Detection and prevalence analysis model description

Joshua Hewitt
Utah State University, Department of Wildland Resources

22 October 2024

1 Overview

The detection and prevalence analysis and power study modules use the same underlying model. The model extends a common, Beta-Binomial Bayesian model to account for presence/absence of the process that generates successes (i.e., disease) moderated by the ability to accurately observe successes (i.e., sensitivity and specificity). The Binomial sampling model, extensions thereof, or its underlying assumptions are widespread in the disease detection, freedom, and prevalence estimation literature and applications (cf. Rogan and Gladen, 1978; Martin et al., 2007; Tabak et al., 2019; Podgórski et al., 2020; Habibzadeh et al., 2022; Hewitt et al., 2024a,b).

2 Model description

Let $Z = 1$ denote a wildlife system that has disease, and $Z = 0$ otherwise. Assume there are m distinct sampling populations in the system, each of which has prevalence $p_i \in [0, 1]$ if $Z = 1$. In this context, a sampling population is a group of individuals, any one of whom is equally representative of others and can be sampled randomly. For example, a system may have distinct sampling populations that represent animals collected via roadkill vs. hunter harvest. Sampling populations can also stratify the overall system by additional information, such as age and sex. So, a complex system may have distinct sampling populations for adult male animals collected via roadkill, adult female animals collected via hunter harvest, and other categories. A necessary simplifying assumption for the purposes of the workshop is that the categories are conditionally independent of each other. The simplification is important for helping the analyses run quickly in real time.

Assume n_i samples are uniformly collected from population i, yielding $Y_i \in \{0, \ldots, n_i\}$ positive test results. We assume the ith population's test results are obtained from an assay with sensitivity $\varphi_i \in [0, 1]$ and specificity $\phi_i \in [0, 1]$. The relationship between Z, $p_1, \ldots, p_m$, and $Y_1, \ldots, Y_n$ can be described via the hierarchical model

$$\begin{aligned} Z &\sim \text{Bernoulli}(\pi), \\ p_i | Z = 0 &\sim \text{Bernoulli}(0), \\ p_i | Z = 1 &\sim \text{Beta}(\alpha_i, \beta_i), \\ Y_i | p_i, Z = 0 &\sim \text{Binomial}(n_i, 1 - \phi_i), \\ Y_i | p_i, Z = 1 &\sim \text{Binomial}(n_i, \varphi_i p_i + (1 - \phi_i)(1 - p_i)), \end{aligned} \tag{1}$$

in which $\pi \in [0, 1]$ is the prior probability that the wildlife system has disease, and $\alpha_i > 0$ and $\beta_i > 0$ parameterize the prior distribution for the ith sampling population's prevalence.

Inference for the model (1) relies on the joint posterior distribution. The joint posterior distribution is proportional to the likelihood times the prior, specified via

$$[p_1, \ldots, p_m, Z|Y_1, \ldots, Y_m] \propto [Z|\pi] \prod_{i=1}^{m} [p_i|Z, \alpha_i, \beta_i][Y_i|p_i, Z, \varphi_i, \psi_i, n_i], \tag{2}$$

in which brackets $[\cdot]$ represent probability density or distribution functions, as appropriate. Marginal posterior distributions for (2) can be numerically evaluated via quadrature when $m = 1$, and via standard Markov Chain Monte Carlo (MCMC) methods when $m > 1$.

3 Analysis outputs

3.1 Detection module

We present the posterior probabilities for disease freedom $P(Z = 0|Y_1, \ldots, Y_m)$ and disease presence $P(Z = 1|Y_1, \ldots, Y_m)$. The app output also includes the prior probabilities $P(Z = 0)$ and $P(Z = 1)$ for comparison. We also present an estimate of the upper bound for prevalence if disease is present—i.e., we also present the 95% quantile of the posterior conditional distribution $[p_i|Y_1, \ldots, Y_m, Z = 1]$.

3.2 Prevalence module

We present the posterior mean and 95% highest posterior density interval for each sampling population—i.e., we present common summary statistics for the marginal posterior distributions $[p_i|Y_1, \ldots, Y_m]$ for $i \in \{1, \ldots, m\}$. The app also includes the same summaries of the prior distributions $[p_1], \ldots, [p_m]$ for comparison.

4 Power study outputs

We also conduct a power study to provide sample size recommendations and other evaluations of the analytic assumptions. To simplify the power study for the workshop so that the tools can be interactive, we only provide recommendations for a large, but ultimately limited set of assumptions about a single sampling population—i.e., for scenarios in which $m = 1$.

