J. Parasitol., 99(2), 2013, pp. 386–389

Estimating Prevalence: A Confidence Game

Derek A. Zelmer, Department of Biology and Geology, University of South Carolina Aiken, Aiken, South Carolina 29801. Correspondence should be sent to: *derekz@usca.edu*.

ABSTRACT: Prevalence is one of the few estimates that rarely are reported with an appropriate measure of error in the parasitological literature. A minimum sample size recommendation of 15 samples, based on the relationship between sample size and standard error, likely has led to a false degree of confidence because of the nonlinear relationship between standard error and "true" 95% confidence intervals (as determined by Monte Carlo simulation or integration of the Bayesian posterior). Given that 95% confidence intervals for proportions are influenced by both sample size and the actual estimate of the proportion, there is no "gold standard" sample size beyond which estimates of binomial proportions can be considered "reliable." This necessitates the reporting of confidence interval estimates that have been shown to be conservative, such as the Clopper-Pearson estimate, or robust, such as the Wilson score approximation, or the computationally intensive integration of the Bayesian posterior.

Bush et al. (1997) characterized prevalence as one of the least misused ecological descriptors in the field of parasite ecology. While this is almost certainly true in terms of consistency in how it is calculated, sample prevalence and other proportional measures are all too frequently reported with no measure of the error for the estimate. In a review of the first 5 issues of Volume 97 of the *Journal of Parasitology* (ignoring taxonomic data, where ranges are reported by convention) 49 papers reported sample means, and, in all but 3, standard deviation or standard error was reported as a measure of dispersion or accuracy. In contrast, data were presented as binomial proportions (generally as percentages) in 58 papers, 45 of which failed to report an error term for the estimates. Of the 13 papers that did account for some measure of error, 5 included the error calculated for the average prevalence, which measures the precision of the prevalences among the samples, and not the accuracy of the estimate itself; 5 others reported the standard error for a binomial proportion. Estimates of 95%, or 99%, confidence intervals (CIs) were given for binomial proportions in only 3 instances. The present investigation addresses some of the peculiarities of prevalence estimation and the means by which such estimates can best be rendered interpretable. The focus of this investigation is on producing meaningful error estimates but does not address the methods of testing for differences among such estimates.

Almost without exception, papers that did not include an estimate of error did report sample sizes, which does allow for calculation of CIs after publication. It is not clear, however, why a practice that would be considered unacceptable for other ecological parameters, e.g., abundance, is so readily accepted for estimates of prevalence. There is no justification for providing any estimate without some measure of its error (Krebs, 1998). For binomial proportions (such as prevalence), 95% CIs are the most interpretable, indicating the range within which one is 95% confident that the actual value of the parameter can be found. There are also other error estimates that can be associated with estimates of binomial proportions, but most do not have a consistent relationship with the 95% CI. The interpretation of a 95% CI does not depend on context, such as sample size or the proportion estimated, and so error estimates that do not covary in a linear fashion with 95% CIs are context-dependent descriptors and, therefore, cannot effectively communicate the error inherent in the estimate. Determining the mean and associated error from a sample of prevalences is a common practice but is unlikely to produce a meaningful measure of the error associated with the resulting prevalence estimate. While the mean of the sample prevalences from equally sized samples (or the weighted mean for unequal sample sizes) produces the same estimate of prevalence that would be obtained by pooling the samples, the error terms associated with these means do not necessarily relate to the confidence in the estimate of prevalence. The standard deviation of such samples describes only the precision among these samples and, rather than increasing in concert, does not covary with the 95% CIs generated by pooling the samples and calculating a single estimate (Fig. 1). Thus, the standard deviation of "mean prevalence" does not provide any indication of the accuracy of the prevalence estimate.

