Minimizing Uncertainty in Prevalence Estimates

Paul N. Patrone, Anthony Kearsley

National Institute of Standards and Technology

Abstract

Estimating prevalence q, the fraction of a population with a specific medical condition, is fundamental to epidemiology. Traditional methods address this task by using a diagnostic test outcome $r \in \Gamma \subset R^n$ to classify and count samples, but such approaches suffer from bias and uncontrolled uncertainty. Recently we showed that unbiased prevalence estimates can be constructed in terms of measures of conditional probability distributions on arbitrary sets $D_\pm$ that partition Γ. In this work, we minimize the variance of this estimator by optimizing the partition itself. In particular, we employ a bathtub principle to recast optimization over the $D_\pm$ in terms of a one-dimensional problem depending only on the total probability mass in these sets. Using symmetry, we show that the resulting objective function is well behaved and can be numerically optimized.

1. Introduction

Prevalence – the fraction q ($0 \leq q \leq 1$) of a population Ω with a medical condition – has been a fundamental quantity of interest in epidemiology for more than 100 years. Nonetheless, many core mathematical issues associated with this concept have only been recently understood [1, 2]. For example, it has long been assumed that q can only be estimated by *classifying* a subset of individuals ω taken at random from Ω. However, in Ref. [1] we demonstrated that this is false: unbiased estimators of prevalence can be constructed from class-agnostic arguments based solely on conditional probability.

The core idea of Ref. [1] is to recognize that individuals or "outcomes" ω in the underlying sample space Ω are often comprised of two parts:[1] an unknown true class; and a diagnostic measurement result that maps onto random variable $r(w) \in \Gamma \subset \mathbb{R}^n$, $n \geq 1$. By the law of total probability, this induces the probability density function (PDF) $Q(r)$ that an individual at random from Ω yields a measurement r. One finds

$$Q(r) = qP(r) + (1-q)N(r), \tag{1}$$

where $P(r)$ and $N(r)$ are *known* PDFs conditioned on the sample being positive or negative. [See Sec. 4 for details of constructing $P(r)$ and $N(r)$.] Integrating Eq. (1) over an arbitrary set $D \in \Gamma$ then yields the exact expression

$$q = (Q_D - N_D)/(P_D - N_D), \tag{2}$$

Email address: paul.patrone@nist.gov (Paul N. Patrone)

[1] In an abuse of notation, we refer to the population and sample space using the same symbol.

where S_D is the probability mass of density S in D.[2] Given M random samples ω_j with measurements $r_j = r(\omega_j)$, one can construct the corresponding estimator

$$q \approx \tilde{q} = \frac{\tilde{Q}_D - N_D}{P_D - N_D}, \qquad \tilde{Q}_D = \frac{1}{M}\sum_{j=1}^{M} \mathbb{I}(r_j \in D), \tag{3}$$

where $\mathbb{I}$ is the indicator function. This $\tilde{q}$ is an unbiased estimator of prevalence [1].

In this manuscript, we address the related problem of *optimal* prevalence estimation by finding the D that minimizes the variance σ^2 of $\tilde{q}$. Mathematically we aim to solve

$$(\sigma^\star)^2 = \min_D \sigma^2 \qquad \text{for} \qquad \sigma^2 = \frac{Q_D(1-Q_D)}{M(P_D - N_D)^2}. \tag{4}$$

The σ^2 reflects properties of the binomial random variable $\tilde{Q}_D$ but is unusual in its dependence on D, which can be deformed arbitrarily. To overcome this challenge, we reduce Eq. (4) to a one-dimensional (1D) problem by maximizing the difference $(P_D - N_D)^2$ for fixed Q_D. The "bathtub principle" is our main analytical tool.

2. Bathtub Principle

Distinct sets D may yield the same Q_D but different values of $P_D - N_D$. This motivates replacing Q_D by a new variable $\hat{Q}$ in Eq. (4), subject to the constraints

$$\hat{Q} = q(P_D - N_D) + N_D, \qquad 0 \le \hat{Q} \le 1. \tag{5}$$

Clearly the collection of sets $\{D^\star\}_{\hat{Q}}$ whose elements maximize $|P_D - N_D|$ for fixed $\hat{Q}$ is an equivalence class. Since, each element in this class yields the same $F(\hat{Q}) = (P_{D^\star} - N_{D^\star})^2$, Eq. (4) is equivalent to the 1D problem

$$\hat{Q}^\star = \operatorname*{argmin}_{\hat{Q}} \frac{\hat{Q}(1-\hat{Q})}{MF(\hat{Q})}. \tag{6}$$

Our main task is therefore to construct $F(\hat{Q})$ by finding the equivalence class $\{D^\star\}_{\hat{Q}}$.

