Supplementary materials for this article are available at https://doi.org/10.1007/s13253-023-00578-7.

Sample Size for Estimating Disease Prevalence in Free-Ranging Wildlife Populations: A Bayesian Modeling Approach

James G. Booth, Brenda J. Hanley, Florian H. Hodel, Christopher S. Jennelle, Joseph Guinness, Cara E. Them, Corey I. Mitchell, Md Sohel Ahmed, and Krysten L. Schuler

A two-parameter model and a Bayesian statistical framework are proposed for estimating prevalence and determining sample size requirements for detecting disease in free-ranging wildlife. Current approaches tend to rely on random (ideal) sampling conditions or on highly specialized computer simulations. The model-based approach presented here can accommodate a range of different sampling schemes and allows for complications that arise in the free-ranging wildlife setting including the natural clustering of individuals on the landscape and correlation in disease status from transmission among individuals. Correlation between individuals and the sampling scheme have important consequences for the sample size requirements. Specifically, high within cluster correlations in disease status can reduce sample size requirements by reducing the effective population size. However, disproportionate sampling of small subsets of subjects from the greater target population, combined with high correlation of disease status, tends to inflate sample size requirements, because it increases the likelihood of sampling multiple animals within the same highly correlated clusters, resulting in little additional information gleaned from those samples. Our results are consistent with those generated using both previously established approaches and extend their ability to adapt to

J. G. Booth · J. Guinness, Department of Statistics and Data Science, Cornell University, 129 Garden Avenue, Ithaca, NY 14853, USA.
B. J. Hanley · Md S. Ahmed, Cornell Wildlife Health Lab, Public and Ecosystems Health, Cornell University, 240 Farrier Road, Ithaca, NY 14853, USA.
F. H. Hodel, Department of Fisheries and Wildlife, Michigan State University, 480 Wilson Road, East Lansing, MI 48824, USA.
C. S. Jennelle, Minnesota Department of Natural Resources, Nongame Wildlife Program, 500 Lafayette Road, Saint Paul, MN 55155, USA.
C. E. Them, Cara Them Consulting, LLC, Corvallis, OR 97330, USA.
C. I. Mitchell, Desert Centered Ecology, LLC, Tucson, AZ 85716, USA.
J. G. Booth (✉), Cornell University, 1172 Comstock Hall, Ithaca, USA
(E-mail: *jb383@cornell.edu*).
K. L. Schuler (✉), Cornell University, A2-230 Animal Health Diagnostic Center, Ithaca, USA
(E-mail: *ks833@cornell.edu*).
Present address: C. I. Mitchell, U.S. Fish and Wildlife Service, Tucson, AZ, USA.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 29, Number 3, Pages 438–454
https://doi.org/10.1007/s13253-023-00578-7

additional biological, epidemiological, or societal sampling complications specific to wildlife health.

Key Words: Beta-binomial; Correlation; Sampling design; Wildlife disease surveillance; Wildlife health.

1. INTRODUCTION

The discipline of wildlife health is concerned with surveillance and monitoring of diseases that originate in wildlife, cause morbidity or mortality in wildlife, affect conservation status of wildlife at the population level, or cross the species barrier to affect domestic animals or humans. Notable diseases such as rabies, Ebola, dengue, and HIV have arisen in free-ranging wildlife, then crossed the species barrier to humans (Wolfe et al. 2007). Few technical standards exist to ensure the statistical reliability of surveillance data collection (or its sampling methodologies) before specimens are transferred to an approved laboratory for testing. Because health surveillance begins in the field, not the laboratory (Ryser-Degiorgis 2013), there exists a need to establish rigorous fundamental statistical guidance that allows investigators in wildlife health to substantiate claims regarding freedom-from-disease when surveillance data are used for population-scale inference (Martin et al. 2007).

The collection of samples to study disease in free-ranging wildlife is complicated, and practical limitations to the collection of subjects are numerous. Difficulties include: imbalanced migrations or dispersal of wildlife; (un)availability of samples by season or year; differential access to land parcels for data collection; discrepant land use policies or practices; and a patchwork of regulations or laws limiting harvest, collection, or take (Wobeser 1994; Cannon 2001; Stallknecht 2007; Ryser-Degiorgis 2013; Belsare et al. 2020). Consequently, historical inferences on free-ranging populations have overwhelmingly relied on the analysis of data sets that frequently include biased, non-random, or opportunistically-collected data (Belsare et al. 2020).

Appropriate sample sizes are crucial components to produce statistically rigorous inferences in quantitative explorations (Cochran 1977, Chapter 4). Officials in animal health studies have long computed target sample sizes using questions such as (1) what is the necessary sample size to substantiate freedom from disease?, (2) how many subjects should we sample to detect the presence (of a disease) at a low prevalence threshold?, and (3) how many samples need we collect to adequately monitor disease distribution or prevalence (Cannon and Roe 1982)? Until the turn of the century, sample sizes were computed using models containing simplistic assumptions such as perfect diagnostic tests and infinite populations (Cameron and Baldock 1998).

In recent years, scientists have accommodated more complexity using a variety of inferential strategies, including Bayesian approaches to infer disease prevalence under imperfect diagnostic tests (Joseph et al. 1995; Johnson et al. 2004; Branscum et al. 2004; Messam et al. 2008), and simulation-based approaches to compute desired sample sizes for particular pathogens and species under specific circumstances (Belsare et al. 2020). However, each of these strategies lack the generalizability necessary to render them foundational for the computation of target sample sizes in the field of wildlife health.

Correlation in disease status between individuals has important consequences in terms of sample size requirements for making inferences about disease prevalence that depend upon the sampling scheme. The influence of the sampling scheme on the sample size requirements has been highlighted recently by Belsare et al. (2020) who contrast simple random (SR) sampling with high-harvest (HH) sampling of clustered populations for surveillance of chronic wasting disease (CWD) in white-tailed deer (*Odocoileus virginianus*). High-harvest sampling occurs when a large proportion of the sample is taken from a small proportion of the clusters, reflecting the fact the hunters disproportionately concentrate their efforts in certain (presumably more accessible) habitats. While not explicitly modeling correlation, Belsare et al. (2020) simulate samples from a population in which 1% of the clusters are entirely diseased and the other 99% are entirely free from disease and show that HH sampling requires a significantly larger sample size to achieve the same level of disease detection as SR sampling.

