm, which alternates between fitting the mean model in Eq. 2.7 for given weights \(1/\phi\) and fitting the dispersion model by using the response variable \(d_i\). The detailed fitting procedure can be found in McCullagh and Nelder (1989).

\textbf{Testing for size-based sample-unit selection.}—A test for size-based sample-unit selection can be constructed by comparing the fit of two nested models with one model lacking the parameter \(\beta\). A preliminary fit of Eq. 2.7 is performed to estimate \(\phi\) and determine whether a model with \(\phi \neq 1\) should be fit. If significant overdispersion is detected, then a model with \(\phi \neq 1\) is fit as described above.

If the slope \(\beta_i\) in Eq. 2.7 is equal to one, the model becomes
log(μ) = β₀ + log(A); \quad (2.12)

the covariate log(A) in this model is called an offset in generalized linear models. The deviance reduction by using Eq. 2.7 when Eq. 2.12 is correct has an approximate χ² distribution, which can be used for testing the null hypothesis β₁ = 1 against the alternative hypothesis β₁ ≠ 1. Values of β₁ significantly greater (less) than one would indicate selection of large (small) sample units.

In order to examine the effects of other covariates on aggregation and size-based sample-unit selection, we examined the generalized linear model (Eq. 2.7) with additional covariates for the effects of year (Year) and host tree species (HTS) for the six populations combined. The covariate Year has three levels (1984, 1991, and 1994), and the covariate HTS has two levels (I = Q. alba, and II = Q. macrocarpa). The full generalized linear model has the form:

\log(μ) = β₀ + β₁ log(A) + αᵢ + δᵢ + (αA)ᵢ
+ (δA)ᵢ + (αδ)ᵢ + (αδA)ᵢ, \quad (2.13)

where αᵢ and δᵢ represent the main effects of the covariate Year and HTS, respectively, and (αA)ᵢ, (δA)ᵢ, (αδ)ᵢ, and (αδA)ᵢ are the two-factor and three-factor interactions.

Table 2. Observed variance/mean ratio of the number of C. hamadryadella mines/leaf, estimated variance/mean ratio (from Eq. 2.4a), and variance/mean ratio (φ) estimated by fitting the generalized linear model (Eq. 2.7) assuming φ = 1. Application of the χ² test based on φ indicates that all populations are significantly aggregated even after adjusting for variation in the size of the sample units. For all tests, p < 1 × 10⁻⁶.

| Population | Observed var./mean (Poisson) | Estimated var./mean (Poisson) | Dispersion (φ) | df | χ²
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.54</td>
<td>2.29</td>
<td>3.64</td>
<td>551</td>
<td>2005.64</td>
</tr>
<tr>
<td>2</td>
<td>3.39</td>
<td>1.78</td>
<td>2.56</td>
<td>148</td>
<td>378.88</td>
</tr>
<tr>
<td>3</td>
<td>5.07</td>
<td>3.05</td>
<td>2.40</td>
<td>337</td>
<td>808.80</td>
</tr>
<tr>
<td>4</td>
<td>2.51</td>
<td>1.68</td>
<td>1.74</td>
<td>498</td>
<td>866.52</td>
</tr>
<tr>
<td>5</td>
<td>11.21</td>
<td>5.12</td>
<td>1.20</td>
<td>683</td>
<td>12430.60</td>
</tr>
<tr>
<td>6</td>
<td>44.61</td>
<td>17.80</td>
<td>21.54</td>
<td>962</td>
<td>20442.50</td>
</tr>
</tbody>
</table>

It is well known that the effects in Eq. 2.13 are confounded and constraints must be imposed to estimate the parameters. Following McCullagh and Nelder (1989:64), we set the estimates of the first main effects of the covariates Year and HTS to be zero, i.e., αᵢ = δᵢ = 0. The constraints on the estimates of the two-factor interactions are (αA)ᵢ = 0, (δA)ᵢ = 0, and constraints on the three-factor interactions are imposed similarly:

(αδA)ᵢ₁ = (αδA)ᵢ₂ = (αδA)ᵢ₃ = (αδA)ᵢ₄ = 0.

We then fit the model in Eq. 2.13 with a dispersion parameter (e.g., φ ≠ 1) to the six populations of C. hamadryadella combined.