Equation (5) indicates that "loading" points r into D to increase $|P_D - N_D|$ simultaneously increases N_D. We should therefore only include r with the largest (or smallest) ratios $|P(r) - N(r)|/N(r)$. This motivates the sets

$$D_\pm = \{r : \pm q[P(r) - N(r)] > \pm \Delta_\pm N(r)\} \cup S_\pm, \tag{7}$$

$$S_\pm \subset \{r : q[P(r) - N(r)] = \Delta_\pm N(r)\}, \tag{8}$$

where $\Delta_\pm$ are constants and $S_\pm$ are arbitrary subsets of the "boundary set" subject to

$$\hat{Q} = \int_{D_\pm} Q(r)dr. \tag{9}$$

[2] It is necessary to assume that: (i) D has neither zero measure nor unit mass with respect to Q; and (ii) the masses with respect to P and N are not equal. These conditions are trivial to ensure in practice.

Remark 1. The constants $\Delta_\pm$ determines the level sets up to which we include corresponding values of r in $D_\pm$. The sets $S_\pm$ must be chosen and the constants $\Delta_\pm$ determined so as to solve Eq. (9) for a fixed value of $\hat{Q}$.

A proof that the $D_\pm$ are optimal is closely related to the bathtub principle [3]. However, the following lemma is more naturally applicable to the setting of diagnostic classification.

Lemma 1. *Let $0 < q < 1$, and assume every point r has zero measure with respect to $P(r)$ and $N(r)$. For fixed $\hat{Q}$ satisfying $0 < \hat{Q} < 1$, either D_+ or D_- maximizes $F(\hat{Q})$.*

PROOF. First we show that Eq. (9) defines $\Delta_\pm$. By construction, $\Delta_\pm$ can be restricted to $-q \leq \Delta_\pm < \infty$. Consider Δ_+, and let $D(\Delta) = \{r : q[P(r) - N(r)] > \Delta N(r)\}$, where $-\infty < \Delta < \infty$. Define $D(\infty) = \lim_{\Delta\to\infty} D(\Delta) = \{r : P(r) > 0, N(r) = 0\}$. Clearly,

$$\mathcal{Q}(\Delta) = \left[\int_{D(\Delta)} Q(r)dr\right] - \hat{Q} \tag{10}$$

is a monotone decreasing function of Δ. Assume first that $\mathcal{Q}(\Delta)$ crosses zero at a finite Δ. Then $\mathcal{Q}(\Delta)$ is either continuous or discontinuous at this crossing, which yields Δ_+ as the corresponding limit from the right; that is $\Delta_+ = \inf_\Delta\{\Delta : \mathcal{Q}(\Delta) < 0\}$. The S_+ can then be chosen as any subset of $\{q[P(r) - N(r)] = \Delta_+ N(r)\}$ for which $\mathcal{Q}(\Delta_+) + \int_{S+} Q(r)dr = 0$. This extends to the case that $\mathcal{Q}(\Delta) = 0$ is true for more than one Δ. If instead $\mathcal{Q}(\Delta)$ does not vanish for any finite Δ, then Δ_+ is not finite and S_+ is a subset of $D(\infty)$. A similar argument yields existence of Δ_- and S_-.

Let D be any set that differs from D_+ by positive measure not on S_+. Requiring that Eq. (5) also hold for D yields

$$P_{D_+} - N_{D_+} - P_D + N_D = q^{-1}[N_{D/D_+} - N_{D_+/D}], \tag{11}$$

where / is the set difference operator. Together with Eq. (5), Eq. (7), and the restriction that $-q \leq \Delta_+ < \infty$, one finds that $N_{D/D_+} > N_{D_+/D}$, implying $P_{D_+} - N_{D_+} > P_D - N_D$. A similar argument can be used to show that D_- maximizes the difference $N_D - P_D$. These differences are unique, as can be verified by the definition of $S_\pm$. Finally, note that $F(\hat{Q})$ is maximized by $\max_{\hat{Q}}[P_{D_+} - N_{D_+}, N_{D_-} - P_{D_-}]$. □

Lemma 1 demonstrates that for a fixed $\hat{Q}$ satisfying $0 < \hat{Q} < 1$, one of the variances

$$\sigma_+^2(\hat{Q}) = \frac{\hat{Q}(1-\hat{Q})}{M(P_{D_+} - N_{D_+})^2}, \qquad \sigma_-^2(\hat{Q}) = \frac{\hat{Q}(1-\hat{Q})}{M(P_{D_-} - N_{D_-})^2} \tag{12}$$

minimizes σ^2. However, we have yet to uniquely define the objective function in Eq. (6), since it not clear when to use σ_+^2 or σ_-^2. We now prove that both are equivalent.