Here, we describe a Bayesian statistical modeling approach for situations in which correlation between the disease status of individuals exists in a naturally clustered population. Inferences about disease prevalence, detection and required sample size are determined from the predictive distribution of population prevalence given either observed or hypothetical data obtained under a specified sampling design. In Sect. 2 we use chronic wasting disease (CWD) in white-tailed deer as motivation and case study for our modeling approach, although the methodology is intended to be quite general, applicable in a wide range of wildlife systems, and not limited to CWD in deer. The two parameter beta-binomial distribution (Rosner 2005), introduced in Sect. 3, is the key ingredient needed to model both prevalence and correlation. This distribution also arises as the posterior for the disease prevalence under the independence (no correlation) assumption which underlies most of the literature on detection probabilities and sample size requirements for disease monitoring. Bayesian and classical methods and formulas in the independence case are reviewed in Sect. 4. The Bayesian model that allows for within cluster correlations is described in Sect. 5 as well as details of how the Bayesian analysis can be implemented. Computational results are given in Sect. 6, and we conclude with some discussion, including possible extensions of the model, and recommendations for usage in Sect. 7.

2. CASE STUDY

Chronic wasting disease is a fatal neurodegenerative disease first characterized in the late 1960s in Colorado, USA that affects ungulate species from the family Cervidae (Williams and Young 1980). It has been detected in free-ranging cervids in 29 U.S. states and three Canadian provinces in North America, as well as Finland, Norway, South Korea, and Sweden. The disease is of particular concern due to potential long-term demographic impacts to infected cervid populations [Edmunds et al. (2016), DeVivo et al. (2017)] and the economic importance of deer hunting in the USA, which generated over 27 billion dollars in 2016 (Southwick-Associates 2018). Generally, wildlife agencies lead efforts to conduct all disease surveillance and management activities in their jurisdiction. Despite a variety of significant efforts however, CWD has proven extremely difficult to control (Uehlinger et al.

2016), meaning that early detection to improve the chance of successful management has emerged as a widespread priority across agencies.

Since 2010 in the U.S. state of Minnesota (MN), the Minnesota Department of Natural Resources (MNDNR) has repeatedly detected CWD in free-ranging white-tailed deer, leading to the creation of multiple management zones and surveillance areas considered to be high-risk areas for CWD invasion (MNDNR 2023). The MNDNR uses deer permit areas (DPA), which range in size from 88 to 3,662 km^2, to assign harvest regulations for deer population management goals. Tissues from these harvested deer are simultaneously used to surveil for CWD, and therefore, harvest and CWD sample effort are regularly reported at a DPA spatial scale.

We used harvest registration data generated in 2022 by the MNDNR at the DPA scale to parameterize the clusters in our sample size model. Specifically, we calculated the proportion of adult males (1.5 years and older) and antlerless deer (all females and male fawns) that might be expected in a surveillance sample in MN. Since deer harvest in MN occurs typically in the northern fall during rutting season, adult males are competing for female mates and are not generally associated with any particular group during that time (Hawkins and Klimstra 1970). We therefore treated these males as singletons (social groups of size 1). On the other hand, during that time females form matrilineal social groups centered upon an adult doe, very commonly containing a yearling doe, and two fawns (Hawkins and Klimstra 1970), with structuring ranging from 3 to 9 deer per female social group (Porter et al. 1991). In 2022, the MNDNR established CWD surveillance areas in 13 DPAs, reporting a total of 28,008 deer harvested, 51.6% were adult males (singletons) and 48.4% were antlerless deer (social groupings containing more than one animal).

3. BASIC NOTATION

Let Y denote the number of infected (positive) individuals in a population of size N. The goal is to describe in a probabilistic manner what can be said about the prevalence in the population, Y/N, based on a sample of size n. Towards this, let Y_s and Y_{-s} denote respectively the number of diseased individuals in the sample and the number in the remaining (unsampled) $N - n$ individuals. We will assume in what follows that positive diagnostic tests are validated and so that the specificity, p_{sp}, is equal to 1. To simplify the exposition we will also assume, initially, that there are no false negatives (i.e., sensitivity, $p_{se} = 1$), although this assumption will be relaxed later.

A Bayesian solution to this problem is to determine the predictive probability distribution, $P(Y = y|Y_s = y_s)$, for $Y = y_s + Y_{-s}$ based on a specific data generating mechanism and prior distribution on the model parameters; that is, we determine the probability there are y cases in the entire population given we observed y_s cases among the n sampled individuals. For a given sample size, n, we can then estimate the probability that the population prevalence is less than a threshold value π_0, or alternatively determine what sample size is necessary to conclude that the prevalence is low with high posterior probability. We note that the Bayesian approach emphasized in this paper is inherently conditional on an observed (or perhaps hypothetical) sample. An alternative, unconditional approach, is based on sampling from populations that were assumed to have been generated under a specific data-generating model, and to construct Monte Carlo estimates of quantities such as the detection probability. The latter is the approach taken, for example, by Belsare et al. (2020).

The main contribution of this work is to allow for correlation in the disease status between individuals in the same social group (also referred to as a cluster), using a beta-binomial model parameterized in terms of a correlation, ρ and a probability π. The beta-binomial distribution (Rosner 2005) provides a two-parameter probabilistic model for a nonnegative count, Y, with probability π and (positive) correlation ρ. Specifically, suppose that given π and ρ

$$P(Y = y|\pi, \rho) = \binom{N}{y} \frac{Be(y + \alpha, N - y + \beta)}{Be(\alpha, \beta)}, \quad y = 0, 1, \ldots, N, \tag{1}$$

where $\alpha > 0$ and $\beta > 0$ are related to π and ρ via the 1–1 transformation

$$\alpha = \frac{1-\rho}{\rho}\pi \quad \text{and} \quad \beta = \frac{1-\rho}{\rho}(1-\pi), \tag{2}$$

and $Be(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta)$ is the beta function. Note that the inverse of the parameter transformation in (2) is

$$\pi = \frac{\alpha}{\alpha + \beta} \quad \text{and} \quad \rho = \frac{1}{\alpha + \beta + 1}. \tag{3}$$

In what follows we write $Y \sim \text{BetaB}(N, \pi, \rho)$ and denote the probability mass function in (1) by $\text{BetaB}(y; N, \pi, \rho)$. Similarly, $\text{Beta}(\alpha, \beta)$ denotes a beta distribution and $\text{Beta}(x; \alpha, \beta)$ denotes the beta density function. As its name suggests, the beta-binomial distribution can be derived as a beta mixture of binomials. Specifically, suppose that Y is conditionally a binomial count, $Y|P \sim B(N, P)$, and that the 'success' probability, P, has a beta distribution with parameters $\alpha > 0$ and $\beta > 0$ satisfying (2). Then, the marginal distribution of Y is the beta-binomial distribution given in (1).

