

## IMPROVING THE PRECISION OF ESTIMATES OF THE FREQUENCY OF RARE EVENTS

PHILIP M. DIXON,<sup>1,4</sup> AARON M. ELLISON,<sup>2</sup> AND NICHOLAS J. GOTELLI<sup>3</sup>

<sup>1</sup>Department of Statistics, Iowa State University, Ames, Iowa 50011 USA

<sup>2</sup>Harvard Forest, Harvard University, P.O. Box 68, Petersham, Massachusetts 01366 USA

<sup>3</sup>Department of Biology, University of Vermont, Burlington Vermont 05405 USA

**Abstract.** The probability of a rare event is usually estimated directly as the number of times the event occurs divided by the total sample size. Unfortunately, the precision of this estimate is low. For typical sample sizes of  $N < 100$  in ecological studies, the coefficient of variation (cv) of this estimate of the probability of a rare event can exceed 300%. Sample sizes on the order of  $10^3$ – $10^4$  observations are needed to reduce the cv to below 10%. If it is impractical or impossible to increase the sample size, auxiliary data can be used to improve the precision of the estimate. We describe four approaches for using auxiliary data to improve the precision of estimates of the probability of a rare event: (1) Bayesian analysis that includes prior information about the probability; (2) stratification that incorporates information on the heterogeneity in the population; (3) regression models that account for information correlated with the probability; and (4) inclusion of aggregated data collected at larger spatial or temporal scales. These approaches are illustrated using data on the probability of capture of vespulid wasps by the insectivorous plant *Darlingtonia californica*. All four methods increase the precision of the estimate relative to the simple frequency-based estimate (absolute precision = 1.26, relative precision [cv] = 70%): stratification (absolute precision = 1.10, cv = 62%); regression models (absolute precision = 1.59, cv = 55%); Bayesian analysis with an informative prior probability distribution (absolute precision = 4.28, cv = 47%); and using temporally aggregated data (absolute precision = 6.75, cv = 36%). When informative auxiliary data is available, we recommend including it when estimating the probability of rare events.

**Key words:** aggregation; Bayesian inference, coefficient of variation; estimators; precision; rare events; regression; sampling, stratification.

### INTRODUCTION

Rare events are important in ecology and evolution. Familiar examples include genetic drift in founding populations (Mayr 1963), seedling establishment in plant populations with low growth rates (Harper 1977), successful establishment of seedlings following long-distance dispersal (Clark et al. 2001), species extinction (Roberts and Solow 2003), and extreme meteorological events such as ice storms, wildfires, or hurricanes (Whelan 1995, Foster and Aber 2003). The ecology (Rabinowitz 1981) and biogeography (Jetz and Rahbek 2002) of rare species may be very different from that of common species, and the statistical distribution of rare species is a key prediction that distinguishes many models of community structure (Williams 1964, Hubbell 2001, Magurran 2003, Chave 2004).

Precisely estimating the probability of rare events is a statistical challenge. If the true probability of a discrete rare event is  $\pi$ , the standard frequentist estimate of this probability,  $\hat{p}$ , is calculated as the number of

observations  $n$  of the rare event divided by the total number of observations (or trials)  $N$  (Gotelli and Ellison 2004):  $\hat{p} = n/N$ . The standard error of this estimate  $SE_{\hat{p}} = \sqrt{\hat{p}(1 - \hat{p})/N}$ , and its coefficient of variation  $CV_{\hat{p}} = SE_{\hat{p}}/\hat{p}$ .

If an event is truly rare ( $\pi < 0.01$ ), its frequentist estimate  $\hat{p}$  has reasonable precision ( $CV_{\hat{p}} \leq 10\%$ ) only when the sample size  $N$  exceeds 1000 total observations or trials. For  $N < 100$  total observations, typical for many ecological studies,  $CV_{\hat{p}}$  may exceed 300%. When the precision is low, it can be difficult to detect trends in the frequency or differences between groups. Increasing the precision provides better estimates and higher power for statistical tests of trends and differences. The precision of an estimate can also be important for policy decisions (e.g., Lewison et al. 2004). In this article, we describe four methods that can provide more precise estimates of the probability of a rare event. All of these methods require auxiliary data, but obtaining this auxiliary data usually requires less effort or cost than obtaining larger samples of the rare event itself.

### EXAMPLE DATA

We use data on the capture efficiency of wasps by the insectivorous pitcher plant, *Darlingtonia californica*.

Manuscript received 31 March 2004; revised 16 June 2004; accepted 12 July 2004; final version received 20 September 2004. Corresponding Editor: A. A. Agrawal. For reprints of this Special Feature, see footnote 1, p. 1079.

<sup>4</sup> E-mail: pdixon@iastate.edu



PLATE 1. *Darlingtonia californica*, a rare carnivorous plant species endemic to the Siskiyou Mountains of Oregon and northern California, which grows in a threatened plant community type—serpentine fen. Photo credit: A. M. Ellison.

*nica* (Sarraceniaceae), to illustrate methods by which the precision of estimates of the probability of rare events can be increased (see Plate 1). Although prey capture by carnivorous plants provides nutrients required for successful sexual reproduction (reviewed in Ellison and Gotelli 2001), prey capture may be infrequent or rare (Zamora 1995, Zamora et al. 1998); most insects that enter pitcher-plant traps are not captured (Newell and Nastase 1998).