Results

Between 1984 and 1994, C. hamadryadella experienced drastic fluctuations in abundance in the arboretum at Blandy Farm, but not in the natural woodlot (Table 1). Mean densities in the arboretum ranged between 0.032 and 34.04 mines/leaf, while mean densities in the woodlot never exceeded 0.02 mines/leaf. The variance in leaf area within populations was between 12 and 58 times the mean leaf size. Quercus alba leaves were smaller and less variable in size than were Q. macrocarpa leaves. We excluded populations 2, 3, and 4 from further analysis because their densities were so low (a total of 13 insects) as to prevent the detection of aggregation or leaf size selection, even were these processes present.

Observed and Poisson-based variance/mean ratios

The observed variance/mean ratios for the number of C. hamadryadella mines/leaf are all >1, which on face value would suggest that C. hamadryadella mines are aggregated among leaves (Table 2). However, because the variance/mean ratio of leaf areas is high in each population, the Poisson-based variance/mean ratio estimated by substituting sample estimates of the parameters in Eq. 2.4a under an hypothesis of independent random placement is also substantially greater than one in each population.

Using Eq. 2.4a, we plotted the effects of various
ratio in sample-unit size lead to large increases in the variance/mean ratio in the number of individuals per sample unit relative to that expected when all sample units are equal in size. For Q. alba and Q. macrocarpa, which have variances in leaf areas that are at least 12-fold greater than the mean leaf area, Fig. 1 shows that only at population densities below 0.01 individuals/unit area would the variance/mean ratio approach one.

### Generalized linear models

As a preliminary analysis, we first assumed that the number of *C. hamadyadella*, $Y$, has a Poisson distribution specified in Eq. 1.5. The log-linear model (Eq. 2.7) was fitted to the data from the six populations, and the estimated slopes, $\beta_1$, along with their standard errors are listed in Table 3. The dispersion parameter for each population is estimated by $\phi = X^2/(n - p - 1)$ where $X^2$ is the Pearson statistic (Table 2). All six of the populations are overdispersed since the estimated dispersion parameters are considerably greater than 1. The estimate of $\phi$ obtained when the model in Eq. 2.13 is fit to the six populations combined is 11.6, which is similar to the average $\phi$ estimated from the six populations individually. It is notable that the estimated dispersion parameters are substantially smaller than the observed variance/mean ratios presented in column 2 of Table 2. In fact, for each population the estimated dispersion parameter, $\phi$, is a measure of the degree of aggregation in the distribution of *C. hamadyadella* among leaves adjusted for leaf size, since we assume that the mean number on a leaf is proportional to a power of leaf size.

Since the slopes ($\beta_1$) for the six populations are estimated assuming the dispersion ratio $\phi = 1$, the estimated standard errors presented in column 4 of Table 3 are misleading. For example, the estimated slope for population 1 is 1.14 with an estimated standard error of 0.043, which in fact leads to the wrong conclusion, i.e., that ovipositing *C. hamadyadella* prefer large leaves.

### Table 3

<table>
<thead>
<tr>
<th>Population</th>
<th>$df$</th>
<th>Model with $\phi = 1$</th>
<th>Model with $\phi \neq 1$</th>
<th>Likelihood ratio test statistic (deviance difference)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Slope ($\beta_1$)</td>
<td>SE</td>
<td>Deviance</td>
<td>Slope ($\beta_1$)</td>
</tr>
<tr>
<td>1</td>
<td>1.14</td>
<td>0.043</td>
<td>1788</td>
<td>1.14</td>
</tr>
<tr>
<td>5</td>
<td>0.88</td>
<td>0.098</td>
<td>361</td>
<td>0.88</td>
</tr>
<tr>
<td>6</td>
<td>0.97</td>
<td>0.041</td>
<td>798</td>
<td>0.97</td>
</tr>
<tr>
<td>7</td>
<td>0.83</td>
<td>0.058</td>
<td>886</td>
<td>0.83</td>
</tr>
<tr>
<td>8</td>
<td>1.01</td>
<td>0.013</td>
<td>11114</td>
<td>1.01</td>
</tr>
<tr>
<td>9</td>
<td>1.04</td>
<td>0.009</td>
<td>19728</td>
<td>1.04</td>
</tr>
</tbody>
</table>

* Indicates test significant at $\alpha = 0.05$ level.