As ρ approaches zero with π fixed, the beta-binomial distribution (1) converges to the ordinary binomial distribution with N trials and success probability π; i.e., $Y \sim B(N, \pi)$. This limiting case corresponds to the model in which each individual's disease status is independent of other members of the population and positive with probability π. The sample size formulas that appear in the wildlife monitoring literature are largely based on this assumption. In the next subsection we describe some of these methods.

4. STANDARD APPROACHES

Under the binomial (independence) model the number of diseased individuals in a sample of size n, as well as the number in the unsampled portion of the population, are also binomial specifically, $Y_s \sim B(n, \pi)$ and $Y_{-s} \sim B(N - n, \pi)$. Since the beta distribution is conjugate to the binomial, the prior $\pi \sim \text{Beta}(\alpha_\pi, \beta_\pi)$ results in a beta posterior, $\pi|Y_s = y_s \sim \text{Beta}(y_s + \alpha_\pi, n - y_s + \beta_\pi)$ for π, and a beta-binomial posterior predictive distribution for

Table 1. Sample size requirements to conclude the disease prevalence is less than or equal to a specified threshold with 95% confidence, assuming no positives in the sample and no correlation in disease state among individuals

Theshold	Population Size						
	100	500	1,000	2,000	5,000	10,000	∞
1%	78	195	237	264	283	290	298
2%	63	118	131	139	144	146	148

the number of diseased individuals in the unsampled portion of the population, $Y_{-s}|Y_s = y_s \sim \text{BetaB}(N-n, \pi_{y_s,n}, \rho_n)$ where $\pi_{y_s,n} = (y_s + \alpha_\pi)/(\alpha_\pi + n + \beta_\pi)$ and $\rho_n = 1/(1 + \alpha_\pi + n + \beta_\pi)$. In particular, with a uniform (Bayes-Laplace) prior on π, if there are no diseased individuals in the sample, the predictive distribution for the number diseased in the population is $\text{BetaB}[N-n, 1/(n+2), 1/(n+3)]$. This predictive distribution can be used to directly evaluate the sample size required to declare the population prevalence less than any threshold, π_0, with high probability. For example, if $N = 1000$ and the sample size is $n = 237$, the predictive probability that the prevalence is less than or equal 1% is 0.9504. The corresponding sample size to declare the prevalence less than or equal to 2% with 95% probability is only 131, a fact that illustrates the diminishing return from additional sampling in the effort to gain greater certainty. Sample size requirements for a variety of other population sizes are given in Table 1

If the sample size is negligible relative to the population size, one needs to consider the limit of the predictive distribution of Y/N as $N \to \infty$. This limiting distribution is beta with mean $\pi_{y_s,n}$ and variance $\pi_{y_s,n}(1-\pi_{y_s,n})\rho_n$, or beta with parameters determined from $\pi_{y_s,n}$ and ρ_n using the the transformation in (2) (see Appendix A.2). For example, if there are zero positives in a sample of size $n = 298$, then $\alpha_s = 1$ and $\beta_s = 299$, and the predictive probability that the prevalence is less than 0.01 is 0.9505.

There is in fact a large existing literature on Bayesian inference about disease prevalence based on diagnostic tests. Examples include Joseph et al. (1995), Johnson et al. (2004), Branscum et al. (2004) and Messam et al. (2008). A common theme is the use of a beta prior for the prevalence, as we have here, but also beta priors for the sensitivity and specificity to account for imperfections in the diagnostic test. The posterior for prevalence given the testing data can then be computed using MCMC methods such as the Gibbs sampler. However, these papers all assume the binomial or multinomial model, or a hypergeometric model in the finite population setting, so they do not account for the correlation between individuals who cluster into social groups.

The standard non-Bayesian approach to the sample size question is based on the hypergeometric distribution of the number of diseased individuals in a sample of size n from a population of size N containing $Y = y$ diseased individuals given by

$$P(Y_s = y_s|Y = y) = \frac{\binom{y}{y_s}\binom{N-y}{n-y_s}}{\binom{N}{n}}, \quad \max(0, n+y-N) \le y_s \le \min(n, y). \tag{4}$$

In particular, for an entirely negative sample, we have

$$P(Y_s = 0 | Y = y) = \frac{\binom{y}{0}\binom{N-y}{n}}{\binom{N}{n}} = \prod_{k=0}^{y-1} \left(1 - \frac{n}{N-k}\right) \tag{5}$$

(see e.g., Cannon 2001, equation 4.2). For a given target prevalence, π_0, we can determine the sample size necessary for this probability to be less than, say $a = 0.05$; that is, find the minimum n such that $P(Y_s = 0 | Y = N\pi_0) \leq a$.

If the sample size is negligible relative to the population size, the probability in (5) can be approximated by $(1 - \pi)^n$, since each individual sampled has a probability $1 - \pi$ of testing negative, independently of others in the sample. In this case, the sample size problem reduces to finding the maximum n such that $(1 - \pi_0)^n \leq a$, which implies $n \geq \frac{\log a}{\log(1-\pi_0)}$ (see Cannon 2001, Eq. 4.5). For example, the number of negative tests required to declare the prevalence less than 1% with 95% confidence is $n = \log(0.05)/\log(0.99) = 298$, which agrees almost exactly with the Bayesian solution with a uniform prior on the prevalence.

5. A BETA BINOMIAL MODEL FOR POPULATIONS WITH CORRELATED DISEASE STATES AMONG INDIVIDUALS

We now consider a model in which the disease status of individuals is correlated. Specifically, suppose that $Y \sim \text{BetaB}(N, \pi, \rho)$. A Bayesian approach to predictive inference about the prevalence requires the specification of a prior distribution for both π and ρ. Since both lie in the interval (0, 1), a natural choice of prior for the pair is a product of independent beta distributions. Figure 1 shows the marginal distribution of Y for various choices of the beta priors including uniform distributions and priors that concentrate the latent prevalence near zero and/or the correlation near one. Of particular note is that, when the correlation is high (1b and 1d), the distribution of Y is concentrated at $Y = 0$ (all disease free) and $Y = N$ (all diseased).

Under the beta-binomial model, the distribution of the number of positives in a sample of size n is also beta-binomial; specifically, $Y_s | (\pi, \rho) \sim \text{BetaB}(n, \pi, \rho)$ (see Appendix A.1). If the sensitivity is less than one, the probability of $X \leq Y_s$ positive tests is given by

$$P(X = x | \pi, \rho) = \sum_{k=x}^{n} \binom{k}{x} p_{se}^x q_{se}^{k-x} \text{BetaB}(k; n, \pi, \rho), \tag{6}$$

where $q_{se} = 1 - p_{se}$.