Like other pitcher plants in this family, *Darlingtonia* grows as a rosette of leaves that are modified to form pitcher-shaped traps (Arber 1941). These pitchers secrete copious nectar that attracts foraging insects, especially vespid wasps (*Vespula atropilosa*) and ants (*Tapinoma sessile*). As part of a long-term study of the demography of *Darlingtonia*, we recorded the frequency with which *Darlingtonia* captures wasps and estimated the conditional probability of a successful capture:  $\pi = P(\text{capture} \mid \text{visit})$ . During July 2002, Ellison, Gotelli and their colleagues observed 753 *Darlingtonia* plants for one-half hour each, for a total of 376.5 plant-hours of observation (A. M. Ellison, R. J. Emerson, E. J. Farnsworth, N. J. Gotelli, C. M. Hart, H. R. Steinhoff, and S. E. Wittman, *unpublished data*). During this time,  $N = 157$  wasps were seen to visit the pitchers, and  $n = 2$  of these wasps were captured. For each visit, we also recorded the time a wasp spent in each pitcher, and we measured the orientation of the pitcher's opening (as degrees east of north). Assuming that the observed visits are a simple random sample of visits, the frequentist

estimate of  $\pi$  is  $\hat{p} = n/N = 2 \text{ captures}/157 \text{ visits} = 0.0127$ . The estimated standard error for  $\hat{p}$  is  $SE_{\hat{p}} = \sqrt{\hat{p}(1 - \hat{p})/N} = 0.0089$ . These estimates do not assume that the per-visit probability of capture is the same for all visits. When the sample is a simple random sample, heterogeneity in the population is irrelevant (Thompson 2002). Because the probability of capturing a wasp is very low and the total sample size is small, the precision of this frequentist estimate ( $\hat{p}$ ) of capture probability by *Darlingtonia* is poor:  $CV_{\hat{p}} = 70.2\%$  and precision (defined below) = 1.26.

#### MEASURES OF PRECISION OF A PROBABILITY

Precision “refers to the dispersion of the observations” (Marriott 1990). It can be quantified by at least four different, but related, measures (Table 1). The most familiar are measures of absolute precision, the standard error (SE) and variance ( $s^2$ ). Because more precise estimates have smaller SEs, it is also common to define precision as  $1/s^2$ , especially in the Bayesian literature (Gelman et al. 1995:43).

When used to compare events that have different probabilities, absolute measures of precision have the counter-intuitive property that rarer events are more precisely measured. To illustrate this, consider an event (such as a visit of a wasp to a *Darlingtonia* pitcher) that occurs as an independent Poisson process over time with a constant rate of 0.1 visits/hour. If a plant is watched for a one-hour period, the probability of a visit during that hour is  $P(\text{visit}) = 1 - e^{-0.1} = 0.095$ . If 100

TABLE 1. Measures of absolute and relative precision when a proportion,  $p$ , is estimated from a simple random sample of  $n$  observations.

Measure	Absolute measures	Relative measures
Variability	Standard error (SE): $\sqrt{p(1-p)/n}$	Coefficient of variation: $\sqrt{(1-p)/(pn)}$
Precision	$1/\sigma^2 = (1/SE)^2$	$(1/cv)^2 = p^2/\sigma^2$

plants are watched, each for one hour, the SE of the visit probability is 0.029. If a plant is watched for a one-minute period, the probability of a visit during that one minute is  $P(\text{visit}) = 1 - e^{-0.1/60} = 0.00167$ . If 100 plants are watched, each for one minute, the SE of the one-minute-visit probability is much smaller, 0.00408. This apparent increase in precision is an artifact of a rarer event.

Measures of relative precision, including the cv or relative variance (Table 1), avoid this counterintuitive behavior by expressing the precision relative to the probability of the event. Rarer events are less precisely estimated, on a relative scale, than are more common events. Measures of relative precision are also unitless, unlike absolute measures of precision. To continue the example from the previous paragraph, when 100 plants are watched for an hour each, the cv is  $0.029/0.095 = 32\%$ . When 100 plants are watched for a minute each, the cv is much larger,  $0.00408/0.00167 = 244\%$ . The estimate from the shorter observation period is less precise, when measured using relative precision. The cv is one of many possible measures of relative precision. Others include the reciprocal of the cv, which is larger for more precise estimates, or the reciprocal of the  $cv^2$ , which is the relative analog of the Bayesian measure of precision ( $1/s^2$ ).

#### HOW CAN THE PRECISION BE INCREASED?

Imprecise estimates of a probability are not unique to the *Darlingtonia* example. They are common whenever events are rare. When an event has a probability of occurring less than five times in a hundred trials ( $P(\text{event}) = \pi = 0.05$ ), the coefficient of variation (cv) of estimates of this probability from samples of  $N = 100$  are larger than 50%. The cv can exceed 300% when the event is very rare ( $\pi < 0.01$ ; Fig. 1). On the other hand, when events are common,  $\pi$  can be estimated with high precision even with moderate sample sizes. If  $\pi = 0.5$ , a cv of 10% can be obtained with a sample size of  $N = 100$ .

Increasing the total sample size  $N$  increases the precision of the estimate of the probability of a rare event. For example, increasing  $N$  from 100 to 500 independent observations decreases the cv by a factor of  $\sqrt{1/5}$ . However, if an event is rare, a precise estimate (cv  $\leq 10\%$ ) requires very large sample sizes. For example, if  $\pi = 0.01$ , a sample size of  $N = 9900$  is required to achieve cv = 10%. Such large sample sizes may be expensive, difficult, or impossible to obtain.

Alternatively, the precision of the estimate of a rare event can be increased by combining the primary data (e.g., the observed numbers of visits and captures in the *Darlingtonia* dataset) with auxiliary data that provides additional information about the probability of the rare event. Auxiliary data may come from many different sources, of which we discuss four: prior information, stratified sampling, covariates, and aggregated data. We use the *Darlingtonia* data set to illustrate the methods by which auxiliary data can be used to improve the precision of point estimates of probability. We will compare methods using relative precision (cv) and absolute precision ( $0.0001/s^2$ ). We use  $0.0001/s^2$  instead of  $1/s^2$  as a measure of absolute precision because the factor of 0.0001 converts the variance of a proportion to the variance of a percentage, which provides a more intuitive scale for interpreting absolute precision.

#### *Incorporating prior information using Bayesian methods*

Prior information about the probability  $\pi$  of a rare event can be derived from other studies of the same or related species, in the same or in different locations. Bayesian inference can be used to combine this prior information with the observed data (Ellison 1996). If probability estimates from the primary data are similar to those provided by the prior information, the combined estimate will have greater precision than the estimate based on the primary data alone.