Combinations of density ($\lambda$) and the variance/mean ratio of the sample unit sizes on the Poisson-based variance/mean ratio of the number of individuals per sample unit (Fig. 1). The isopleths in Fig. 1 show combinations of density and the variance/mean ratio in sample-unit sizes that increase the variance/mean ratio of number of individuals per sample unit (Eq. 2.4a) by a given percentage. Alternatively, using Eq. 2.4b the ordinate can be the (cv)$^2$ of the sample unit sizes, if the abscissa is interpreted as the average number of organisms per sample unit. High densities or a large variance/mean.
The deviances and their degrees of freedom for each population for the model with $\phi = 1$ are listed in columns 5 and 2 of Table 3, respectively. If the fitted model with $\phi = 1$ were adequate for each population, the deviance should have an approximate $\chi^2$ distribution. Moreover, if the degrees of freedom (df) is large (say, $>30$) then $(\text{deviance} - \text{df})/\sqrt{2 \times \text{df}}$ will be accurately approximated by the standard normal distribution. The deviance values for each population for the models with $\phi = 1$ are obviously highly significant (Table 3). For example, $(\text{deviance} - \text{df})/\sqrt{2 \times \text{df}}$ for population 1 is $(1788 - 552)/\sqrt{1104} = 37.2$, which indicates strongly that the fitted model for this population is inadequate.

Since all six populations show overdispersion, the mean and the dispersion should be modeled jointly as described above. As an illustration, we show the modeling procedure for population 1. At the initial step, Eq. 2.7 is fitted to the data without taking into account the overdispersion ($\phi = 1$). The squared Pearson residuals $d_i = (Y_i - \hat{\mu}_i)/\hat{\mu}_i$; $i = 1, \ldots, n$ are calculated from the fitted model as estimates of the dispersion parameters $\phi_i$. Then $d_i$ are examined to see whether the dispersion parameter estimates vary among leaves or not. Plots of $d_i$ against leaf area show that the dispersion estimates for different leaf sizes are fairly constant (Fig. 2c). Therefore, the $Y_i$'s are assumed to have a common dispersion parameter, i.e., $V(Y) = \phi_i E(Y)$.

The common dispersion parameter is estimated by $\hat{\phi} = X^2/(n - p - 1)$ and the model with $\phi \neq 1$ is fitted. The deviance for the model with $\phi \neq 1$ is 491 with df = 552. Since the value $(491 - 552)/\sqrt{2 \times 552} = -1.8$ shows no evidence that the deviance differs from the expected $\chi^2$ distribution, the fitted model appears to be adequate for population 1. The deviance residuals from the model for population 1 with $\phi \neq 1$ are shown in Fig. 2d. The "striped" pattern is caused by the fact that a residual is always the difference between a low integer and a function of the continuous variable $A$. The estimated slope for population 1 is unchanged in the model with $\phi \neq 1$, but its estimated standard error increases from 0.043 to 0.083, which indicates that the slope is not significantly different from 1. In other words, after taking into account the overdispersion, the

**Fig. 2.** Plots of (a) the number of individual *Cameraria hamadryadella* per leaf as a function of leaf area, (b) the log(number) of *C. hamadryadella* as a function of log(leaf area), (c) the squared Pearson residual from the model assuming $\phi = 1$ vs. area, and (d) the deviance residual from the model assuming $\phi 
eq 1$ vs. the fitted value.
distribution of *C. hamadryadella* among leaves does not show that ovipositing females display a preference for small or large leaves.

The same procedure was applied to the data from each of the populations and the dispersion estimates were found to be fairly constant. The estimated slopes and their standard errors from the model with \( \phi \neq 1 \) are presented in Table 3. In populations 1, 5, 6, 8, and 9, *C. hamadryadella* display no preference for large or small leaves, but it is interesting to note that in population 7, the distribution of *C. hamadryadella* is consistent with the hypothesis that they prefer small leaves.

The likelihood ratio test is used to compare Eq. 2.7 with Eq. 2.12 taking into account overdispersion. The differences of deviances from the two models with \( \phi \neq 1 \) for each of the six populations are listed in Table 3. Since the deviance difference has an approximate \( \chi^2 \) distribution that has a critical value of 3.84 for \( \alpha = 0.05 \), only the slope for population 7 is different from 1.