4.1 Detection module

We present the frequency that detection analyses (Section 3.1) will declare disease freedom—i.e., how often $P(Z = 0|Y_1) > .95$ would occur if the true prevalence were p_1^*. Formally, the frequency is specified via the expectation

$$\mathrm{E}_{Y_1|p_1^*} \left[\mathbb{1} \left\{ P(Z = 0|Y_1) > .95 \right\} |p_1^* \right], \tag{3}$$

which can be evaluated numerically by a) evaluating the posterior event $\mathbb{1}\{P(Z = 0|Y_1) > .95\}$ for all possible values of $Y_1 \in \{0, \ldots, n_1\}$, then b) weighting the posterior events by the frequency that each outcome would occur if the true prevalence were p_1^*. The frequency can also be evaluated via Monte Carlo simulation, by simulating Y_1 many times starting from an assumption that prevalence is p_1^*, but unknown to the "analyst" evaluating the data for each simulated Y_1 value. However, direct evaluation has computational complexity that is $\mathcal{O}(n_i)$ where n_i is the largest sample size explored, so will be substantially faster than Monte Carlo simulation whenever n_i is smaller than the number of Monte Carlo samples needed to provide sufficiently precise numerical approximation.

We also present the frequency that detection analyses declare disease presence—i.e., how often $P(Z = 1|Y_1) > .95$ would occur if the true prevalence were p_1^*. Formally, the frequency is specified via the expectation

$$\mathrm{E}_{Y_1|p_1^*} \left[\mathbb{1} \left\{ P(Z = 1|Y_1) > .95 \right\} |p_1^* \right],$$

which can be evaluated in the same manner as (3).

4.2 Prevalence module

We present the average width of the 95% highest posterior density interval for p_1 when, in reality, true prevalence is p_1^*. Let $I_1|Y_1$ represent the 95% highest posterior density interval for the first sampling population in which $I_1 = (l_1, u_1)$ for $0 \leq l_1 < u_1 \leq 1$. Let $W_1 = u_1 - l_1$ represent the width of I_1. Formally, the average width of the 95% posterior highest density interval is specified via the expectation

$$\mathrm{E}_{Y_1|p_1^*}\left[W_1|Y_1, p_1^*\right],$$

which can be evaluated in a similar manner as (3).

References

Habibzadeh, F., Habibzadeh, P., and Yadollahie, M. (2022). The apparent prevalence, the true prevalence. *Biochemia Medica*, 32(2):020101.

Hewitt, J., Wilson-Henjum, G., Collins, D. T., Linder, T. J., Lenoch, J. B., Heale, J. D., Quintanal, C. A., Pleszewski, R., McBride, D. S., Bowman, A. S., Chandler, J. C., Shriner, S. A., Bevins, S. N., Kohler, D. J., Chipman, R. B., Gosser, A. L., Bergman, D. L., DeLiberto, T. J., and Pepin, K. M. (2024a). Landscape-scale epidemiological dynamics of SARS-CoV-2 in White-tailed deer. *Transboundary and Emerging Diseases*, 7589509.

Hewitt, J., Wilson-Henjum, G., Collins, D. T., Ringenberg, J. M., Quintanal, C. A., Pleszewski, R., Chandler, J. C., DeLiberto, T. J., and Pepin, K. M. (2024b). A method for characterizing disease emergence curves from paired pathogen detection and serology data. *Methods in Ecology and Evolution*, 15:1677–1690.

Martin, P., Cameron, A., and Greiner, M. (2007). Demonstrating freedom from disease using multiple complex data sources: 1: A new methodology based on scenario trees. *Preventive Veterinary Medicine*, 79:71–97.

Podgórski, T., Borowik, T., Łyjak, M., and Woźniakowski, G. (2020). Spatial epidemiology of African swine fever: Host, landscape and anthropogenic drivers of disease occurrence in wild boar. *Preventive Veterinary Medicine*, 177:104691.

Rogan, W. J. and Gladen, B. (1978). Estimating prevalence from the results of a screening test. *American Journal of Epidemiology*, 107(1):71–76.

Tabak, M. A., Pedersen, K., and Miller, R. S. (2019). Detection error influences both temporal seroprevalence predictions and risk factors associations in wildlife disease models. *Ecology and evolution*, 9(18):10404–10414.