Bayesian inference treats parameters, such as the probability  $\pi$  that *Darlingtonia* captures a wasp, as random variables described by statistical distributions (Barnett 1999, Ellison 2004). The distribution of each parameter summarizes both the expected value of the parameter and its variance. Bayesian inference uses the data (observations), along with information known about the parameter(s) before the data are analyzed (the prior probability distribution, or simply the prior) to construct a new distribution (the posterior probability distribution, or simply the posterior) that expresses what is known about the parameter after the data are analyzed.

The posterior is computed from the data and the prior using Bayes' Theorem (Ellison 2004):

$$f(\pi | C, V) = \frac{f(C | V, \pi)f(\pi)}{\int f(C | V, \pi)f(\pi) d\pi}. \quad (1)$$

In Eq. 1,  $f(\pi)$  is the prior,  $f(C | V, \pi)$  is the likelihood

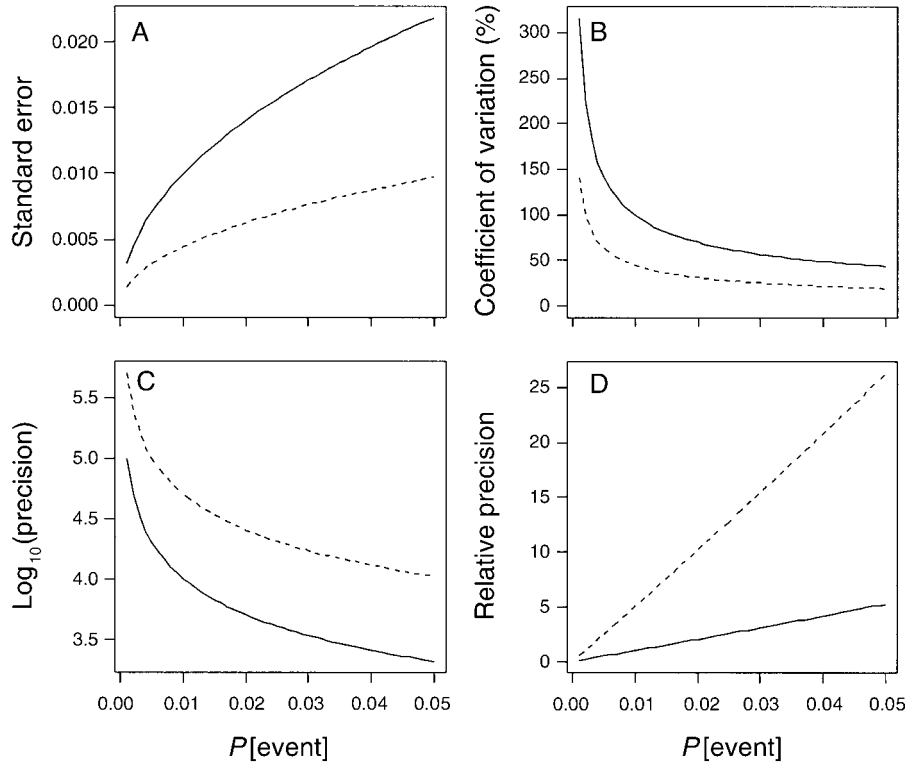


FIG. 1. The relationship between four measures of precision (Table 1) and the probability of an event when an event is rare (probability  $\pi < 5\%$ ). (A) standard error (SE), (B) coefficient of variation (CV), (C) absolute precision (presented as a log value), (D) relative precision,  $(1/ CV)^2$ . Each measure of precision is calculated for sample sizes of  $N = 100$  (solid line) and  $N = 500$  (dashed line).

of the observations, and  $f(\pi | C, V)$  is the posterior. The vertical bars indicate which quantities are considered fixed. That is,  $f(C | V, \pi)$  is the probability distribution of  $C$  (the number of captures), conditional on the fixed values of  $V$  (the number of visits) and  $\pi$  (the capture probability). The integral in the denominator is a normalizing constant that ensures that the posterior distribution is a valid probability distribution (i.e.,  $0 \leq f(\pi | C, V) \leq 1$ ). Using Bayes' Theorem requires that the distributions of both the data and the prior be specified.

In the *Darlingtonia* data set, the data ( $f(C | V, \pi)$ ) are the number of captures observed in a certain number of visits. A binomial distribution is commonly used to model count data when the outcomes (capture or not) are independent, the probability of success (capture) is the same for all visits, and where the number of success (captures) cannot exceed the number of visits (Gotelli and Ellison 2004).

When the data follow a binomial distribution, a Beta distribution,

$$\pi \sim \text{Beta}(\alpha, \beta) \tag{2}$$

is a convenient choice for the prior because the integral in the denominator of Eq. 1 can be evaluated analytically (Gelman et al. 1995). The values of the parameters  $\alpha$  and  $\beta$  in the beta distribution (Eq. 2) summarize our

knowledge of the capture probability before the data are analyzed. When  $\alpha > 1$  and  $\beta > 1$ , the mean  $\mu$  of the Beta( $\alpha, \beta$ ) distribution equals  $\alpha/(\alpha + \beta)$  and the mode equals  $(\alpha - 1)/(\alpha + \beta - 2)$ . Because the Beta distribution is skewed, the mode is the more appropriate measure of location. The variance  $\sigma^2$  of a Beta( $\alpha, \beta$ ) distribution is  $\mu(1 - \mu)/(\alpha + \beta + 1)$ . The posterior distribution given by Eq. 1 is also a Beta distribution (Gelman et al. 1995), and simulation of the posterior (e.g., with Markov chain Monte Carlo methods; Gilks et al. 1996) is not required. The parameters of the posterior depend on the parameters of the prior distribution ( $\alpha, \beta$ ) and the data ( $C, V$ ):

$$\pi | C, V \sim \text{Beta}(\alpha + C, \beta + V - C). \tag{3}$$

The mode of the posterior is an updated estimate of the capture probability, and the standard deviation is an updated estimate of the variability:

$$\text{mode} = \frac{\alpha + C - 1}{\alpha + \beta + V - 2} \tag{4}$$

$$\text{SD} = \sqrt{\frac{(\alpha + C)(\beta + V - C)}{(\alpha + \beta + V)^2(\alpha + \beta + V + 1)}}. \tag{5}$$

The choice of prior distribution (i.e., of  $\alpha$  and  $\beta$ ) influences the posterior distribution, although the influence of the prior is small when  $V$  and  $C$  are large.