When the generalized linear model with covariates for Year and HTS was fitted to the six populations combined, the main effect of HTS, the two-factor interactions between leaf size and the other two covariates, and the three-factor interactions are statistically insignificant (\( \chi^2 \) test of change in deviance). The final fitted model includes the main effects of year and leaf size [log(A)], and the two-factor interactions between Year and HTS. The residual deviance of the fitted model is 2986.4 with 3196 degrees of freedom, and the \( \chi^2 \) test indicates that the fitted model is adequate.

The estimated coefficients in the final model are listed in Table 4. The estimated coefficients for years 1991 and 1994 are 0.810 and 1.237, respectively, which implies that relative to year 1984 (where the coefficient is set to zero), the number of *C. hamadryadella* per square centimeter of leaf area decreased in 1991 and increased in 1994. In fact, based on Table 1, the density of *C. hamadryadella* in 1984 is 0.07795 individuals/cm², while the average density for 1991 (among populations 5, 6, and 7) and for 1994 (among populations 8 and 9) are 0.04571 and 0.2950 individuals/cm², respectively. The estimated coefficient for log(A) is 1.029 with an estimated standard error of 0.024, which indicates that *C. hamadryadella* do not display a preference for small or large leaves. The estimated interaction between year 1991 and HTS II is 0.442, which shows that the density of *C. hamadryadella* in 1991 was greater on *Q. macrocarpa* than on *Q. alba*. Similarly, the density of *C. hamadryadella* on *Q. macrocarpa* in 1994 was less than on *Q. alba*, although the estimated coefficient of \(-0.056\) was not significantly different from zero. The results of the analyses based on the fitted model are consistent with the sample statistics presented in Table 1.

**Discussion**

**Aggregation**

The difference between the observed variance/mean ratios of the number of *C. hamadryadella* per leaf and the expected values calculated explicitly from Eq. 2.4a, and also from the dispersion parameter \( \phi \) estimated by fitting the generalized linear models (Eqs. 2.7 and 2.13) with \( \phi \) set equal to 1, both indicate that *C. hamadryadella* are aggregated among leaves (Table 2). Both of these methods adjust for variation in the size of the sample unit and provide roughly comparable estimates of the actual degree of aggregation among leaves by *C. hamadryadella*. The observed variance/mean ratio overestimates the degree of aggregation in *C. hamadryadella* by between 50 and 100% due to variation in the size of the sample unit. The greater the heterogeneity in the sizes of the sample units, the greater the tendency to overestimate the degree of aggregation in a population, or to conclude that a population is aggregated when it is not (Fig. 1).

The tendency for the observed variance/mean ratios to overestimate the actual degree of aggregation is partly due to the interaction of the average density of organisms per unit sample area and the variance in sample-unit size. At high densities (>0.1 organisms/unit area), even small variances in sample-unit size result in estimated variance/mean ratios that are significantly greater than the Poisson expectation of one. When densities are low (<0.01 organisms/unit area), even high variances in sample-unit size have little influence on the estimated variance/mean ratio. Therefore, the model in Eq. 1.1 remains appropriate when densities (per unit sample area) are very low, or when the variance in sample-unit size approaches zero.

Reports that the distribution of insects among sample units is aggregated are common in the literature (Kobayashi 1968, Dixon and McKay 1970, Sugimoto 1976, Atkinson and Shorrocks 1984, Xu 1985, Thistlewood 1989, Auerbach and Simberloff 1989, Faeth 1991a, b, Horne 1995, and many others [see Taylor 1984 for a review]). These conclusions have usually been reached by fitting the model in Eq. 1.1 to sample data or using a number of other plot-based techniques to calculate
some index of dispersion (Taylor 1961, Green 1966, Lloyd 1967, Iwao 1968, 1972; see Pielou 1977, Southwood 1978, and Ludwig and Reynolds 1988 for reviews of these techniques). The sample variance in abundance among sampling units is a key component of all the common tests and indices of aggregation. All of these techniques are appropriate when applied to data that uses fixed-size sampling units, such as quadrats, because the variance among sampling units is due entirely to processes leading to departures from a Poisson random distribution and error. However, when the sampling unit varies in size as in "natural sampling units" such as leaves, rock pools, or some other naturally occurring unit of the environment (Pielou 1977, Ludwig and Reynolds 1988), variation in abundance caused by processes leading to aggregation or repulsion is confounded with variation in abundance caused by variation in the size of the sample unit. Estimates of the degree of aggregation in biological populations, regardless of the test or index used, that are based on natural sampling units without adjustment for sample unit size generally overestimate the degree of aggregation, and underestimate the degree of repulsion. The magnitude of misestimation will depend on the variance in the sizes of the sampling unit, the density of the population, and possibly on the specific test or index used. The extent of this problem is difficult to predict without more information on the variances in the sizes of natural sampling units. However, for insects on the leaves of plants, our data on the variances in leaf sizes of Quercus alba and Q. macrocarpa suggest that overestimates of the degree of aggregation may be large and common. Faeth (1991a, b) also reports that variances in size among attacked or unattacked leaves of Quercus emoryi Torrey were between 4 and 10 times greater than the mean leaf size, which indicates that the overall variance in leaf size is considerably greater.