Because the prevalence estimates are themselves means, the calculation for sample standard deviation essentially represents what conventionally is referred to as standard error, i.e., the standard deviation of sample means around a population mean. When treating the prevalences as observations rather than means, the standard error of such samples (generated by dividing the standard deviation by the square root of the number of samples used to determine mean prevalence) does covary in a nearly linear fashion with 95% CIs when dealing with more than 10 samples (Fig. 2A) but becomes less predictable with 10, or fewer, samples (Fig. 2B). The standard error can provide an estimate of the accuracy of the prevalence estimate, but only if the sample sizes are equal and the number of samples is sufficient. Even when the aforementioned conditions are met, the estimate of confidence is indirect. More importantly, it does not seem logical to partition confidence into subsets, when it is possible to pool the samples into a single estimate, and generate appropriate and meaningful 95% CIs. Therefore, estimates of the error generated from multiple prevalence estimates cannot be recommended.

Another alternative to estimating the error associated with prevalence estimates is the determination of the standard error specific to binomial estimates. In essence, a proportion represents an average of zeros and ones, which means that a standard error can be calculated for an estimated binomial proportion ($\hat{p}$), such as prevalence, as

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Unfortunately, this error estimate does not have a linear or consistent (across proportions) relationship with 95% CIs for binomial proportions (Fig. 3) and is, therefore, not a useful measure of confidence in relation to sample size.

Jovani and Tella (2006) provide an excellent critique of some of the methods that have been employed in dealing with estimates of prevalence based on low sample sizes. They also used the relationship between standard error of binomial estimates and sample size to suggest that a minimum sample size of 15 host individuals for estimating prevalence was justified. Since that time, a number of papers have used that recommendation to justify or contextualize small sample sizes (e.g., Jovani et al., 2006; Geue and Partecke, 2008; Illera and Emerson, 2008; Klomp et al., 2008; Carrete et al., 2009; Raharivololna and Ganzhorn, 2009; Alcaise et al., 2010; Garamszegi, 2010; Lymbery et al., 2010; Rubio-Godoy et al., 2011). As noted above, the problem with relating this error to appropriate sample sizes is that it does not relate directly to accuracy of the estimate. Although reliance on a minimum sample size would be convenient, 95% CIs for binomial data vary continuously with sample size and with the estimate of prevalence. The 95% CI calculated from a sample size of 15 might well be "small enough" to address certain questions or comparisons,

DOI: 10.1645/GE-3168.1

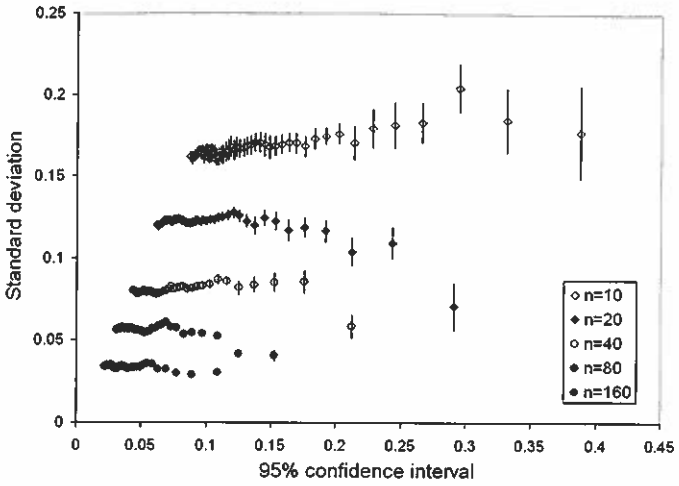

FIGURE 1. Relationship between the cumulative standard deviation (mean ± SE) of 10 trials of 2 (right side of graph) to 50 (left side of graph) random samples drawn from a population with a prevalence of 0.5 (50%), and the cumulative 95% CIs calculated from the pooled samples, for sample sizes of 10, 20, 40, 80, and 160 individuals. Error terms were calculated from proportions. The relationships shown are consistent with the relationships observed for other proportions.

but the calculation of 95% CIs is not exceptionally difficult and provides a direct indication of the types of questions or comparisons that can be addressed with a given dataset.