TABLE 2. Parameters of Beta distributions used as prior distributions in the Bayesian analysis of the *Darlingtonia* data, along with the resulting posterior distributions of the capture probability.

Parameter	Prior				Posterior			
	Mode	SD	$\alpha$	$\beta$	Mode	SD	CV (%)	Precision
Data					0.0127	0.0089	70	12.5
Prior A	0.00931	0.0018	28	2873	0.0095	0.0018	19	31.5
Prior B	0.00931	0.0056	4.335	355.8	0.0104	0.0048	47	4.28
Prior C	0.00931	0.018	1.622	67.24	0.0117	0.0083	71	1.44
Prior D	0.00931	0.056	1.145	16.38	0.0124	0.010	80	0.99
Prior E	0.50	0.0833	1	1	0.0127	0.0108	85	0.86

Notes: All distributions except the uninformative prior (prior E) have a mode at the capture probability estimated by Newell and Nastase (1998) for the confamilial species *Sarracenia purpurea*. Prior A has a standard deviation (SD) equal to the sampling uncertainty reported by Newell and Nastase (1998). Priors B, C, and D have larger standard deviations to reflect uncertainty in the extrapolation across species and study sites. Prior E is the uninformative prior. The mode, SD, and CV reported in the first line ("Data") of the table are the frequentist estimates for these parameters.

The prior distribution can be determined in many ways (Berger 1985). One is to use an uninformative prior: a prior for which any value of capture probability is equally likely. For a probability between 0 and 1, the uninformative prior is a uniform(0, 1) distribution which is equivalent to a Beta(1, 1).

Another approach is to use previous research to determine a prior distribution. Newell and Nastase (1998) estimated the per-visit probability of insect capture by a related pitcher plant, *Sarracenia purpurea*, to be 0.0093 (27 captures in 2899 visits with observed outcomes). If *S. purpurea* and *Darlingtonia* are assumed to have similar per-visit probabilities of insect capture, these data can be used to specify the prior distribution for the analysis of the *Darlingtonia* data. One approach is to do a Bayesian analysis of Newell and Nastase's data, using a noninformative hyperprior ( $\alpha = 1$ ,  $\beta = 1$ ) and the data ( $C = 27$ ,  $V = 2899$ ) in Eq. 3. The resulting posterior distribution, Beta(28, 2873) can be used as the prior for the *Darlingtonia* analysis. This distribution has a mode = 0.0093 and SD = 0.0018. There is some uncertainty introduced by extrapolating between species and between studies. This uncertainty can be expressed by increasing the standard deviation of the prior. Accordingly, we used three additional prior distributions with the same mode but with increasingly larger standard deviations (Table 2). If multiple prior data sets are available, the variability among the data sets can be used to estimate the parameters of the prior distribution (Birkes and Dodge 1993).

The posterior modes for the five choices of prior are given in Table 2. The posterior mode lies between the mode of the prior distribution and the capture probability estimated solely from the data. When the prior distribution has a small standard deviation (e.g., Newell and Nastase's prior [A] in Table 2), the posterior mode is very close to the prior mode (Fig. 2). As uncertainty in the prior increases, the posterior mode approaches the estimate based on the data (Table 2, Fig. 2).

The SD, CV, and precision ( $0.0001/s^2$ ) of the posterior distribution summarize the uncertainty in the estimated capture probability. The improvement in precision

gained by incorporating prior information depends on the SD of the prior and on the difference between the expectation of the prior (the probability of prey capture by *Sarracenia*) and the expectation of the primary data (the probability of prey capture by *Darlingtonia*). If the two species are very similar (priors A or B in Table 1) the Bayesian estimate is considerably more precise. When the two species are less similar (priors C or D in Table 1) or if the prior information is uninformative (prior E in Table 1), the Bayesian estimate is less precise than the estimate based on the primary data alone.

#### Stratified sampling

Stratification, dividing the population into more homogeneous strata, can lead to a more precise estimate of a proportion when heterogeneity in  $\pi$  is associated with identifiable characteristics of the events (Thompson 2002). Stratification can be used to estimate the probability of a rare event by dividing the population (e.g., all possible visits by wasps to *Darlingtonia*) into subgroups that have different capture probabilities. For example, one stratum may have a very small capture probability, another may have a slightly larger capture probability, and a third stratum may have a large capture probability. Stratification increases the precision of the estimated probability by removing the variability between strata. In this example we assume a simple random sample of observations within each stratum, but many other sampling designs could be used (Thompson 2002).

Strata cannot be defined on the basis of the response variable itself. In other words, it is not appropriate to define one stratum as those plants that captured a wasp ( $N_1 = 2$ ) and the other as those plants that did not ( $N_2 = 155$ ). Instead, strata should be defined a priori based on knowledge specific to the system. For example, the size of the plant or the orientation of the pitcher might be associated with the capture probability. As an illustration, we will use strata defined by the orientation of the pitcher. Two different definitions of strata will be used to illustrate the importance of between-strata heterogeneity in capture probabilities (Table 3). One