However, the assumption of equicorrelated disease status for all pairs of individuals in the entire population is unrealistic in practice. Social structuring in wild animal populations is ubiquitous across taxa (Wilson 1975), and accounting for this biological realism is critical for estimating disease prevalence or calculating sample sizes to determine disease-free thresholds (Cannon and Roe 1982). To this end, we consider a population consisting of c clusters of sizes, $N_1, \ldots, N_c$, so that $N = \sum_{i=1}^{c} N_i$, and suppose that the number of infected animals in cluster i is $Y_i \sim \text{BetaB}(N_i; \pi, \rho)$ independently for $i = 1, \ldots, c$. Suppose that a

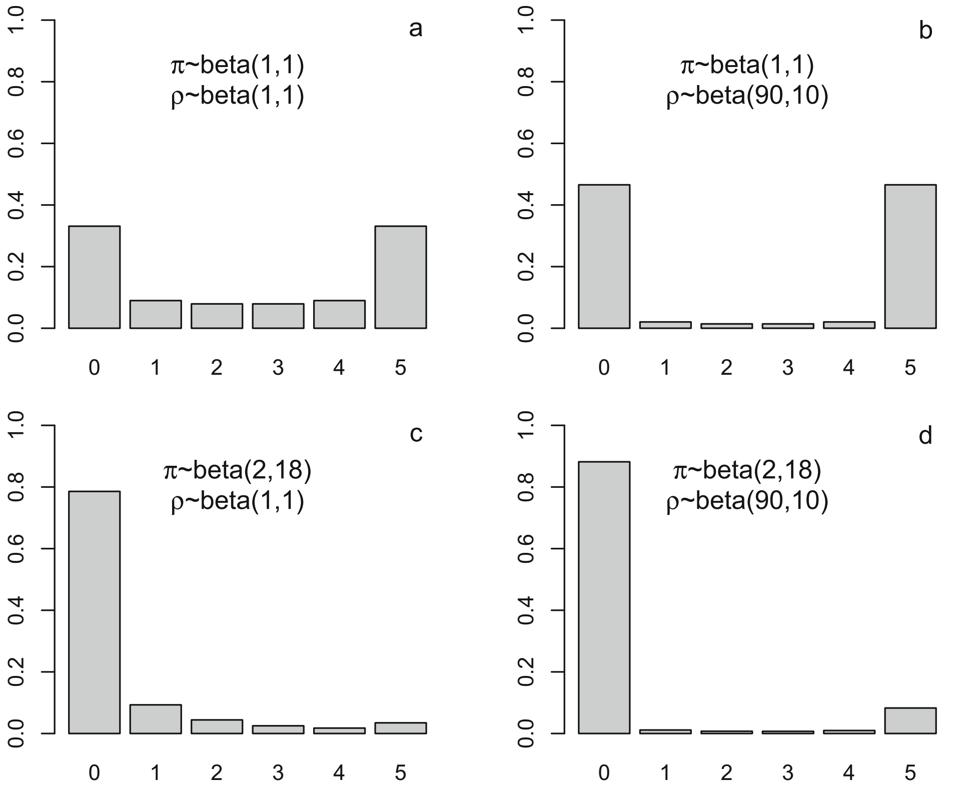

Figure 1. Marginal distributions of the number of positives in populations of size five with independent beta priors on the prevalence and correlation parameters .

sample of size n is selected from the population. Let $\mathbf{n} = (n_1, \ldots, n_c)$ denote the vector of counts selected from the clusters and let $P(\mathbf{n})$ denote the probability of this configuration of counts. For example, under simple random sampling

$$P(\mathbf{n}) = \binom{N}{n}^{-1} \prod_{i=1}^{c} \binom{N_i}{n_i}, \quad \mathbf{n} \in S, \tag{7}$$

where $S = \{\mathbf{n}|0 \leq n_i \leq N_i, i = 1, \ldots, c, \sum_{i=1}^{c} n_i = n\}$. However, in general, all that is required is for the conditional probability $P(\mathbf{n})$, of any sample configuration, given the cluster sizes, to have a known form. For example, in situations in which the clusters can be identified in advance, then two-stage cluster sampling, or a stratified design could be used. Let Y_{is} denote the number of infected animals selected from cluster i. Then, $Y_{1s}, \ldots, Y_{cs}$ are conditionally independent given $(n_1, \ldots, n_c)$ and (π, ρ), with $Y_{is}|n_i, (\pi, \rho) \sim \text{BetaB}(n_i, \pi, \rho)$ for $i = 1, \ldots, c$. In particular, given $\mathbf{n}$, the posterior distribution of (π, ρ) if $Y_s = \sum_{i=1}^{c} Y_{is} = 0$, that is, if the sample is all negative, is

$$p_c(\pi, \rho|\mathbf{n}, Y_s = 0) \propto \left\{ \prod_{i=1}^{c} \text{BetaB}(0; n_i, \pi, \rho) \right\} \text{Beta}(\pi; \alpha_\pi, \beta_\pi)\text{Beta}(\rho; \alpha_\rho, \beta_\rho), \tag{8}$$

with modifications to the bracketed product terms of the form (6) if the sensitivity is less than one. The predictive distribution of $\mathbf{Y}_{-s} = (Y_{1-s}, \ldots, Y_{c-s})$ is therefore

$$P(\mathbf{Y}_{-s} = \mathbf{y}_{-s} | Y_s = 0) = \sum_{\mathbf{n} \in S} \int_0^1 \int_0^1 P(\mathbf{n}) p_c(\pi, \rho | \mathbf{n}, Y_s = 0) \prod_{i=1}^{c} \left\{ \text{BetaB}(y_{i-s}; N_i - n_i, \pi_{0,n_i}, \rho_{n_i}) \right\} d\pi \, d\rho, \tag{9}$$

where $\pi_{0,n_i} = \alpha/(\alpha + n_i + \beta)$ and $\rho_{n_i} = 1/(1 + \alpha + n_i + \beta)$. Equation (9) assumes that the actual configuration of the sample, $\mathbf{n}$, is unobserved and so it is necessary to sum over the set S of all possible configurations. This will always be the case for determining the sample size required for a desired detection probability, for example, where the sample is hypothetical (as opposed to observed). In this setting exact calculation of the posterior probabilities in (9) is infeasible. However, a Monte Carlo approach is possible as follows: a) simulate cluster sample sizes, $\mathbf{n} = (n_1, \ldots, n_c)$, from the sampling probability model P; b) for a given sample configuration, simulate (π, ρ) from the conditional posterior (8); and c) for given $\mathbf{n}$ and (π, ρ) simulate independent beta-binomials, $Y_{i-s} \sim \text{BetaB}(N_i - n_i, \pi_{0,n_i}, \rho_{n_i})$, $i = 1, \ldots, c$.