In his study of the distribution of a leaf-mining fly, Phytomyza ranunculi among the leaves of Ranunculus asiaticus, Sugimoto (1976) recognized that variation in leaf size led to overestimation of the degree of aggregation of P. ranunculi among leaves. He used a graphical method to estimate the degree of aggregation caused by variation in leaf size with an index derived from Iwao's mean crowding. Sugimoto (1976) concluded that P. ranunculi was aggregated even after accounting for variation in leaf sizes. Sugimoto's (1976) pioneering delineation of this problem and his solution have been ignored.

Size-based sample-unit selection

Tests of the hypothesis that the slope $\beta_1 = 1$, based on fitting the generalized linear model (Eq. 2.7) with $\phi \neq 1$, indicate that five of the six populations of C. hamadryadella show no evidence for leaf-size selection and the distribution of one population (number 7) is consistent with the hypothesis that females preferentially deposit eggs on small leaves. This contrasts with the results one would obtain by fitting the model in Eq. 2.7 with $\phi = 1$, and therefore not accounting for overdispersion. Similarly, a test of the hypothesis that $\beta_1 = 1$, based on fitting the generalized linear model (Eq. 2.13) with $\phi \neq 1$ to the six populations combined, also shows no evidence for leaf-size selection. Given that the distribution of insects on leaves is usually reported to be overdispersed relative to the Poisson distribution, we advise fitting generalized linear models with $\phi \neq 1$, such as the models in Eqs. 2.7 and 2.13 to test for leaf-size selection.

Among leaf-mining and leaf-galling insects the selection of leaves for juvenile development is accomplished at oviposition by the female parent. Most species of leaf miners and leaf galls are sedentary as larvae, obtaining all of their nutrition from a single leaf. As a result, there should be strong selection for females to recognize leaves that will provide adequate nutrition for the development of their offspring. Most studies of leaf selection in folivorous insects suggest that aspects of leaf chemistry, physical attributes of leaves such as toughness, trichome density, etc., position on a plant, or the presence or absence of eggs of conspecifics are the critical cues used in leaf selection (Whitham 1978, 1980, Zucker 1982, Nef 1988, McNeil and Quiring 1983, Faeth 1991a, b, Horne 1995). However, some studies suggest that leaf size may be a trait important in leaf selection by ovipositing females (Whitham 1978, 1980, Faeth 1985, 1991a, b, Bultman and Faeth 1986), and others suggest that apparent leaf-size selection is an artifact of combining data from foliage of different phenological (and therefore size) classes (Rhomberg 1984, Borgacheva 1994). Leaf size could affect juvenile growth and survival directly if the resources available from small leaves are inadequate to support complete development. Alternatively, the impact of competitors, natural enemies, or chemical or physical attributes of leaves could be related to leaf size, leading ovipositing females to select leaves to avoid any detrimental effects (Bultman and Faeth 1986).