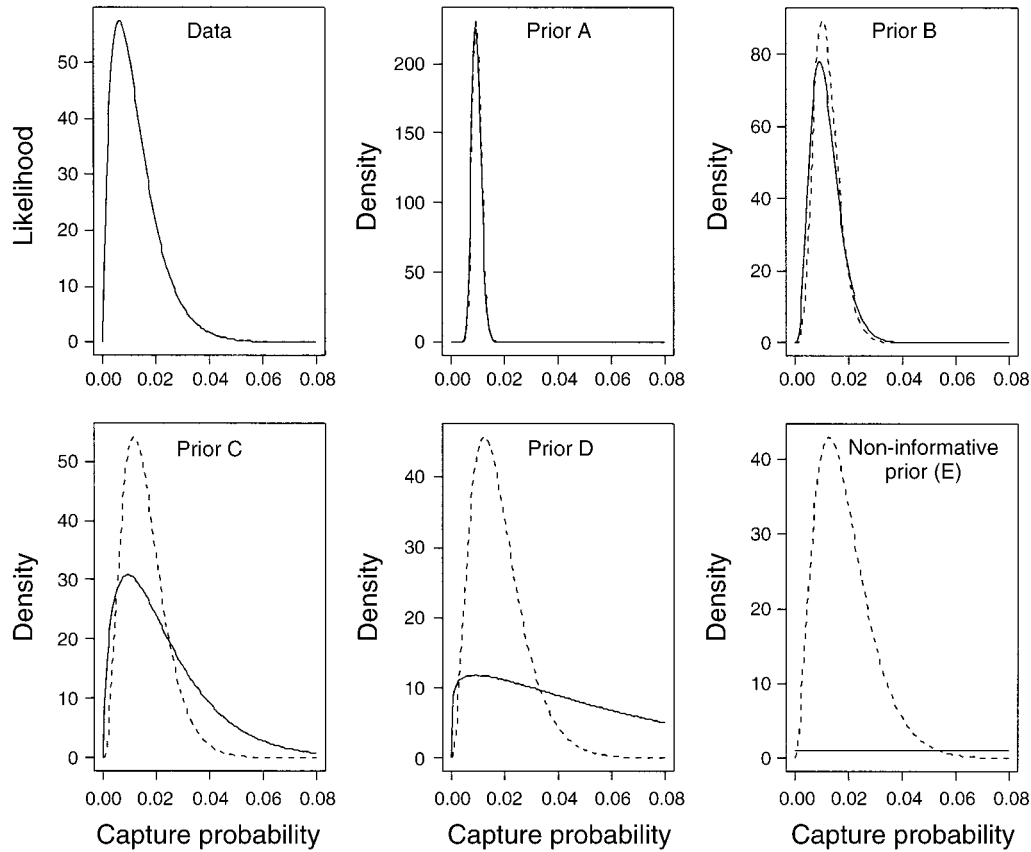


FIG. 2. Plots of the likelihood and prior (solid lines in density plots) and posterior (dashed lines in density plots) distributions for the five choices of prior distribution in Table 2. Note the different y-axis scales.

strata definition separates pitchers facing either 20° or 30° east of north from plants with all other orientations. The second strata definition separates those plants with orientations between 10° and 40° east of north from all other plants.

The estimated capture probability for the entire population from a stratified random sample is

$$\hat{p} = \frac{N_A \hat{p}_A + N_B \hat{p}_B}{N_A + N_B} \quad (6)$$

where  $N_A$  and  $N_B$  are the population sizes in the two strata, and  $\hat{p}_A$  and  $\hat{p}_B$  are the within-stratum estimates of the capture probability (Thompson 2002). When the event of interest only occurs in one stratum, the variance of the estimated probability is

$$\hat{s}^2(\hat{p}) = \left( \frac{N_A}{N_A + N_B} \right)^2 \left[ \frac{\hat{p}_A(1 - \hat{p}_A)}{n_A} \right] \quad (7)$$

where stratum A is the stratum including all the events and  $n_A$  is the sample size of that stratum. This variance estimator assumes that the population size is large relative to the sample size. If this assumption is not appropriate, a finite population correction factor should be included in the variance estimate (see Thompson 2002 for details).

TABLE 3. Precision of the estimate of the probability of capture using two different stratifications of the data.

Parameter	Stratum		Total
	A	B	
<b>Stratification 1†</b>			
Sample size (visits $V$ )	9	148	157
Number of captures ( $C$ )	2	0	2
Capture probability ( $\hat{p}$ )	0.222	0	0.0127
$SE_{\hat{p}}$	0.138	0	0.00792
$CV_{\hat{p}}$			62.1%
Precision			1.59
<b>Stratification 2‡</b>			
Sample size (visits $V$ )	19	138	157
Number of captures ( $C$ )	2	0	2
Capture probability ( $\hat{p}$ )	0.105	0	0.0127
$SE_{\hat{p}}$	0.0070	0	0.00852
$CV_{\hat{p}}$			66.9%
Precision			1.38

† Stratum A, plants with orientations of 20° or 30°; stratum B, all other plants.

‡ Stratum A, plants with orientations from 10° to 40°; stratum B, all other plants.

Estimating either  $\hat{p}$  (Eq. 6) or its variance  $s^2(\hat{p})$  (Eq. 7) requires knowledge of the relative sizes of the strata:  $N_A/(N_A + N_B)$ . The relative size of the strata may be estimated by independent criteria, such as stratum coverage or frequency in GIS databases. Because such information is lacking for the *Darlingtonia* population, we assume that the size of each stratum in the population is proportional to the size of each stratum in the sample:  $N_A/(N_A + N_B) = 9/157$  for the first stratum (pitchers oriented either 20° or 30°), and 19/157 for the second (pitchers oriented between 10° and 40°).

Because stratum sizes were estimated from the sample itself, the capture probability  $\hat{p}$  from either stratified sample (0.0127) is exactly the same as the estimate from the entire sample (Table 3). However, stratified sampling provides slightly more precise estimates of  $\pi$  (cv = 62.1% and 66.9% and absolute precision = 1.59 and 1.38 for the two definitions, respectively; Table 3) than do estimates based on the unstratified data (cv = 70.2%, absolute precision = 1.26). The first stratification (pitchers oriented either 20° or 30° vs. all others) is more precise than the second (Table 3) because the former has a larger between-strata difference in capture probability.

Stratification is especially useful when the probability of a rare event varies greatly among a small number of strata. However, if there are many strata, the number of observations per stratum is likely to be small and the stratum-specific probability will be poorly estimated.

#### Models incorporating covariates

Additional characteristics of the individuals may be measured. If these characteristics are associated with the rare event, they could be used either to stratify the observations (as in the previous approach) or to construct a model, e.g., a logistic regression model (Hosmer and Lemeshow 1989), to predict  $\pi$  for a specified set of covariates. The overall capture probability can be estimated by combining the model with information about the distribution of covariates in the population. The distribution can be enumerated when covariate information is available for all elements of the population or estimated from a simple random sample of the population. If the event is very rare (<10 events per covariate incorporated in the logistic regression; Van Belle 2002), logistic regression may not be useful for modeling the probability of very rare events.

