Bayesian Beta-Binomial Prevalence Estimation Using an Imperfect Test

Jonathan Baxter
covid@baxters.biz

April 23, 2020

Abstract

Following [2, 3], we give a simple formula for the Bayesian posterior density of a prevalence parameter based on unreliable testing of a population. This problem is of particular importance when the false positive test rate is close to the prevalence in the population being tested. An efficient Monte Carlo algorithm for approximating the posterior density is presented, and applied to estimating the Covid-19 infection rate in Santa Clara county, CA using the data reported in [1]. We show that the true Bayesian posterior places considerably more mass near zero, resulting in a prevalence estimate of 5,000–70,000 infections (median: 42,000) (2.17% (95CI 0.27%–3.63%)), compared to the estimate of 48,000–81,000 infections derived in [1] using the delta method.

1 Introduction

We consider the problem of estimating disease prevalence in a population of interest using an unreliable test: i.e. a test with specificity or sensitivity less than 1. Imperfect sensitivity causes the raw positive rate to underestimate the true population prevalence, while imperfect specificity results in overestimating the population prevalence and underestimating the dispersion in the prevalence estimate.

Following [2, 3], we take a Bayesian approach and model uncertainty in the test characteristics (sensitivity and specificity) as well as the uncertainty due to a finite testing population. We extend [2] by deriving a simple expression for the posterior prevalence probabilty density in the common case of a test that has been validated against a number of known positive and negative subjects. A Monte Carlo algorithm for computing the prevalence posterior is presented, and applied to Covid-19 infection data from Santa Clara county CA [1], where the false positive test calibration rate (0.5%) was close to the measured prevalence (1.5%). The posterior distribution on prevalence in this case acquires a second mode at zero, which results in substantial broadening of the credible interval on prevalence. The appearance of a second mode also explains why local approximation methods such as the delta-method can fail to capture all the posterior variance.

2 Known test performance

Suppose we know the test false-positive rate (1 - specificity) u and sensitivity v. Denote the (unknown) population prevalence by θ. The probability p of a positive test is the probability of a positive test given the subject has the disease, plus the probability of a positive test given the subject is disease-free:

$$p = v\theta + u(1-\theta) = u + \theta(v-u) \tag{1}$$

The probability of k positive tests out of n subjects tested follows a binomial distribution with parameter p:

$$\Pr(k|n,\theta,u,v) = \binom{n}{k} p^k (1-p)^{n-k} \tag{2}$$

By application of Bayes' rule, the distribution of θ is given by:

$$\Pr(\theta|k,n,u,v) = \frac{\Pr(k|n,\theta,u,v)\Pr(\theta)}{\Pr(k|n,u,v)} \tag{3}$$

where $\Pr(\theta)$ is our prior probability on θ and

$$\Pr(k|n,u,v) = \int_0^1 \Pr(k|n,\theta,u,v)\Pr(\theta)\,d\theta \tag{4}$$

Choosing a uniform prior on θ, and applying $d\theta = \frac{dp}{v-u}$:

$$\begin{aligned}
\Pr(k|n,u,v) &= \int_0^1 \binom{n}{k} p^k (1-p)^{n-k}\,d\theta \\
&= \frac{1}{v-u}\binom{n}{k}\int_u^v p^k(1-p)^{n-k}\,dp \\
&= \frac{1}{v-u}\binom{n}{k}\left[\int_0^v p^k(1-p)^{n-k}\,dp - \int_0^u p^k(1-p)^{n-k}\,dp\right] \\
&= \frac{1}{v-u}\binom{n}{k}\left[B(v;k+1,n-k+1) - B(u;k+1,n-k+1)\right] \\
&=: \frac{B(v)-B(u)}{v-u}\binom{n}{k}
\end{aligned} \tag{5}$$

where $B(x;\alpha,\beta) := \int_0^x t^{\alpha-1}(1-t)^{\beta-1}\,dt$ is the incomplete Beta function, and for notational brevity we drop the dependence on k and n from $B(v;k+1,n-k+1)$ and just write $B(v)$.