Several studies have reported evidence of leaf-size selection among leaf-mining and leaf-galling insects (Sugimoto 1976, Whitham 1978, 1980, Hileman and Lieto 1981, Tuomi et al. 1981, Mopper et al. 1984, Faeth 1985, 1991a, b, Bultman and Faeth 1986, Simberloff and Stiling 1987, Auerbach and Simberloff 1989, Clancy et al. 1993). The evidence used to infer leaf-size selection has been based either on correlations between the abundance of insects and leaf size, or on differences in the average sizes of leaves with leaf-mining or leaf-galling insects relative to leaves with no leaf miners or leaf galler. However, in every instance the statistical test performed has been inappropriate because the tests do not account for the fact that, even were eggs placed independently and at random, large leaves would be more likely to be attacked than would small leaves.
For example, in his examination of "habitat selection" in *Pemphigus* aphids on cottonwood, Whitham (1978, 1980) performed a test to determine whether stem mothers choose the size of leaves they colonize in proportion to their availability on trees, or whether they select small or large leaves disproportionately. From a random sample of leaves, Whitham (1978, 1980) calculated the average and standard deviation of leaf size for all leaves and for leaves with galls, and reported based on a $t$ test that the average size of leaves with galls was significantly greater than for leaves in general. Based on this result he concluded that stem mothers selectively colonize large leaves. Bultman and Faeth (1986), Faeth (1985, 1991a, b), and Simberloff and Stiling (1987) performed similar tests with leaf-mining insects, but compared the average size of leaves with mines to the average size of leaves without mines. The null hypothesis implicit in each of these tests is that all leaves, regardless of size, are equally likely to be colonized. For example, a leaf of 1 cm$^2$ would have the same chance of being colonized as a leaf of 10 cm$^2$.

Sugimoto (1976), Tuomi et al. (1981), Mopper et al. (1984), and Auerbach and Simberloff (1989) attempted to test for leaf-size selection in leaf-mining insects by examining the ordinary least-squares regression of number of mines on a leaf with leaf size. They concluded that since the number of mines on a leaf was positively correlated with leaf size that females prefer to oviposit on large leaves.

The null hypotheses examined in the tests for leaf-size selection outlined above do not weight the probability that an individual leaf will be selected for leaf size. However, if each unit area of every leaf has a constant probability of being attacked as proposed in Eq. 1.5, large leaves will be attacked more often than will small leaves. Given the Poisson null hypothesis (Eq. 1.5), the expected number of eggs, leaf mines, leaf galls, stem mothers, or animals per leaf is proportional to leaf area. This hypothesis is analogous to the random placement hypothesis developed by Coleman (1981) for species-area relationships.

There are at least three ways to test for leaf-size selection under Eq. 1.5: (1) performing tests comparing the mean sizes of attacked and unattacked leaves, but using randomization or bootstrapping to generate the null distribution of the test statistic under Eq. 1.5, (2) using an ordinary least-squares regression of the relationship between the density or number of individuals per leaf and leaf area, and (3) using a test that $\beta_1$ estimated from fits of generalized linear models with $\phi \neq 1$ such as Eqs. 2.7 or 2.13 to data on counts and leaf areas, equals one.

**Tests on mean leaf size.**—The $t$ tests or Smirnov tests previously performed by Whitham (1978, 1980), Faeth (1985, 1991a, b), Bultman and Faeth (1986), Simberloff and Stiling (1987), and Clancy et al. (1993) estimate the tail probability of the observed test statistic using the wrong null distribution. If animals are more likely to encounter larger targets (leaves), then large leaves should be encountered more than small leaves even if no leaf-size selection is occurring. Therefore, the expected value of a $t$ or Smirnov statistic will be positively biased, and tabulated values of the probability of $t$ or the Smirnov statistic will be incorrect. A computer-intensive randomization or bootstrap procedure would be necessary to generate the appropriate null distribution of $t$ or the Smirnov statistic to test for leaf-size selection. One could perform a computer simulation repeatedly selecting $I_n = n - I_0$ leaves to constitute the partition of the sample of $n$ leaves that are attacked, and $I_0 = n - I_n$ leaves that are not attacked, and calculate a $t$ test or Smirnov test to compare the mean sizes of attacked and unattacked leaves. On each random draw the probability that leaf $i$ is selected would be $A_i / \Sigma_{j=1}^{n} A_j$, where $A_i$ is the area of the individual leaf and $\Sigma_{j=1}^{n} A_j$ is the total leaf area of the sample. From these repeated simulations the null distribution of the test statistic could be elaborated, and the test statistic computed from the actual data compared to the percentiles of the simulated null distribution to determine whether the observed mean leaf sizes differ sufficiently to reject the null hypothesis of no leaf-size selection.