In some cases, the event of interest is determined by an underlying continuous random variable. One example of this approach is the analysis of flood frequencies (Haan 2002). Floods are defined when water level exceeds a critical height for a specific patch of ground. The probability of flooding is the probability that the water level exceeds the critical height. Flood-frequency analysis uses a model for the distribution of water levels to estimate the probability of flooding (Hahn 2002).

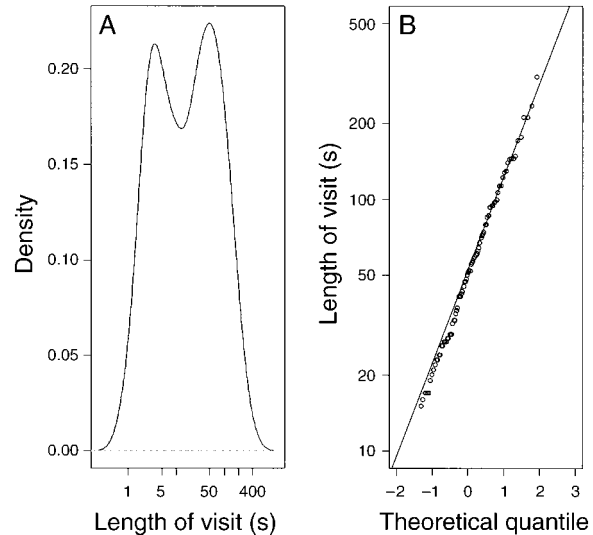


FIG. 3. (A) Probability density of visit length, estimated using a kernel smoother. The two modes are at 4 s and 50 s. The trough between the two peaks is centered at 13.5 s. (B) A log-normal quantile-quantile plot of the 86 visit lengths in the upper peak (visit lengths > 13.5 s). The theoretical quantiles were calculated after accounting for truncation (no value less than 13.5 s) and censoring (two captures with lengths > 307 s).

We used this last approach to estimate the probability that a wasp is captured by modeling the distribution of the length of time (visit lengths) that a wasp spends in a single pitcher. Visit lengths for noncaptured wasps ranged from a minimum of 1 s to a maximum of 307 s, with a median of 17 s. The empirical distribution of logarithmically transformed visit lengths is bimodal, with peaks at 4 and 50 s (Fig. 3A). The distribution of the logarithmically transformed values in the upper peak is very close to a normal distribution, as shown by a quantile-quantile plot (Fig. 3B). A two-component normal mixture model was fit to the log-transformed visit lengths by maximum likelihood. The two observed captures were considered censored observations, i.e., visit length > 307 s. The upper peak was estimated to contain  $\pi = 59.0\%$  of the visit lengths and have a normal distribution with  $\bar{x} = 3.9$  and  $SD = 0.94$ .

The probability that a visit length exceeds  $s$  seconds is estimated using the normal cumulative distribution function,  $\Phi(z)$ :

$$\hat{P}(\text{visit length} > s) = v \left[ 1 - \Phi \left( \frac{\log s - \bar{x}}{SD} \right) \right] \quad (8)$$

where  $v$  is the probability that a visit is in the upper peak. Eq. (8) only applies to long visits where the contribution from the lower peak can be ignored.

Calculating the capture probability from this distribution requires specifying a critical visit length; any visit longer than that critical length is assumed to be a capture. This critical value could be determined from

knowledge of wasp behavior and energetics. Lacking that information, we used a critical visit length of 307 s, the longest observed visit that did not result in a capture. The estimated capture probability  $\hat{p}$  is the probability that a visit exceeds 307 s:  $\hat{p} = 0.59[1 - \Phi(1.99)] = 0.0137$ . The estimate  $\hat{p}$  of  $\pi$  is very sensitive to the choice of critical visit length. For example, if the critical length is 360 s, the estimate  $\hat{p}$  decreases to 0.0090.

Bootstrap resampling can be used to estimate the precision of  $\hat{p}$  (Efron 1981, Dixon 2001). The bootstrapped standard error of the capture probability is estimated to be 0.0095, corresponding to a cv of 66% and precision of 1.10. The estimate from the threshold model is less precise than the frequentist estimate if precision is measured using an absolute measure ( $0.0001/s^2$ ) and more precise if precision is measured by a relative measure (cv).

#### Using aggregated data from larger scales

The primary data used to estimate the probability of a rare event come from observations of individuals, such as detailed observations of 753 individual *Darlingtonia* plants. Such data provide information about both the number of events (e.g., captures) and the number of trials (e.g., visits). At larger spatial or temporal scales, we can obtain samples of entire populations and observe the total number of rare events over a given interval of time or space (e.g., Lawson and Williams 1994, Plummer and Clayton 1996). This sample yields the product of the rate of occurrence of the event  $\times$  the number of trials (e.g., capture rate  $\times$  visitation rate). We can glean indirect information about the rate at which the rare event occurs from this product. Combining the direct and indirect information using a statistical model provides a more precise estimate of the capture probability.

We collected aggregate data on the total number of wasps captured by *Darlingtonia* individuals at several nearby sites over one-hour and two-day periods (A. M. Ellison, R. J. Emerson, E. J. Farnsworth, N. J. Gotelli, C. M. Hart, H. R. Steinhoff, and S. E. Wittman, *unpublished data*). These aggregate data were much easier to collect; we simply counted the number of wasps trapped in each pitcher after one hour or two days, rather than collecting direct behavioral observations. However in the aggregate data, we only recorded the number of wasps successfully captured per pitcher; the number of visits to each pitcher by wasps was not recorded.

Direct observations of wasp behavior suggests that wasps are actively foraging at *Darlingtonia* pitchers only for a 4-h period (10:00–14:00 hours) each day, so the 2-d aggregate data were assumed to reflect all captures made during 8 h of wasp activity. In the aggregate data, a total of six wasps were captured in a total of 1416 plant-hours (162 plants in the 2-d sample = 1296 plant-hours + 120 plants in the 1-h sample).