Confidence intervals generated using the standard error, such as the Wald estimator (a Gaussian approximation of the binomial distribution), fare poorly relative to the actual confidence limits, as determined by Monte Carlo simulation or integration of the Bayesian posterior (Agresti and Coull, 1998). They underestimate the true limits by a factor that can be several times the value of the type I error rate (Ross, 2003). The Wald CI is not recommended for sample sizes of less than 600, and even then only for proportions ranging from 0.2 to 0.8 (Ross, 2003).

Clopper-Pearson "exact" CIs (Clopper and Pearson, 1934) avoid approximation by relying on the relationship between the F distribution and the binomial distribution, taking the form

$$\left[1+\frac{n-x+1}{xF_{\nu_1,\nu_2,1-\alpha/2}}\right]^{-1}<\hat{p}<\left[1+\frac{n-x}{(x+1)F_{\nu_1',\nu_2',\alpha/2}}\right]^{-1}$$

for a proportion ($\hat{p}$) estimated as x/n, where

$$\nu_1 = 2x,$$

$$\nu_2 = 2(n-x+1),$$

$$\nu_1' = 2(x+1),$$

$$\nu_2' = 2(n-x).$$

The values of F can be determined from a table or, more conveniently, determined in Microsoft Excel using the formulas

$$= \mathrm{FINV}\left(1-\left(\frac{\alpha}{2}\right),\nu_1,\nu_2\right) \quad \text{and} \quad = \mathrm{FINV}\left(\frac{\alpha}{2},\nu_1',\nu_2'\right).$$

For example, 10 infected hosts in a sample of 23 produce CIs for $\hat{p}=0.43$ that are

$$0.23<0.43<0.66.$$

The performance of the Clopper-Pearson interval greatly exceeds that of Wald intervals (Agresti and Coull, 1998), especially at small sample sizes.

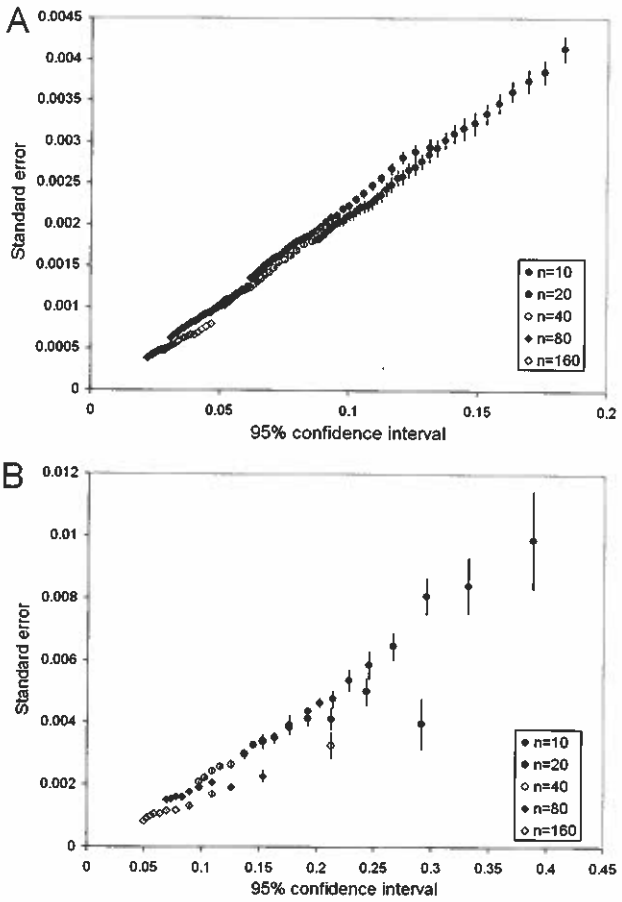

FIGURE 2. Relationship between the cumulative standard error (mean ± SE) of 10 trials of random samples drawn from a population with a prevalence of 0.5 (50%), and the cumulative 95% CIs calculated from the pooled samples, for sample sizes of 10, 20, 40, 80, and 160 individuals. Error terms were calculated from proportions. The relationships shown are consistent with the relationships observed for other proportions. (**A**) Cumulative mean for 11 (right side of graph) to 50 (left side of graph) random samples. (**B**) Cumulative mean for 2 (right side of graph) to10 (left side of graph) random samples.