Simulating from the conditional posterior (8) is accomplished by rejection sampling using the product of priors for π and ρ as the proposal distribution. (A justification for this choice is given later in Sect. 6.) Specifically, let $f(\pi, \rho)$ denote the product of beta-binomials on the right side of Eq. (8) and let $M = \sup_{\pi,\rho} f(\pi, \rho)$. Then

1. Generate three independent variables, $u \sim \text{Beta}(\alpha_\pi, \beta_\pi)$, $v \sim \text{Beta}(\alpha_\rho, \beta_\rho)$ and $w \sim \text{Beta}(1, 1)$.
2. Accept (u, v) as a draw from the conditional posterior if $w < f(u, v)/M$. Otherwise, repeat step 1.

As noted above, sampling schemes other than simple random sampling can be accommodated. For example, a sampling scheme in which a subset of the clusters (e.g., social groups) is over-sampled relative to the remaining clusters (high-harvest sampling, Belsare et al. 2020) can be modeled as a product of independent multivariate hypergeometric distributions. Another modification that is easily accommodated is to allow the cluster sizes to be random with a known probability distribution. This only requires part a) of the Monte Carlo simulation algorithm to be modified.

A complication that arises in the case of positive tests ($Y_s > 0$) is that, if the clusters the positives came from are unknown, then all possible allocations of the positives to the clusters need to be considered. On the other hand, if each sampled animal can be clearly identified with a particular cluster, perhaps because the clusters are distinct locations that are isolated from one another, then it is straightforward to modify (8) and (9) to allow for positive tests. In fact, in this case the summation in (9) is not required and the posterior predictive distribution can, in principle, be calculated via numerical integration. Alternatively, a Monte Carlo approach involving only steps (b) and (c) could be used. Another comment is that, if the sample sizes are negligible relative to the cluster sizes, the sampling distribution,

P, can be replaced by its limiting version; for example, a multinomial distribution in the case of multivariate hypergeometric sampling. Similarly, it is shown in Appendix A.2 that, if $Y \sim \mathrm{BetaB}(N, \pi, \rho)$, then Y/N converges to a beta variable with mean π and variance $\pi(1-\pi)\rho$ as N diverges to infinity. Finally, we note that the clustering model described here includes the situation in which part of the population consists of independent singletons, such as the white-tailed deer population described in the case study, because the beta-binomial distribution (1) reduces to a Bernoulli distribution with success probability $\alpha/(\alpha+\beta)$ when $N = 1$.

6. RESULTS

In order to investigate the sensitivity of the posterior in (8) to the choice of prior we consider a sample of size $n = 100$ drawn from a population of $N = 1000$ consisting of $c = 200$ clusters each of size 5. Suppose that the sample was drawn from 85 clusters, 3 from 1 cluster, 2 from 13 clusters, and 1 from 71 clusters. We consider two outcome scenarios: (i) no disease is detected or $Y_s = 0$; and (ii) $Y_s = 15$ positive tests, 11 associated with singletons and 4 from two of the cluster samples of size 2. In each outcome scenario three prior specifications (each a product of betas) is considered: (I) Beta(1, 1) × Beta(1, 1); (II) Beta(1, 9) × Beta(9, 1); and (III) Beta(1, 19) × Beta(5, 1). Tuyl et al. (2008) discuss examples in which there are no successes in a binomial experiment and argue that the Beta(1, 1) Bayes-Laplace (uniform) prior is appropriate for that setting. However, the two other priors chosen for π can be justified on the grounds that we are interested in detecting and determining the prevalence of rare diseases and so priors that give more weight to values near zero are justified and are likely to be closer to the marginal posterior than a uniform prior. Table 2 tabulates the means and standard deviations of the priors and posteriors for all six combinations of outcome and prior specification. Summarizing the posteriors in terms of their means and standard deviations is justified by the fact that the marginal posteriors resemble beta distributions with the same first two moments; see Fig. 2 for illustrations of this claim. Furthermore, for each of the six settings we generated 1000 draws from the joint posterior and tested the null hypothesis of independence using Hoeffding's D-test (Hoeffding 1948). In every case the test failed to reject independence indicating, given the large sample size, that the posteriors for π and ρ are almost independent.

The take-away from Table 2 is that the posteriors for π are relatively insensitive to the choice of prior. We used a Beta(1, 9) prior for π in the computations described below because it results in a more efficient rejection sampler than a uniform prior. On the other hand, the posteriors for ρ are very sensitive to the choice of prior. In particular, when no disease is detected, the priors and posteriors are almost indistinguishable. Thus, in practice, the choice of prior for the correlation parameter must be based on data from other sources and/or biological considerations concerning the disease and species being investigated.

Figure 3 compares cumulative distributions for the number of diseased individuals estimated under two different sampling scenarios and two different correlation structures for populations consisting of 200 clusters with sizes drawn from a discrete distribution on the integers between 3 and 9 to reflect the sizes of female and fawn social groups of white-tailed

Table 2. Comparisons of the means and standard deviations (in parentheses) between prior and posterior densities under senario (i) (no disease detected) and scenario (ii) (disease detected)

Prior distributions	Probability, π		Correlation, ρ	
$\pi \times \rho$	Prior	Posterior	Prior	Posterior
No disease detected				
Beta(1, 1) × Beta(1, 1)	0.500 (0.289)	0.0106 (0.0106)	0.500 (0.289)	0.513 (0.289)
Beta(1, 9) × Beta(9, 1)	0.100 (0.090)	0.0104 (0.0103)	0.900 (0.090)	0.901 (0.090)
Beta(1, 19) × Beta(5, 1)	0.050 (0.047)	0.0093 (0.0092)	0.833 (0.141)	0.836 (0.139)
Disease detected				
Beta(1, 1) × Beta(1, 1)	0.500 (0.289)	0.1578 (0.0386)	0.500 (0.289)	0.794 (0.179)
Beta(1, 9) × Beta(9, 1)	0.100 (0.090)	0.1463 (0.0360)	0.900 (0.090)	0.925 (0.070)
Beta(1, 19) × Beta(5, 1)	0.050 (0.047)	0.1321 (0.0328)	0.833 (0.141)	0.889 (0.101)