Lemma 2. *The variances σ_+^2 and σ_-^2 satisfy the symmetry $\sigma_+^2(\hat{Q}) = \sigma_-^2(1-\hat{Q})$. In particular, $\sigma_+^2(\hat{Q}^\star)$ minimizes Eq. (4) if and only if $\sigma_-^2(1-\hat{Q}^\star)$ does.*

PROOF. The numerators of σ_+^2 and σ_-^2 are invariant to $\hat{Q} \to 1 - \hat{Q}$. Also, for any D_+,

$$P_{D_+} - N_{D_+} = N_{\Gamma/D_+} - P_{\Gamma/D_+}. \tag{13}$$

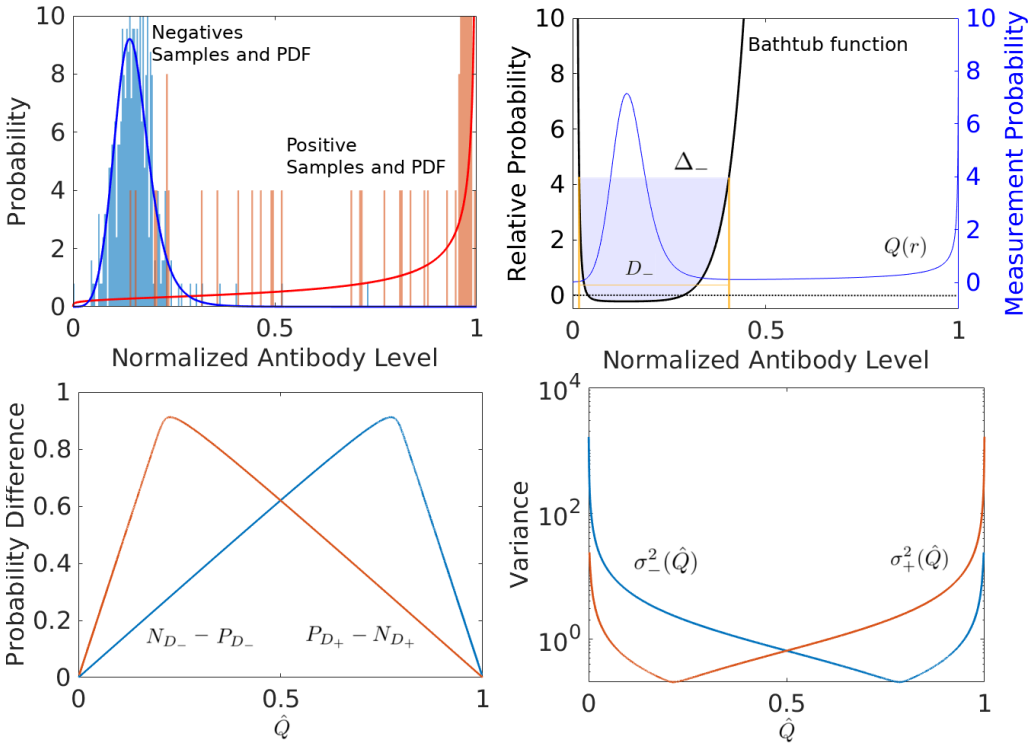

Figure 1: *Top Left*: data and PDFs associated with a SARS-CoV-2 antibody test. Blue and red histograms are relative fractions of negative and positive data from an empirical study. Continuous curves are $N(r)$ and $P(r)$. See main text and supplemental information for modeling details. *Top Right*: "bathtub function" $B(r) = q[P(r) - N(r)]/N(r)$ (black curve) and $Q(r)$ (blue curve) for $q = 30/130$. The domain D_- is constructed by including all r up to the level set $B(r) = \Delta_-$. D_+ is the complement of D_-. *Bottom Left:* The differences $P_{D_+} - N_{D_+}$ and $N_{D_-} - P_{D_-}$ as functions of $\hat{Q}$. These differences are positive on the interior of the domain and reflect the symmetry of Eq. (13). *Bottom Right:* The variances $\sigma_+^2(\hat{Q})$ and $\sigma_-^2(\hat{Q})$ versus $\hat{Q}$ for $M = 1$. Both functions diverge as $\hat{Q} \to 0$ and $\hat{Q} \to 1$, consistent with Lemma 3. Moreover, the variances exhibit the same symmetry as in the bottom left plot.