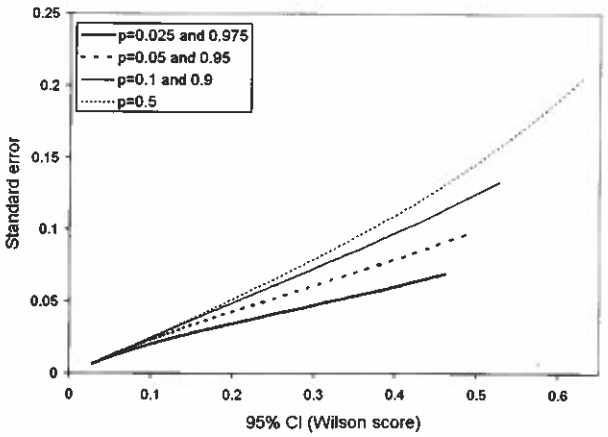

FIGURE 3. The relationship between the standard error of estimates of binomial proportions and the width of the Wilson score 95% CIs for sample sizes ranging from 5 (right side of graph) to 500 (left side of graph).

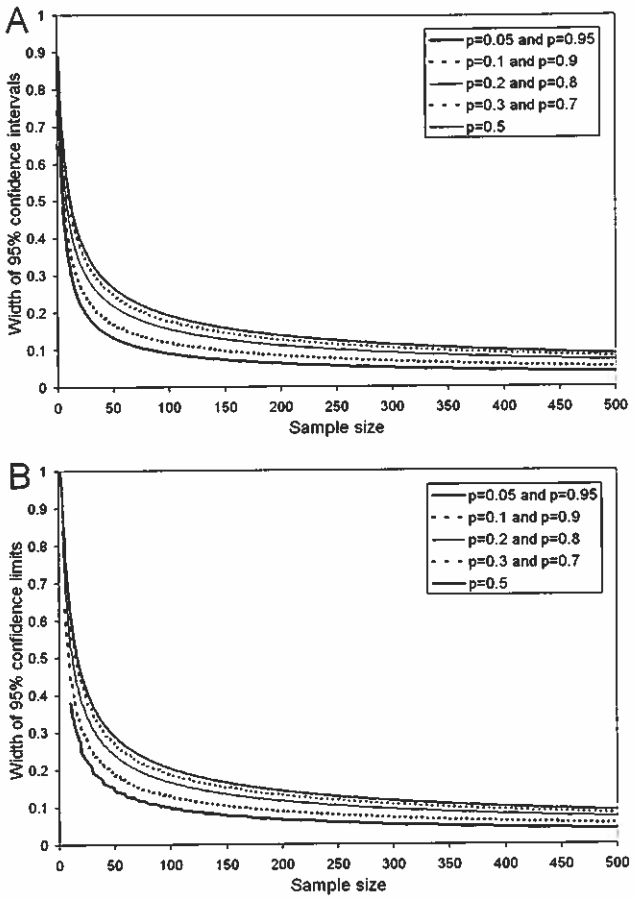

FIGURE 4. Width of 95% CIs for binomial proportions as a function of sample size and estimated proportion. (A) Wilson score intervals. (B) Clopper-Pearson intervals.

Although the error relative to actual confidence limits can be as high as 60% of the type I error rate, it typically is on the order of 20%, and the error is always positive, i.e., never smaller than the true CI, making it a conservative estimator (Ross, 2003).

The Wilson score interval, although based on a Gaussian approximation, is more robust and less biased than the Clopper-Pearson and Wald estimates (Agresti and Coull, 1998) and is computationally simple. The standard error employed for the Wilson score interval is based on the null hypothesis of the parameter, rather than on its estimate (Wilson, 1927; Agresti and Coull, 1998), taking the form

$$\frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p}) + \frac{z^2_{\alpha/2}}{4n}}{n}}}{1 + \frac{z^2_{\alpha/2}}{n}},$$

where $\hat{p}$ is the proportion estimated from a sample size of n, and $z_{\alpha/2}$ is the normalized value (z-score) for a type I error rate of α (where the CI is $1 - \alpha$). The normalized value for the percentile can be determined from a table or recovered in Microsoft Excel using the function

$$= \text{NORMSINV}[1 - (\alpha/2)].$$

Thus, the Wilson score interval for 10 infected hosts in a sample of 23 is

$$0.26 < 0.43 < 0.63,$$

which is a slightly narrower CI than the Clopper-Pearson estimate.