Substituting (1), (2) and (5) into (3) yields:

$$\Pr(\theta|k,n,u,v) = \frac{v-u}{B(v)-B(u)}[u+\theta(v-u)]^k[1-u-\theta(v-u)]^{n-k} \tag{6}$$

3 Estimated test performance

Equation (6) expresses the distribution over population prevalence θ given known test characteristics u and v. However, the false-positive rate u and sensitivity v are usually themselves estimates based on validation against known positive and negative subjects. Specifically, suppose the test has been validated with k_u false positives out of n_u known negative samples, and k_v true positives out of n_v known positive samples.

Assuming a beta prior on u with parameters α_u, β_u, the posterior density on u given the validation data is proportional to a beta density with parameters $k_u + \alpha_u, n_u - k_u + \beta_u$. Abusing notation for clarity, write $\mathrm{Beta}_u(u)$ for this density and similarly $\mathrm{Beta}_v(v)$ for the corresponding density on v. Let $\mathrm{Beta}_p(u + \theta(v - u))$ denote the density at $u + \theta(v - u)$ of the beta distribution with parameters $k + 1, n - k + 1$.

With this notation, integrating out u, v from (6), we obtain the following expression for the posterior distribution on prevalence θ that accounts for uncertainty in the test characteristics:

$$\Pr(\theta | k, n, k_u, n_u, k_v, n_v) \propto$$
$$\int_0^1 \int_0^v \frac{v - u}{B(v) - B(u)} \mathrm{Beta}_p(u + \theta(v - u)) \, \mathrm{Beta}_u(u) \, \mathrm{Beta}_v(v) \, du \, dv \quad (7)$$

The domain of integration has been restricted to the region $v - u > 0$, reflecting the fact that a test with false-positive rate u in excess of sensitivity v is not a usable test (this can also be thought of as an adjustment of the joint posterior on u and v to capture a dependence between u and v).

To the author's knowledge there is no closed-form expression for the right-hand-side of (7) in terms of hypergeometric or related functions. In the next section we will describe an algorithm for evaluating the integral using Monte Carlo integration.

4 Computing the Posterior Distribution

Let $I(\theta)$ denote the integral on the right-hand-side of (7). Draw N samples u_i, v_i from $\mathrm{Beta}_u(u)$ and $\mathrm{Beta}_v(v)$ (any pairs such that $u_i > v_i$ are rejected and resampled). Then with error $\sim \frac{1}{\sqrt{N}}$,

$$I(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \frac{v_i - u_i}{B(v_i) - B(u_i)} \mathrm{Beta}_p(u_i + \theta(v_i - u_i)) \quad (8)$$

This expression is calculated for a discrete grid of values θ_j and then normalized to generate the posterior density $\Pr(\theta)$. The samples u_i, v_i can be reused for each estimate $I(\theta_j)$, which allows computation of $\frac{v_i - u_i}{B(v_i) - B(u_i)}$ to be performed once and reused.

To avoid numerical problems in the case $B(u) \approx B(v) \approx 1$, observe that $B(u; k+1, n-k+1) = B(k+1, n-k+1) - B(1-u; n-k+1, k+1)$ where $B(k+1, n-k+1)$ is the complete beta function with parameters $k+1, n-k+1$. Thus

$$B(v) - B(u) = B(1 - u; n - k + 1, k + 1) - B(1 - v; n - k + 1, k + 1). \quad (9)$$

The right-hand-side of (9) is the difference of two values close to zero when the left-hand-side is the difference of two values close to 1. Differencing two small numbers has better numerical stability than differencing two numbers that may be indistinguishable from 1 within machine precision.

With this substitution, Algorithm 1 gives pseudocode for computing the full prevalance posterior given the results of an imperfect test.