By Eqs. (7)–(8), the set Γ/D_+ is equal to a set D_- for which $\Delta_- = \Delta_+$. Moreover, since $\hat{Q}$ is the Q-measure of domain D_+, the complement Γ/D_+ has Q-measure $1 - \hat{Q}$. That is, $\Gamma/D_+ \in \{D_-\}_{(1-\hat{Q})}$. In light of Eq. (13), one finds $\sigma_+^2(\hat{Q}) = \sigma_-^2(1 - \hat{Q})$. □

3. Optimal Prevalence Estimation

Lemma 2 proves that we may select σ_+^2 (or σ_-^2) as the objective function in Eq. (6). We now show that this objective function has desirable properties.

Lemma 3. *Assume that: (i) any point r has zero measure with respect to $P(r)$ and $N(r)$; and (ii) there is a set of positive measure with respect to P for which $P(r) > N(r)$. Then on the open domain $0 < \hat{Q} < 1$, the function $\sigma_\pm^2(\hat{Q})$ is continuous and attains a minimum. Moreover, $\sigma_\pm^2 \to \infty$ as $\hat{Q} \to 0$ or $\hat{Q} \to 1$.*

PROOF. By Lemma 2, it is sufficient to only consider $\sigma_+^2(\hat{Q})$. Because any point r has zero measure, Eq. (7) implies the difference $P_{D_+} - N_{D_+}$ is a continuous function of $\hat{Q}$. Thus, $\sigma_+^2(\hat{Q})$ is continuous on $(0,1)$ provided that $P_{D_+} - N_{D_+} > 0$ for every $\hat{Q} \in (0,1)$. To demonstrate this last inequality, assume there exists a $\hat{Q}_\infty \in (0,1)$ for which $P_{D_\infty} - N_{D_\infty} = 0$, where $D_\infty = D_+(\hat{Q}_\infty)$. By assumption (ii) and Eq. (7), the remaining points in Γ/D_∞ must have the property that $P(r) < N(r)$. Integrating over this set yields

$$P_\Gamma - N_\Gamma = P_{D_\infty} - N_{D_\infty} + P_{\Gamma/D_\infty} - N_{\Gamma/D_\infty} = P_{\Gamma/D_\infty} - N_{\Gamma/D_\infty} < 0, \tag{14}$$

which violates the assumption that P and N are probability densities. Thus $P_{D_+} - N_{D_+}$ cannot be zero in the interior of the domain, and $\sigma_+^2(\hat{Q})$ is continuous on $(0,1)$. By definition, $P_{D_+} \leq \hat{Q}$ and $N_{D_+} \leq \hat{Q}$, which implies that $\sigma_+^2(\hat{Q}) \to \infty$ in the limit $\hat{Q} \to 0$. The corresponding result in the limit $\hat{Q} \to 1$ is proved in the same way by considering $\hat{Q} \to 0$ for σ_-^2 and then using Lemma 2.

Existence of the minimum follows from divergence of $\sigma_+^2(\hat{Q})$ at the boundaries and continuity in the interior $(0,1)$. Specifically, for any $0 < \epsilon < 1/2$, σ_+^2 is continuous on $D_\epsilon = [\epsilon, 1-\epsilon]$. By the extreme value theorem, there exists a value $\hat{Q}_\epsilon \in D_\epsilon$ for which $\sigma_+^2(\hat{Q})$ attains a minimum. Clearly there exists an ϵ_0 such that for all $\epsilon < \epsilon_0$, $\sigma_+^2(\hat{Q}_\epsilon)$ is constant, which defines the minimum. □

Remark 2. Assumption (ii) is an important feature of diagnostic tests. It implies: the ability to distinguish populations via the random variable r; and that there is a set of positive measure with respect to N for which $N(r) > P(r)$. The limiting case in which $P(r)$ and $N(r)$ have disjoint supports corresponds to the lowest possible uncertainty $\sigma^2 = \hat{Q}(1-\hat{Q})/M$, since $D_\pm$ can be chosen as the supports of either PDF. Lemma 3 also does not in general establish uniqueness of the minimum. However, uniqueness may not be inherently useful. The domain $0 \leq \hat{Q} \leq 1$ is compact and one-dimensional, suggesting that straightforward sampling can estimate the minimum value of σ^2

4. Validation and Discussion

4.1. Practical Considerations, Limitations, and Open Questions

The analysis herein assumes that $P(r)$ and $N(r)$ are known exactly. In practice, we need to empirically model these PDFs using *training data*, for which the true classes are known. The analysis is then applied to *test data* having the same conditional distributions as the training data, but a prevalence and sample classes that are unknown *a priori*. Several issues must be considered en route.