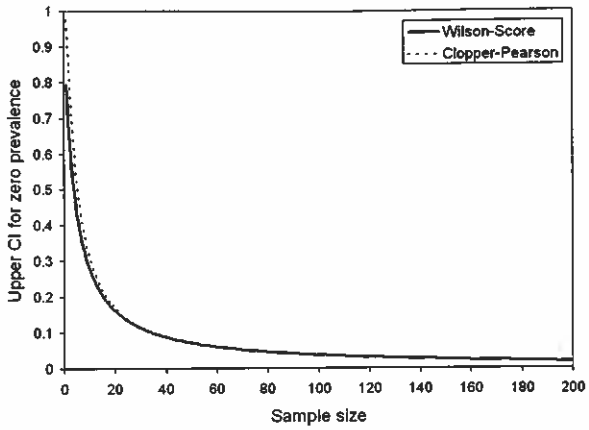

FIGURE 5. Relationship between the upper CI for a binomial proportion estimate of zero, and the size of the sample used to produce the estimate.

CIs based on integration of the Bayesian posterior (IBP) have a vanishingly small degree of error (within 10^{-5} of the desired type I error rate) that is independent of sample size and the estimated proportion but computationally intensive (Ross, 2003). Although IBP is the best choice for estimation of 95% CIs, both the Clopper-Pearson interval and the Wilson score interval can be recommended for those without access to the appropriate software, with the former being more conservative (its value is never less than the true 95% CIs) and, therefore, producing wider CIs at low proportions (Ross, 2003).

Figure 4 shows the narrowing of the combined 95% CI width with increasing sample size for the Wilson score (Fig. 4A) and Clopper-Pearson (Fig. 4B) estimates. It is important to note that these widths are symmetrical around the estimated proportion only when the estimated proportion is 0.5, and they become progressively more asymmetrical as estimates approach the minimum and maximum values of 0 and 1, respectively. This bounding limits the possible alternative values for a given estimate, e.g., an estimated proportion of 0.1 cannot have a lower CI less than 0, resulting in uncertainty being at a maximum when the estimated proportion is 0.5.

As with any parameter estimate, required sample sizes are determined by both the degree of confidence required and some idea of what the estimated proportion would be, making it difficult, and possibly inappropriate, to recommend a minimum sample size within the practical limits of most parasitological investigations. For a purely descriptive investigation, an estimated proportion of 0.5 produces the greatest interval widths for a given sample size, and a sample size of 40 hosts for the Wilson score interval (47 for the Clopper-Pearson interval) is necessary in order to generate upper and lower 95% CIs smaller than 0.15 around that estimate of 0.5. Reducing upper and lower 95% CIs to less than 0.1 would require sample sizes of 94 for the Wilson score interval and 104 for the Clopper-Pearson interval.

Because prevalence data imply presence or absence of a parasite species in a host population, it also is instructive to consider what can be inferred from an observed prevalence of zero at a given sample size. The upper CI of that estimate is a useful measure of what such an investigation might be overlooking. Figure 5 depicts the relationship between the upper Clopper-Pearson and Wilson score intervals for an estimated proportion of zero. With a sample size of 15 hosts, an observed prevalence of zero could be expected for species with a true prevalence of 20%. In order to be 95% confident that species with a true prevalence of 10% will be detected, i.e., an upper confidence limit for an estimate of 0 that is less than 0.1, a sample size of more than 35 would be required. To reliably detect species with a prevalence of 5%, a sample size of 75 would be appropriate.