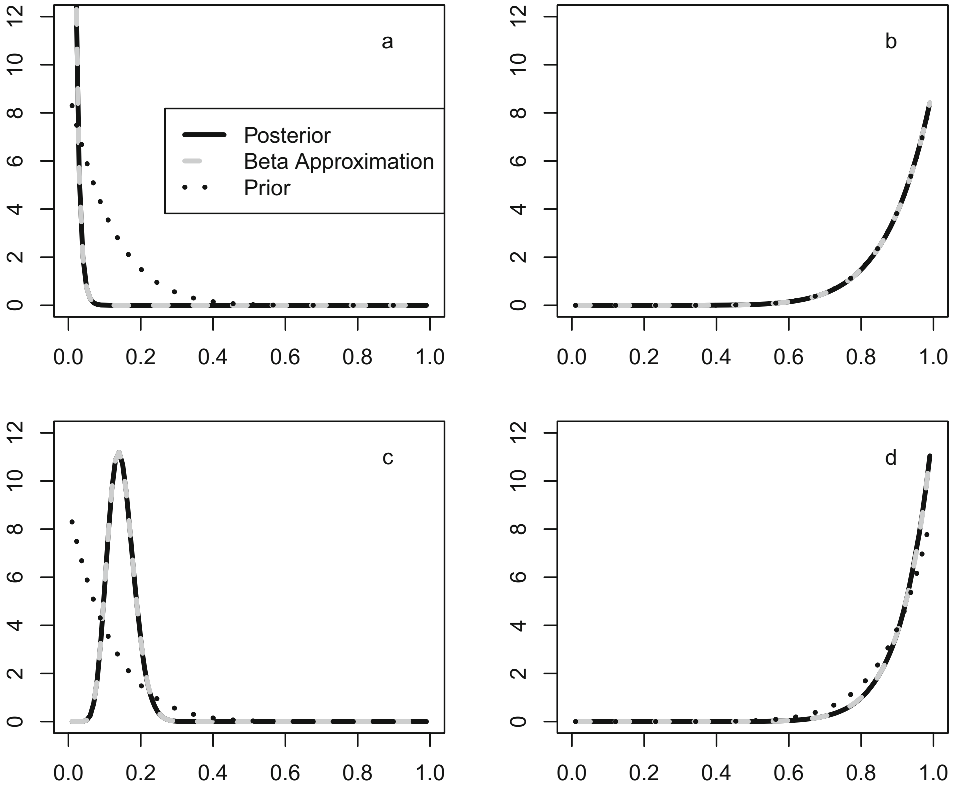

Figure 2. Priors and posterior predictive densities for π (**a**, **c**) and ρ (**b**, **d**). In each case a beta approximation with matching mean and standard deviation is overlaid on the posterior density. Panels 2a and 2b concern scenario (i) (no positive tests) with beta(1, 9) and beta(9, 1) priors for π and ρ respectively. Panels 2c and 2d concern scenario (ii) (15 positive tests) with the same priors. The y-axis in 2a is truncated at 12 .

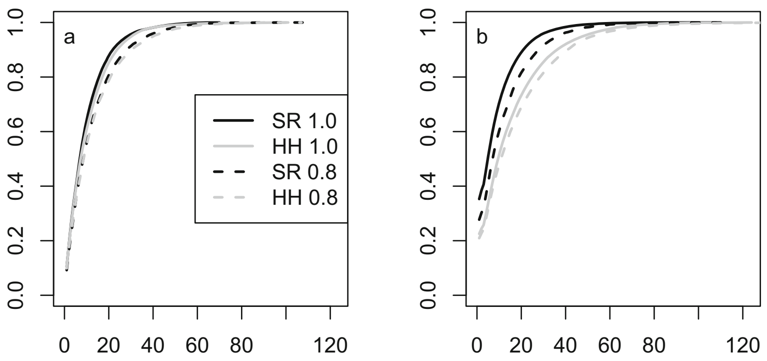

Figure 3. Cumulative predictive distributions of the number of diseased individuals in a population of size 1000 consisting of 200 clusters of random sizes between 3 and 9, based on an all negative sample of size 100. Panels 3a and 3b concern situation in which the within cluster correlation is low and high respectively. The four distributions in each panel correspond to two sampling schemes (SR and HH) and two specificity values (1.0 and 0.8).

deer (i.e., our case study). Specifically, the cluster sizes were drawn using probabilities of integers 3 through 9 proportional to those of a Poisson distribution with a mean of 5. Figure 3a concerns a situation in which the within cluster correlation is low; specifically, $\rho \sim \text{Beta}(1, 9)$ resulting in an average correlation of 0.1. The two sampling schemes being compared are simple random (SR) and high-harvest (HH), and both concern a situation with an all negative sample of size $n = 100$. In the HH case, we assumed disproportionate harvest from available clusters, with 75 individuals drawn from 30 clusters and 25 from the remaining 170 clusters. Figure 3b concerns a situation in which the within cluster correlation is high; specifically, $\rho \sim \text{Beta}(9, 1)$, so the average correlation is 0.9. A takeaway from this comparison is that SR and HH sampling predict prevalence almost equally well when the correlation is low, a result that is consistent with the sampling scheme being irrelevant if the disease statuses of individuals are all independent binary variables with the same probability of being positive. On the other hand, Fig. 3b compares the two sampling schemes when the within cluster correlation is high. In this case, HH sampling results in a predictive distribution that is stochastically larger than SR indicating the need for larger sample sizes with HH sampling. For example, the estimated probabilities that the prevalence is less than 2% are .878 and .852 for SR and HH sampling, respectively, in the low correlation setting, and 0.892 and 0.733 for SR and HH, respectively, in the high correlation setting. Each predictive distribution was based on simulating 1000 samples and took about 10 s on a desktop computer with a single processor.

In fact, the predictive distribution for the prevalence is stochastically smaller for SR sampling with high within cluster correlation relative to the low correlation setting because the effective population size is smaller. On the other hand, HH sampling results in a stochastically larger predictive distribution for the prevalence when there is high within cluster correlation because it is wasteful to sample multiple individuals from the same clusters, and concentrating sampling in a small percentage of clusters does precisely that.

An alternative way to illustrate differences between SR and HH sampling is by sampling directly from simulated populations with low and high within cluster correlations. A key difference, however, is that this approach is not conditional on a real or hypothetical sample, so there is no ‘learning’ from the data. Suppose that the low correlation populations are constructed by (a) again generating 200 cluster sizes independently from the discrete distribution on the integers 3 through 9 described earlier; (b) generating π and ρ as Beta(1, 99) and Beta(1, 9) independently; and (iii) generating the number of diseased individuals in the clusters independently (given π and ρ) as BetaB(N_i, π, ρ). We repeated this process 10,000 times, and, in each case, a sample of size 100 was drawn both by SR and by HH sampling. As in the Bayesian analysis, HH sampling consisted of drawing 75 individuals from 30 high-harvest clusters and 25 from the remaining 170 clusters. High correlation populations were constructed in the same way except that ρ was generated as Beta(9, 1). Thus, in the low population case the average correlation was 0.1 and in the high population it was 0.9. In both cases, the average prevalence was 0.01.