Algorithm 1 Posterior prevalence probability (PPP) estimation from an imperfect test

Inputs:
$k, n, k_u, n_u, k_v, n_v, \alpha_u, \beta_u, \alpha_v, \beta_v, N, M$
$B(\cdot)$: incomplete beta function with parameters $n-k+1, k+1$
Outputs:
Posterior prevalence probability density p_j at $\frac{j}{M}$, $j = 0 \dots M$

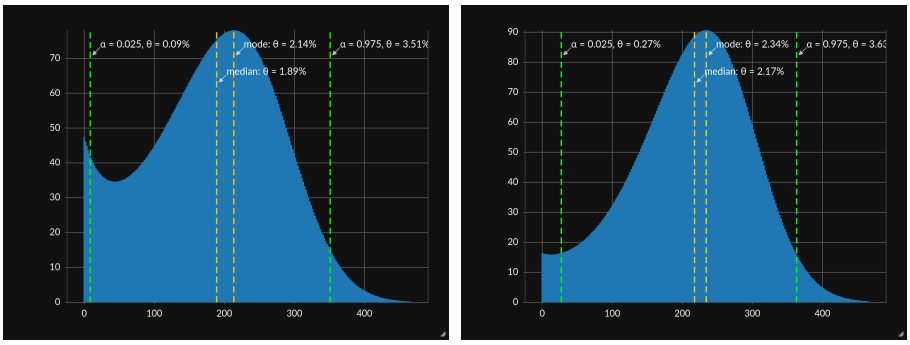
Initialization:
for i=1 to N **do**
  $u_i \leftarrow 0, v_i \leftarrow 0$
  while $u_i \geq v_i$ **do**
    $u_i \leftarrow u \sim \mathrm{Beta}(k_u + \alpha_u, n_u - k_u + \beta_u)$
    $v_i \leftarrow v \sim \mathrm{Beta}(k_v + \alpha_v, n_v - k_v + \beta_v)$
  end while
  $d_i \leftarrow \frac{v_i - u_i}{B(1-u_i) - B(1-v_i)}$
end for

Posterior Density Estimation:
for j=0 to M **do**
  $\theta_j \leftarrow \frac{j}{M}, p_j \leftarrow 0$
  for i=1 to N **do**
    sample $f \sim \mathrm{Beta}(u_i + \theta_j(v_i - u_i); k+1, n-k+1)$
    $p_j \leftarrow p_j + d_i * f$
  end for
  $p_j \leftarrow \frac{p_j}{N}$
end for

Normalization:
$T \leftarrow \frac{1}{M+1} \sum_{j=0}^{M} p_j$
$p_j \leftarrow \frac{p_j}{T}, j = 0, \dots, M$

5 Example

The prevalence of SARS-CoV-2 antibodies in Santa Clara county, CA, was recently measured using an imperfect serological test [1]. Three different calculations were performed based on different estimates of the test's sensitivity and specificity. For brevity, we will focus on their scenario 3 (similar conclusions apply to the other two scenarios).



(a) (b)

Figure 1: Prevalence posterior for Santa Clara county testing data described in [1]. The x-axis is measured in bps, or units of 0.01%. Figure 1a: uniform priors on test false positive rate and sensitivity. Figure 1b: $\mathrm{Beta}(1, 99)$ prior on test false positive.

The relevant parameters for the PPP estimation algorithm are as follows:

- $k = 50$ positive tests out of $n = 3330$ subjects tested.
- $k_u = 2$ false positives out of $n_u = 401$ known negative samples.
- $k_v = 103$ correct positives out of $n_v = 122$ known positive samples.

The authors used the delta method [4] to estimate standard errors for the population prevalence, which accounts for sampling error and propagates the uncertainty in the test sensitivity and specificity. However, the delta method provides only a local approximation to the posterior density, and with small counts this can result in underestimated variance.