First, training the PDFs $P(r)$ and $N(r)$ may introduce "model-form errors," depending on how much information about the underlying physics and biology is incorporated. Addressing these effects is necessary to develop reliable models, but such questions are beyond the scope of this work and the topic of a future manuscript; see Ref. [1] for an in-depth discussion.

In a related vein, the optimal prevalence estimates require knowledge of q itself. Since this quantity is often unknown *a priori*, the $D_\pm$ can at best be estimated. A practical algorithm doing so is to: (i) guess a trial domain D based on prior information;

(ii) estimate q via Eq. (3); (iii) estimate $D_\pm$ from this prevalence estimate; and (iv) recompute q and its uncertainty from Eq. (4). While iterating this approach may not yield converging estimates of σ^2, one or two such repetitions should yield reasonable estimates of the minimum uncertainty. Thus, Lemmas 1-3 should be viewed as idealized results that inform best practices for estimating prevalence and uncertainty therein.

4.2. Example Applied to a SARS-CoV-2 Assay

In order to validate our main theoretical results, we consider a SARS-CoV-2 immunoglobulin G (IgG) receptor binding domain (RBD) assay developed in Ref. [4]. The top-left plot of Fig. 1 shows positive and negative training datasets, i.e. collections of r values for which the true underlying class is known. Also shown are $P(r)$ and $N(r)$ modeled on these training datasets. Our model choices are empirical ones that make several simplifying assumptions, primarily for the sake of illustrating notable aspects of $D_\pm$. In short, we assume that $N(r)$ is a Burr distribution and $P(r)$ is a beta distribution [5] and estimate the associated parameters using maximum likelihood estimation. See the supplemental material for more details. The top-right plot of Fig. 1 shows the "bathtub function" $B(r) = q[P(r) - N(r)]/N(r)$ for a prevalence of $q = 30/130$. The figure illustrates that for a given Δ_-, the domain D_- corresponds to the set of all r for which $B(r) < \Delta_-$. The corresponding value of $\hat{Q}$ is given by the integral of $Q(r)$ over D_-. Observe that since D_- does not contain the points $\hat{Q} = 0$ or $\hat{Q} = 1$, the domain $D_+ = \Gamma/D_-$ is not simply connected.

The bottom plots of Fig. 1 illustrate various results derived in Lemma 3. The bottom-left shows the symmetry argument expressed in Eq. (13), as well as the fact that the difference $P_{D_+} - N_{D_+}$ and $N_{D_-} - P_{D_-}$ are positive on $0 < \hat{Q} < 1$. The bottom-right plot illustrates that $\sigma_+^2(\hat{Q}) = \sigma_-^2(1 - \hat{Q})$, as well as the divergence when $\hat{Q} \to 0$ and $\hat{Q} \to 1$. Moreover, the objective function is continuous and bounded from below.

Acknowledgements: This work is a contribution of the National Institutes of Standards and Technology and is therefore not subject to copyright in the United States.

Use of all data deriving from human subjects was approved by the NIST Research Protections Office.

References

[1] P. N. Patrone, A. J. Kearsley, Classification under uncertainty: data analysis for diagnostic antibody testing, Mathematical Medicine and Biology: A Journal of the IMA 38 (3) (2021) 396–416.

[2] L. Böttcher, M. R. D'Orsogna, T. Chou, A statistical model of covid-19 testing in populations: effects of sampling bias and testing errors, Philosophical Transactions of the Royal Society A 380 (2214) (2022) 20210121.

[3] E. Lieb, M. Loss, A. M. Society, Analysis, Crm Proceedings & Lecture Notes, American Mathematical Society, 2001.

[4] T. Liu, J. Hsiung, S. Zhao, J. Kost, D. Sreedhar, C. V. Hanson, K. Olson, D. Keare, S. T. Chang, K. P. Bliden, P. A. Gurbel, U. S. Tantry, J. Roche, C. Press, J. Boggs, J. P. Rodriguez-Soto, J. G. Montoya, M. Tang, H. Dai, Quantification of antibody avidities and accurate detection of sars-cov-2 antibodies in serum and saliva on plasmonic substrates, Nature Biomedical Engineering 4 (12) (2020) 1188–1196.

[5] I. W. Burr, Cumulative Frequency Functions, The Annals of Mathematical Statistics 13 (2) (1942) 215 – 232.