For each sampling scheme and correlation combination we estimated the detection probability by the proportion of times there was at least one positive in the sample. In the low correlation setting the estimates were 0.501 and 0.483 for SR and HH sampling respectively, indicating little difference between the two sampling schemes in agreement with the (conditional) Bayesian analysis. However, in the high correlation setting the corresponding estimates were 0.455 and 0.355, the lower detection probability with HH sampling indicated the need for larger sample sizes. (The Monte Carlo standard errors for these estimates are all less that 0.005.)

7. DISCUSSION

Wildlife health monitoring is the science of seeking out and evaluating a sample of animals from a free-ranging population to obtain a probabilistic statement about disease prevalence (Heisey et al. 2014). The Bayesian modeling approach described here highlights the key roles played by the sampling scheme and correlation between animals. It leads to similar conclusions as classical frequentist approaches in the case of simple random sampling with no correlation, as described in Cannon (2001), for example, and to complex simulation-based, species-specific, approaches such as Belsare et al. (2020).

Unlike previous literature on the topic, our modeling approach is species agnostic, linking only to the biology of the host through the scale of the clusters, the inferred source of the correlation, and the relevancy of the two to the degree to which the host species mixes on the landscape. It allows practitioners of wildlife health to consider how such structures affect sample sizes requirements for diseases that are heterogeneously distributed across large landscapes. We acknowledge, however, that there is a need for additional refinements to better hone our model to the diagnostic test and host. For example, to incorporate uncertainty about the accuracy of the test we could generalize the model by putting a prior distribution on the sensitivity. Other modifications could include modeling host-specific factors such as sex, age, season, or cluster sizes by developing a beta-binomial generalized linear model that allows both the prevalence and correlation parameters to vary as a function of covariates (see e.g., Yee 2015). Such a model would require multiple independent counts, perhaps from spatially distinct areas within a population, and would potentially be much more informative about the correlation parameter. More generally, one might consider allowing for correlation between clusters by modeling the locations at which animals were sampled as a marked spatial point process (Cressie 1993), with the marks being 0/1 indicators of disease status.

More can be done to hone the model to specific disease systems. For example, in our case study we did not address the epidemiological process required for a deer to incur CWD then transmit the disease to others within its social group, nor the ways CWD can transmit between social groups when clusters overlap in home range or individuals mix differently outside of the fall season. One could further argue that depending on the state of the outbreak when certain animals test positive for CWD, professionals may not bother to validate those results, lending the need for an adaptive handling of sensitivity. Indeed, natural next steps in the development of this model are to explicitly accommodate such complications to better mesh the model with specific challenges faced in practice.

While one can argue the details of system complications *ad infinitum*, such focus thwarts our ability to discover key patterns that span systems of interest now. For example, a key take away from this theoretical work is that (high) positive correlation between the disease statuses of individuals in a population decreases the effective population size. As a consequence, the SR sample size required to estimate prevalence to any desired level of accuracy can actually be lower. Taking this to an extreme, if the population is perfectly correlated, then the individuals are either all disease negative or all positive, and a sample size of one is sufficient to infer disease status at the population scale. However, other (perhaps more realistic) sampling schemes, such as HH sampling (Belsare et al. 2020), have the opposite effect in the presence of within cluster correlation by increasing the probability of sampling multiple times from the same clusters, thereby decreasing the effective sample size.

Sampling wildlife populations to detect diseases of interest is nontrivial and requires coordinated efforts in time and space. It is typically not possible or practical to implement optimal sampling design elements, and host animals along with infected individuals are rarely, if ever, distributed randomly throughout a landscape. While baseline sample sizes to detect wildlife diseases at prescribed prevalence and confidence levels are available (Cannon and Roe 1982), they do not account for the biological realities of how the host and infected animals are distributed in space and how disease spreads between individuals. This work provides a way of addressing these issues using a relatively simple, and computationally tractable, Bayesian model, and provides new insight on how practitioners of wildlife health can more appropriately sample given the shared characteristic of clustering across free-ranging wildlife taxa.

ACKNOWLEDGEMENTS

All authors conceived the idea; JGB, FHH, CSJ, and JG designed methodology; data were collected by the MNDNR and collated by CSJ; JGB and FHH wrote the software; JGB, FHH, and JG analyzed the data; JGB and BJH led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication. Data Availability: Data used in this study can be obtained by contacting the MNDNR. Conflict of interest: JGB, BJH, FHH, CSJ, JG, MSA, and KLS declare no conflict of interest. CET is contracted through Cara Them Consulting, LLC; CIM was contracted through Desert Centered Ecology, LLC at the time of this writing. Funding: The Project was funded by a Multistate Conservation Grant # F23AP00488-00, a program funded from the Wildlife and Sport Fish Restoration Program, and jointly managed by the U.S. Fish and Wildlife Service and the Association of Fish and Wildlife Agencies. The findings and conclusions in this article are those of the authors and do not necessarily represent the views of the U.S. Fish and Wildlife Service.

[Received December 2022. Revised June 2023. Accepted October 2023. Published Online November 2023.]

APPENDIX

A.1

Suppose that $Y \sim \text{BetaB}(N, \alpha, \beta)$. Then

1. $Y_s \sim \text{BetaB}(n, \alpha, \beta)$; and
2. the conditional distribution of Y_{-s} given $Y_s = y_s$ is given by

$$P(Y_{-s} = y_{-s} | Y_s = y_s; \alpha, \beta) = \binom{N-n}{y_{-s}} \frac{Be(y+\alpha, N-y+\beta)}{Be(y_s+\alpha, n-y_s+\beta)}.$$

where $y = y_s + y_{-s}$. That is, $Y_{-s} | Y_s, \alpha, \beta \sim \text{BetaB}(N-n, y_s+\alpha, n-y_s+\beta)$.