This aggregate information can be combined with the detailed data using a model that relates captures, visits and aggregated data to capture efficiency and visitation rate.

We again use a binomial random variable to model  $C$  as a function of  $\pi$  and  $V$ :

$$C|V \sim \text{Bin}(V, \pi). \quad (9)$$

If visits are rare and independent of each other, the number of visits in the primary data (direct observation of visits and captures) follows a Poisson distribution:

$$V \sim \text{Poiss}(\mu D) \quad (10)$$

where  $\mu$  is the mean number of visits per plant hour and  $D$  is the total number of plant-hours of detailed observations.

The same model (Eqs. 9 and 10) applies to the aggregate data, except that we did not observe the number of visits  $V$ . A capture in the aggregated data represents two events: a wasp visits a plant, and then the wasp is captured. If the probabilities of visitation and capture are constant,  $W$ , the total number of captured wasps in the aggregate data also has a Poisson distribution:

$$W \sim \text{Poiss}(\pi\mu A) \quad (11)$$

where  $A$  is the total number of plant hours of aggregated observations. Because the aggregated information,  $V$ , and  $W$  follow Poisson distributions, it is also convenient to use a Poisson distribution for the number of captures (cf. Eq. 9):

$$C|V \sim \text{Poiss}(\pi V). \quad (12)$$

Note that the Poisson distribution approximates a binomial distribution when the counts of rare events (e.g., captures) are small (Gotelli and Ellison 2004).

The parameters  $\pi$  and  $\mu$  in Eqs. 9–12 can be estimated using maximum likelihood (Appendix A). When captures are modeled using a Poisson distribution (Eq. 12),  $\pi$  and  $\mu$  can be estimated using standard software for Poisson regression (Appendix B).

The estimated capture probability is  $\hat{p} = 0.0107$ , only slightly smaller than the estimate from using the detailed observational data alone (Table 4). However, incorporating the aggregate data increases the precision of this estimate; the cv is reduced to 36%, nearly 50% smaller than the cv of the estimate from only the detailed observational data (Table 4). The absolute precision is increased to 6.75, slightly more than five times the precision of the estimate from only the detailed observational data. To achieve an equally precise estimate using only direct observations of wasp foraging behavior would require just over 2000 plant-hours of continuous observation.

#### DISCUSSION

Auxiliary data come in many forms. We have illustrated four different methods of using auxiliary data to increase the precision of the estimate of the probability

TABLE 4. Summary of estimated capture probabilities and their coefficients of variation for five estimators of capture probability, ranked from least to most precise.

Estimator	Estimate ( $\hat{p}$ )	CV	Precision
Frequentist (proportion of captures)	0.0127	70%	1.26
Threshold model (visit > 307 s = capture)	0.0137	66%	1.10
Stratification (best)	0.0127	62%	1.59
Bayesian (using Prior B, Table 1)	0.0104	47%	4.28
Aggregated data	0.0107	36%	6.75

of a rare event (Table 4). For the *Darlingtonia* data, the most precise and appropriate estimate of capture probability was estimated from pooling direct observations with temporally aggregated data. This method led to an estimate that was about twice as precise as the estimate derived from the direct observations alone (Table 4). Bayesian inference using informative, narrow priors yielded slightly less precise estimates. Stratification increased the precision only slightly, whereas modeling the distribution of visit lengths or Bayesian inference using informative priors with very large variance or uninformative priors decreased the precision of the estimate (see also Ellison 2004).

Which method is best? The appropriateness of a particular method can be judged by examining the assumptions and the choices that each requires. Bayesian inference assumes that information from previous studies is available and is relevant to the problem at hand. The relevance can be quantified by choosing the standard deviation of the prior distribution; a small standard deviation (i.e., high precision) implies that the prior is strongly informative, whereas a large standard deviation (i.e., low precision) implies little prior information or prior ignorance. If there is more than one previous study, the between-study standard deviation can be used as an estimate of the prior standard deviation, but if only one previous study is available, more care is needed in setting the precision of the prior, and the value may appear to be arbitrary. In the *Darlingtonia* example, as in many studies of rare events, the precision of the prior was important because when the sample size is small, the posterior will reflect more of the prior. In typical Bayesian analyses reported in the literature, data are more abundant, and the posterior reflects the likelihood of the data more strongly than the prior (Gelman et al. 1995, Ellison 1996, 2004).

Stratification requires strata that can be defined by characteristics other than the response variable. Stratification is most effective when event probabilities differ markedly between the strata. Correct use of stratification also requires that the sizes of each stratum in the sampled population are known. In the *Darlingtonia* example, we chose strata and estimated the sizes of the strata from the sample data. Realistic criteria and supporting information should be used to justify whatever strata are chosen.

Similarly, the modeling approach that incorporates covariates depends on a choice of a threshold value at

which a rare event is said to have occurred. In some situations, such as analysis of flood frequencies (Haan 2002) or the probability of structural failure (Heffernan and Tawn 2004), the threshold can be identified clearly and objectively supported. In other situations, such as the *Darlingtonia* example, the threshold must be derived from the data (e.g., the length of a wasp visit designated as a capture was determined from the distribution of visit lengths). Deriving thresholds from the data must be done cautiously, and should be justified whenever possible using independent observations or methods.

Finally, pooling of direct observations and aggregated data assumes that probability of the rare event is the same in both sets of data. Using Poisson distributions for both assumes that there is no between-year or between-site heterogeneity in the rate at which the rare events occur. This assumption of heterogeneity is almost impossible to test when the total number of events is small. In the *Darlingtonia* example, this assumption was reasonable because the two data sets were collected over the same years in the same general area, and each dataset (direct observations and temporally aggregated data) included observations collected from various sites and multiple years.

Estimating the probability  $\pi$  of an event from a series of independent observations is a very common activity in ecology and environmental science. The standard frequentist estimator of  $\pi$ ,  $\hat{p} = \text{number of events } n / \text{number of observations } N$ , is unbiased and straightforward to calculate. However, if the event is rare, the estimate is very imprecise if  $N < 1000$ . By incorporating other kinds of information, some of which may be from other studies, ecologists can increase the precision and the usefulness of these estimates. Ecologists should be alert for ways to incorporate auxiliary data to improve the precision of conventional statistical estimates.