Although a discussion of the analyses appropriate for comparing prevalence data is beyond the scope of this paper, one can make cursory

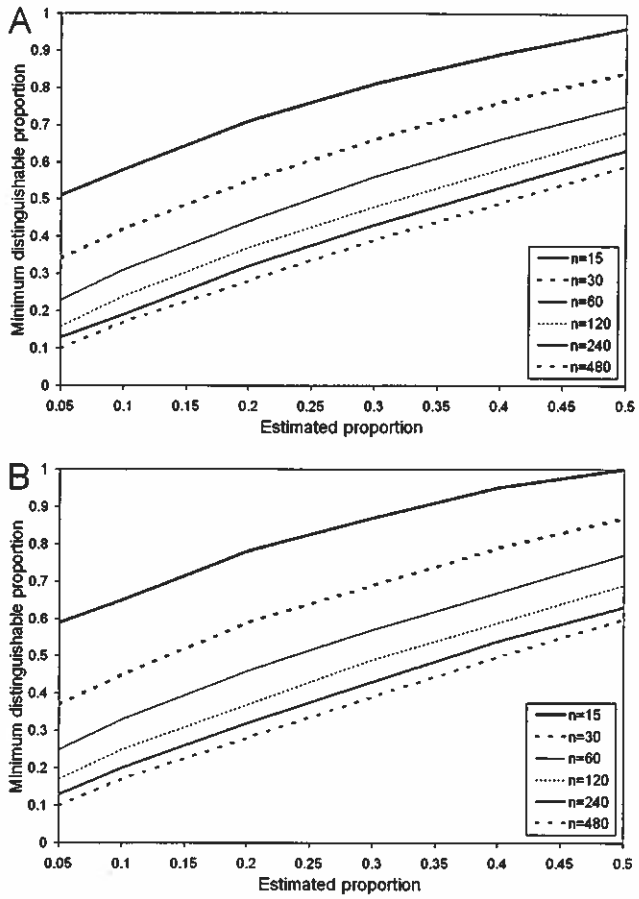

FIGURE 6. The minimum estimated proportion that can be distinguished with 95% confidence, i.e., CIs do not overlap, from an estimated proportion as a function of sample size, assuming the same sample size for both estimates. (A) Wilson score. (B) Clopper-Pearson.

comparisons among estimates of prevalence using the overlap between 2 or more sets of CIs. Thus, in terms of evaluating appropriate sample sizes, a consideration of the overlap between 2 sets of 95% CIs may be warranted. Figure 6 depicts the minimum estimated proportion that can be distinguished with 95% confidence, i.e., the confidence limits do not overlap, from a given estimated proportion as a function of sample size. The minimum sample size recommended by Jovanni and Tela (2006) would make a prevalence of 5% ($0.01 < 0.05 < 0.28$) indistinguishable from estimates ranging up to 50% ($0.27 < 0.5 < 0.73$) when the less conservative Wilson score interval is employed. Considering the maximum interval width for any sample size, which occurs at a prevalence of 50%, Wilson score intervals distinguish a prevalence of 50% ($0.39 < 0.5 < 0.61$) from a prevalence of 73% ($0.62 < 0.73 < 0.82$) at a sample size of 75, and distinguish a prevalence of 50% ($0.46 < 0.5 < 0.54$) from a prevalence of 59% ($0.55 < 0.59 < 0.63$) with a sample size of 480.

Although it seems intuitive that the use of presence-absence data would require smaller sample sizes than estimates generated using abundance data, careful examination of 95% CIs for proportions as a function of sample size make it clear that the sampling requirements, especially where comparisons are to be made, are at least as stringent. Regardless of sample size, which often is constrained by practical and not theoretical issues, an appropriate measure of the error in estimates of prevalence must be reported in order to provide an objective indication of the degree of confidence in that estimate, and the level of confidence with which one can make comparisons among prevalences. None of the commonly applied alternative estimates of error for binomial proportions has a consistent relationship with the 95% CIs, leaving the 95% CIs themselves as the best choice for presenting an estimate of the error associated with estimates of prevalence.