**Regression tests based on density or abundance.**—A regression approach similar to that used by Sugimoto (1976), Tuomi et al. (1981), Mopper et al. (1984), and Auerbach and Simberloff (1989) based on either density of animals per unit leaf area, or on the logarithms of number per leaf and leaf area could also be used to test the constant density hypothesis. These tests are analogous to tests previously applied to determine whether large areas or islands contain more species per unit area than small islands (Abele and Connor 1979, Connor and McCoy 1979). Under the constant density hypothesis, the slope of the regression of the log(number of animals) on a leaf with the log(area) of the leaf should equal 1, and the slope of the regression of the density of animals on a leaf (number per unit area) with leaf area should equal 0. If the slope of the regression of the log(number of animals) on a leaf with the log(area) of the leaf is significantly greater or less than 1, the data indicate selection for large or small leaves, respectively. For regressions of animal densities on leaf area, slopes significantly greater or less than 0 indicate selection for large or small leaves, respectively. Comparison of the observed slope values with the expected slope value using a $t$ test could serve as a test for leaf-size selection (Draper and Smith 1981:26). However, ordinary least-squares regression assumes normally distributed errors, and more seriously, constant variance. For a Poisson random variable such as $Y$, the number of individuals per sample unit, the variance is not constant, but rather equals the mean. Therefore, tests based on linear least-squares regression will be inappropriate.

For regressions using abundance rather than density
data, when the sample contains many leaves with no animals the log(number + 1) transformation must be used since log(0) is undefined. However, the expected slope of the regression of log(number + 1) on log(area) under an hypothesis of no leaf size selection is less than 1 when there are many sample units with no individuals. The exact value of the slope would depend on the specific population being examined. The expected slope also cannot be derived analytically so it would have to be generated by a computer simulation similar to that described above. The previously mentioned violation of the assumption of constant variance in Y at each sample unit size also applies.

None of the studies mentioned above provide adequate data to permit us to test for leaf-size selection under an hypothesis of constant density per unit area. Therefore, we can neither assert that leaf-size selection is or is not a common phenomenon in leaf-mining or leaf-galling insects. However, future tests for leaf-size selection should be designed to account for the fact that even under a Poisson null hypothesis of independent random placement, large leaves are more likely to be attacked than are small leaves. We conclude that fitting Eq. 2.7, with or without additional covariates, as a generalized linear model with \( \phi \neq 1 \) is the best approach to testing for leaf-size selection.

**Synthesis**

Published estimates of the degree of aggregation in animal populations that are derived from sampling units that vary in size are overestimates. Adjusted estimates of the degree of aggregation and tests for departure from the Poisson null hypothesis of independent random placement can be derived from generalized linear models such as Eqs. 2.7 or 2.13 with \( \phi \neq 1 \). Regardless of the approach used, to estimate or test for aggregation, it is imperative that in studies where the size of the sample unit is a random variable, the size of the sample unit be recorded.

The generalized linear models based on Eq. 1.5 provide a unified basis to account for variation in the sizes of sampling units when estimating the degree of aggregation or repulsion in the distribution of organisms among natural sampling units, and when testing for size-based sample-unit selection. The ability to use additional covariates in these models also permits the researcher to account for the specific structure of the sampling design. Currently, it is common practice to pool and treat all sample units as independent regardless of how nested their selection may have been.

While we have illustrated models Eqs. 1.3 and 1.5 and generalized linear modeling techniques using the distribution of an insect among the leaves of its host tree, they may be readily applied to other taxa, others kinds of natural sampling units, and to other questions. For example, to examine the "sampling" hypothesis as an explanation for the species-area relationship for birds in woodlots, Haila et al. (1993) tested the hypothesis that the distribution of individual birds among woodlots was Poisson random. They calculated the Poisson expectation of the number of birds in each woodlot assuming woodlots were of constant size (Eq. 1.1), even though Haila et al. (1993) recognized that large woodlots should have more birds and that Eq. 1.1 was inappropriate. However, whenever a Poisson null hypothesis of independent random placement is examined in a series of sampling units that vary in size, Eq. 1.5 will be the appropriate null hypothesis.

**Acknowledgments**

We would like to thank M. Beck, E. Green, M. Nuckols, S. Vix, and S. Vargo for assistance with the field work. This manuscript benefited from the helpful suggestions of M. Beck, J. Dooley, S. Matter, C. Sacchi, and J. Yoder, and from reviews by L. McDonald and P. Quang. This research was supported by NSF Grants DEB80-21779 and BBS91-12013.

**Literature Cited**