Proof. Recall that, if $Y|P \sim B(N, P)$ and $P \sim \text{Beta}(\alpha, \beta)$, then $Y \sim \text{BetaB}(N, \alpha, \beta)$. But, if $Y|P \sim B(N, P)$ then, given P, Y_s and Y_{-s} are conditionally independent binomials, with $Y_s|P \sim B(n, P)$ and $Y_{-s} \sim B(N-n, P)$. Hence, the marginal distribution of Y_s is $\text{BetaB}(n, \alpha, \beta)$ and the conditional distribution of Y_{-s} given Y_s is the joint

$$\begin{aligned} P(Y_s = y_s, Y_{-s} = y_{-s} | \alpha, \beta) &= \int_0^1 \binom{n}{y_s} p^{y_s} (1-p)^{n-y_s} \binom{N-n}{y_{-s}} p^{y_{-s}} (1-p)^{N-n-y_{-s}} \\ &\quad \frac{p^{\alpha-1}(1-p)^{\beta-1}}{Be(\alpha, \beta)} dp \\ &= \binom{n}{y_s} \binom{N-n}{y_{-s}} \frac{Be(y+\alpha, N-y+\beta)}{Be(\alpha, \beta)} \end{aligned}$$

divided by the beta binomial marginal distribution for Y_s. □

A.2

The moment generating function (mgf) of $Y \sim \text{Beta}(N, \pi, \rho)$ is given by

$$m_Y(t) = {}_2F_1(-N, \alpha, \alpha+\beta, 1-e^t) = 1 + \sum_{k=1}^{N} (-1)^k \binom{N}{k} \frac{(\alpha)_k}{(\alpha+\beta)_k} (1-e^t)^k$$

where α and β are related to π and ρ via the transformation (2) and $(\alpha)_k = \prod_{i=0}^{k-1} (\alpha + i)$ if $k > 0$ (Johnson et al. 2005, Chapter 6). Hence, the moment generating function of Y/N

is $m_Y(t/N)$. Using the approximation, $1 - e^{t/N} \approx -t/N$, for large N we obtain

$$m_Y(t/N) = 1 + \sum_{k=1}^{N} \frac{(\alpha)_k}{(\alpha+\beta)_k} \frac{t^k}{k!} \prod_{i=1}^{k} \frac{N-k+i}{N}$$

which converges to

$$1 + \sum_{k=1}^{N} \frac{(\alpha)_k}{(\alpha+\beta)_k} \frac{t^k}{k!}$$

as $N \to \infty$, the latter being the moment generating function of a Beta(α, β) distribution.

REFERENCES

Belsare A, Gompper M, Keller B, Sumners J, Hanson L, Millspaugh J (2020) Size matters: sample size assessments for chronic wasting disease surveillance using an agent-based modeling framework. MethodsX 7:100953. https://doi.org/10.1016/j.mex.2020.100953

Branscum A, Gardner I, Johnson W (2004) Bayesian modeling of animal- and herd-level prevalences. Prev Vet Med 66:101–102. https://doi.org/10.1016/j.prevetmed.2004.09.009

Cameron A, Baldock F (1998) A new probability formula for surveys to substantiate freedom from disease. Prev Vet Med 34:1–17. https://doi.org/10.1016/s0167-5877(97)00081-0

Cannon R (2001) Sense and sensitivity—designing surveys based on an imperfect test. Prev Vet Med 49:141–163. https://doi.org/10.1016/s0167-5877(01)00184-2

Cannon A, Roe R (1982) Livestock disease surveys—a field manual for veterinarians. Technical report, Australian Government Publishing Service. Canberra

Cochran WG (1977) Sampling techniques. Wiley, New York

Cressie N (1993) Statistics for spatial data. Revised (edn). Wiley, New York

DeVivo MT, Edmunds DR, Kauffman MJ, Schumaker BA, Binfet J, Kreeger TJ, Richards BJ, Schätzl HM, Cornish TE (2017) Endemic chronic wasting disease causes mule deer population decline in Wyoming. PLoS ONE 12:e0186512

Edmunds D, Kauffman M, Schumaker B, Lindzey F, Cook W, Kreeger T, Grogan R, Cornish T (2016) Chronic wasting disease drives population decline of white-tailed deer. PLoS ONE 11:e0161127. https://doi.org/10.1371/journal.pone.0161127

Hawkins RE, Klimstra WD (1970) A preliminary study of the social organization of white-tailed deer. J. Wildl. Manag. 34:407–419

Heisey D, Jennelle C, Russell R, Walsh D (2014) Using auxiliary information to improve wildlife disease surveillance when infected animals are not detected: a bayesian approach. PLoS ONE. https://doi.org/10.1371/journal.pone.0089843

Hoeffding W (1948) A non-parametric test of independence. Ann Math Stat 19:546–557

Johnson W, Su C, Gardner I, Christensen R (2004) Sample size calculations for surveys to substantiate freedom of populations from infectious agents. Biometrics 60:165–171. https://doi.org/10.1111/j.0006-341X.2004.00143.x

Johnson N, Kemp A, Kotz S (2005) Univariate discrete distributions, 3rd edn. Wiley, Hoboken

Joseph L, Gyorkos T, Coupal L (1995) Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. Am J Epidemiol 141:263–272. https://doi.org/10.1093/oxfordjournals.aje.a117428

Martin P, Cameron A, Greiner M (2007) Demonstrating freedom from disease using multiple complex data sources 1: a new methodology based on scenario trees. Prev Vet Med 79:71–97. https://doi.org/10.1016/j.prevetmed.2006.09.008

Messam L, Branscum A, Collins M, Gardner I (2008) Frequentist and Bayesian approaches to prevalence estimation using examples from Johne's disease. Anim Health Res Rev 9:1–23. https://doi.org/10.1017/S1466252307001314

MNDNR (2023) Chronic wasting disease management. https://www.dnr.state.mn.us/cwd/index.html. Accessed: 30 May 2023

Porter WF, Mathews NE, Underwood HB, Sage RW, Behrend DF (1991) Social organization in deer: implications for localized management. Environ Manag 15:809–814

Rosner B (2005) Beta-binomial distribution. In: Encyclopedia of biostatistics. Wiley, ISBN 9780470849071

Ryser-Degiorgis M (2013) Wildlife health investigations: needs, challenges and recommendations. BMC Vet Res 9:223. https://doi.org/10.1186/1746-6148-9-223

Southwick-Associates (2018) Hunting in America: an economic force for conservation

Stallknecht D (2007) Impediments to wildlife disease surveillance, research, and diagnostics. Current Top Microbiol Immunol 315:445–461. https://doi.org/10.1007/978-3-540-70962-6_17

Tuyl F, Gerlach R, Mengersen K (2008) A comparison of Bayes-Laplace, Jeffreys, and other priors: the case of zero events. Am Stat 62(1):40–44

Uehlinger FD, Johnston AC, Bollinger TK, Waldner CL (2016) Systematic review of management strategies to control chronic wasting disease in wild deer populations in north america. BMC Vet Res 12:173

Williams E, Young S (1980) Chronic wasting disease of captive mule deer: a spongiform encephalopathy. J Wildl Dis 16:89–98. https://doi.org/10.7589/0090-3558-16.1.89

Wilson E (1975) Sociobiology: the new synthesis. Belknap Press, Cambridge

Wobeser G (1994) Investigation and management of disease in wild animals. Springer, New York

Wolfe N, Dunavan C, Diamond J (2007) Origins of major human infectious diseases. Nature 447:279–283. https://doi.org/10.1038/nature05775

Yee T (2015) Vector generalized linear and additive models: with an implementation in R. Springer, New York

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.