#### ACKNOWLEDGMENTS

Data collection was supported by grants from the Packard Foundation (to A. M. Ellison) and the National Science Foundation (DEB 03-01361 to A. M. Ellison and DEB 03-01381 to N. J. Gotelli).

#### LITERATURE CITED

Arber, A. 1941. On the morphology of the pitcher-leaves in *Heliamphora*, *Sarracenia*, *Darlingtonia*, *Cephalotus*, and *Nepenthes*. *Annals of Botany* 5:563–578.

- Barnett, V. 1999. Comparative statistical inference. Third edition. John Wiley and Sons, Chichester, UK.
- Berger, J. O. 1985. Statistical decision theory and Bayesian analysis. Second edition. Springer-Verlag, New York, New York, USA.
- Birkes, D., and Y. Dodge. 1993. Alternative methods of regression. John Wiley and Sons, New York, New York, USA.
- Chave, J. 2004. Neutral theory and community ecology. *Ecology Letters* **7**:241–253.
- Clark, J. S., M. Lewis, and L. Horvath. 2001. Population spread with variation in dispersal and reproduction. *American Naturalist* **157**:537–554.
- Dixon, P. M. 2001. The bootstrap and the jackknife: describing the precision of ecological indices. Pages 267–288 in S. M. Scheiner and J. Gurevitch, editors. Design and analysis of ecological experiments, 2nd edition. Oxford University Press, Oxford, UK.
- Efron, B. 1981. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika* **68**:589–599.
- Ellison, A. M. 1996. An introduction to Bayesian inference for ecological research and environmental decision-making. *Ecological Applications* **6**:1036–1046.
- Ellison, A. M. 2004. Bayesian inference in ecology. *Ecology Letters* **7**:509–520.
- Ellison, A. M., and N. J. Gotelli. 2001. Evolutionary ecology of carnivorous plants. *Trends in Ecology and Evolution* **16**:623–629.
- Foster, D. R., and J. D. Aber, editors. 2003. Forests in time: the environmental consequences of 1000 years of change in New England. Yale University Press, New Haven, Connecticut, USA.
- Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 1995. Bayesian data analysis. Chapman and Hall, London, UK.
- Gilks, W. R., S. Richardson, and D. J. Spiegelhalter. 1996. Markov chain Monte Carlo in practice. Chapman and Hall/CRC, Boca Raton, Florida, USA.
- Gotelli, N. J., and A. M. Ellison. 2004. A primer of ecological statistics. Sinauer Associates, Sunderland, Massachusetts, USA.
- Haan, C. T. 2002. Statistical methods in hydrology. Second edition. Iowa State University Press, Ames, Iowa, USA.
- Harper, J. L. 1977. The population biology of plants. Academic Press, New York, New York, USA.
- Heffernan, J., and J. Tawn. 2004. Extreme values in the dock. *Significance* **1**:13–17.
- Hosmer, D. W., and S. Lemeshow. 1989. Applied logistic regression. John Wiley and Sons, New York, New York, USA.
- Hubbell, S. P. 2001. The unified neutral theory of biodiversity and biogeography. Princeton University Press, Princeton, New Jersey, USA.
- Jetz, W., and C. Rahbek. 2002. Geographic range size and determinants of avian species richness. *Science* **297**:1548–1551.
- Lawson, A. B., and F. Williams. 1994. Armadale: a case study in environmental epidemiology. *Journal of the Royal Statistical Society, Series A* **157**:285–298.
- Lewisohn, R. L., S. A. Freeman, and L. B. Crowder. 2004. Quantifying the effects of fisheries on threatened species: the impact of pelagic longlines on loggerhead and leatherback sea turtles. *Ecology Letters* **7**:221–231.
- Magurran, A. E. 2003. Measuring biological diversity. Blackwell Press, Oxford, UK.
- Marriott, F. H. C. 1990. A dictionary of statistical terms. Longman Scientific and Technical, London, UK.
- Mayr, E. 1963. Animal species and evolution. Harvard University Press, Cambridge, Massachusetts, USA.
- Newell, S. J., and A. J. Nastase. 1998. Efficiency of insect capture by *Sarracenia purpurea* (Sarraceniaceae), the northern pitcher plant. *American Journal of Botany* **85**:88–91.
- Plummer, M., and D. G. Clayton. 1996. Estimation of population exposure in ecological studies. *Journal of the Royal Statistical Society, Series B* **58**:113–126.
- Rabinowitz, D. 1981. Seven forms of rarity. Pages 205–207 in H. Synge, editor. The biological aspects of rare plant conservation. John Wiley and Sons, New York, New York, USA.
- Roberts, D. L., and A. R. Solow. 2003. Flightless birds: when did the dodo become extinct? *Nature* **426**:245.
- Thompson, S. K. 2002. Sampling. Second edition. John Wiley and Sons, New York, New York, USA.
- Van Belle, G. 2002. Statistical rules of thumb. John Wiley and Sons, New York, New York, USA.
- Whelan, R. J. 1995. The ecology of fire. Cambridge University Press, Cambridge, UK.
- Williams, C. B. 1964. Patterns in the balance of nature and related problems in quantitative ecology. Academic Press, New York, New York, USA.
- Zamora, R. 1995. The trapping success of a carnivorous plant, *Pinguicula vallisneriifolia*: the cumulative effects of availability, attraction, retention, and robbery of prey. *Oikos* **73**:309–322.
- Zamora, R., J. M. Gomez, and J. A. Hodar. 1998. Fitness responses of a carnivorous plant in contrasting ecological scenarios. *Ecology* **79**:1630–1644.

#### APPENDIX A

A description of the likelihood function for combining detailed and temporally aggregated data is presented in ESA's Electronic Data Archive: *Ecological Archives* E086-059-A1.

#### APPENDIX B

The SAS code used to fit a Poisson regression to detailed and aggregate data is presented in ESA's Electronic Data Archive: *Ecological Archives* E086-059-A2.