This manuscript was improved substantially by the thoughtful comments of Al Shostak and 2 anonymous reviewers.

LITERATURE CITED

AGRESTI, A., AND B. A. COULL. 1998. Approximate is better than "exact" for interval estimation of binomial proportions. American Statistician **52:** 119–126.

ALCAIDE, M., J. A. LEMUS, G. BLANCO, J. L. TELLA, D. SERRANO, J. J. NEGRO, A. RODRÍGUEZ, AND M. GARCÍA-MONTIJANO. 2010. MHC diversity and differential exposure to pathogens in kestrels (Aves: Falconidae). Molecular Ecology **19:** 691–705.

BUSH, A. O., K. D. LAFFERTY, J. M. LOTZ, AND A. W. SHOSTAK. 1997. Parasitology meets ecology on its own terms: Margolis et al. revisited. Journal of Parasitology **83:** 575–583.

CARRETE, M., D. SERRANO, J. C. ILLERA, G. LÓPEZ, M. VÖGELI, A. DELGADO, AND J. L. TELLA. 2009. Goats, birds, and emergent diseases: Apparent and hidden effects of exotic species in an island environment. Ecological Applications **19:** 840–853.

CLOPPER, C. J., AND E. S. PEARSON. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika **26:** 404–413.

GARAMSZEGI, L. Z. 2010. The sensitivity of microscopy and PCR-based detection methods affecting estimates of prevalence of blood parasites in birds. Journal of Parasitology **96:** 1197–1203.

GEUE, D., AND J. PARTECKE. 2008. Reduced parasite infestation in urban Eurasian blackbirds (*Turdus merula*): A factor favoring urbanization? Canadian Journal of Zoology **86:** 1419–1425.

ILLERA, J. C., AND B. EMERSON. 2008. Genetic characterization, distribution and prevalence of avian pox and avian malaria in the Berthelot's pipit (*Anthus berthelotti*) in Macraronesia. Parasitological Research **103:** 1435–1443.

JOVANI, R., D. SERRANO, Ó. FRÍAS, AND G. BLANCO. 2006. Shift in feather mite distribution during the molt of passerines: The case of barn swallows (*Hirundo rustica*). Canadian Journal of Zoology **84:** 729–735.

———, AND J. L. TELLA. 2006. Parasite prevalence and sample size: Misconceptions and solutions. Trends in Parasitology **22:** 214–218.

KLOMP, J. E., M. T. MURPHY, S. B. SMITH, J. E. MCKAY, I. FERRERA, AND A. REYSENBACH. 2008. Cloacal microbial communities of female spotted towhees *Pipilo maculates*: Microgeographic variation and individual sources of variability. Journal of Avian Biology **39:** 530–538.

KREBS, C. J. 1998. Ecological methodology, 2nd ed. Benjamin Cummings, San Francisco, California, 624 p.

LYMBERY, A. J., M. HASSAN, D. L. MORGAN, S. J. BEATTY, AND R. G. DOUPÉ. 2010. Parasites of native and exotic freshwater fishes in south-western Australia. Journal of Fish Biology **76:** 1770–1785.

RAHARIVOLOLNA, B. M., AND J. U. GANZHORN. 2009. Gastrointestinal parasite infection of the Gray mouse lemur (*Microcebus murinus*) in the littoral forest of Mandena, Madagascar: Effects of forest fragmentation and degradation. Madagascar Conservation and Development **4:** 103–112.

ROSS, T. D. 2003. Accurate confidence intervals for binomial proportion and Poisson rate estimation. Computers in Biology and Medicine **33:** 509–531.

RUBIO-GODOY, M., G. PÉREZ-PONCE DE LEÓN, B. MENDOZA-GARFIAS, M. C. CARMONA-ISUNZA, A. NÚÑEZ-DE LA MORA, AND HUGH DRUMMOND. 2011. Helminth parasites of the blue-footed booby on Isla Isabel, México. Journal of Parasitology **97:** 636–641.

WILSON, E. B. 1927. Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association **22:** 209–212.