The raw positive test count $k = 50$ was also reweighted to account for demographic differences between the test sample and the overall Santa Clara population, yielding a considerably larger population-adjusted count of $k = 94$. The reweighting was applied before the sensitivity/specificity adjustments, which also has potential to underestimate variance in the final result.

After all adjustments, the authors reported a prevalence estimate of 2.75% (95CI 2.01%–3.49%).

In order to avoid potentially biasing our results, we applied the PPP algorithm to the raw counts, and then the population reweighting was applied to the estimated posterior prevalence distribution. To compare with the delta method used in [1], we reran their methodology adjusting first for uncertainty in the test characteristics, and then for population. Omitting the details, we arrive at a prevalence estimate of 2.81% (95CI 1.74% – 3.88%). Observe that while the lower bound of the CI has dropped from 2.01% to 1.74%, it is still well above zero.

Figure 1a shows the prevalence posterior density computed using the PPP algorithm with uniform priors ($\alpha_u = \beta_u = \alpha_v = \beta_v = 1$), a Monte Carlo sample size N of 10,000, and a grid size M of 10,000. The estimated prevalence of Covid-19 in Santa Clara county is 1.89% (95CI 0.09% – 3.51%), with a notably reduced lower bound on the credible interval of 0.09%. This translates to an infected population range of 1,800–68,000, considerably wider than the 38,000–76,000 range derived using the delta method in [1].

The shape of the prevalence posterior near zero helps explain the difference between the two results. The full Bayes approach assigns considerably more mass towards zero, such that the posterior distribution becomes bimodal. This is driven by two factors:

- The uncertainty in the false positive rate u, which is derived from only $k_u = 2$ false positves out of $n_u = 401$ known negative samples (0.5%).
- The underlying infection prevalence rate (estimated at 1.5%) is close to the test false positive rate (0.5%).

The uncertainty in false positive rate is exacerbated by using a uniform prior on u, which is arguably too conservative in this case. Figure 1b shows the posterior generated with $\alpha_u = 1, \beta_u = 99$, which corresponds to a beta prior with mean and standard deviation of 1%. Note that the bimodality is almost eliminated, but the credible interval is still considerably wider than that derived via the delta method: 2.17% (95CI 0.27%–3.63%). This corresponds to an infected range of 5,000–70,000 with median 42,000.

6 Discussion

Reliably estimating infection prevalence with an unreliable diagnostic test is of particular importance during the Covid-19 pandemic, expecially when the infection prevalence is not much greater than the test's false positive rate. Following [2, 3], we derived a simple expression (7) for the posterior prevalence distribution given the results of an unreliable diagnostic test. A Monte Carlo algorithm (Posterior Prevalence Probability or PPP) for efficiently computing the posterior was given. Application of the algorithm to the Santa Clara county, CA Covid-19 test data in [1] generates credible intervals with considerably more mass at zero than the delta method used in the same paper. This is primarily due to the appearance of a second mode in the posterior density at zero, which is not captured by local methods such as the delta method.

Acknowledgements

Thanks to David Joerg and Charlie Graham for helpful comments on an earlier draft of this paper.

References

[1] E. Bendavid, B. Mulaney, N. Sood, S. Shah, E. Ling, R. Bromley-Dulfano, C. Lai, Z. Weissberg, R. Saavedra, J. Tedrow, D. Tversky, A. Bogan, T. Kupiec, D. Eichner, R. Gupta, J. Ioannidis, and J. Bhattacharya. COVID-19 Antibody Seroprevalence in Santa Clara County, California. *medRxiv*, 2020.

[2] P. J. Diggle. Estimating Prevalence Using an Imperfect Test. *Epidemiology Research International*, 2011, 2011.

[3] S. Greenland. Basic Methods for Sensitivity Analysis of Biases. *International Journal of Epidemiology*, 25:1107–1116, 1996.

[4] G. W. Oehlert. A Note on the Delta Method. *The American Statistician*, 46:27–29, 1992.