Confidence Intervals for Seroprevalence *

Thomas J. DiCiccio
Department of Social Statistics
Cornell University
tjd9@cornell.edu

David M. Ritzwoller
Graduate School of Business
Stanford University
ritzwoll@stanford.edu

Joseph P. Romano
Departments of Statistics and Economics
Stanford University
romano@stanford.edu

Azeem M. Shaikh
Department of Economics
University of Chicago
amshaikh@uchicago.edu

November 3, 2021

Abstract

This paper concerns the construction of confidence intervals in standard seroprevalence surveys. In particular, we discuss methods for constructing confidence intervals for the proportion of individuals in a population infected with a disease using a sample of antibody test results and measurements of the test's false positive and false negative rates. We begin by documenting erratic behavior in the coverage probabilities of standard Wald and percentile bootstrap intervals when applied to this problem. We then consider two alternative sets of intervals constructed with test inversion. The first set of intervals are approximate, using either asymptotic or bootstrap approximation to the finite-sample distribution of a chosen test statistic. We consider several choices of test statistic, including maximum likelihood estimators and generalized likelihood ratio statistics. We show with simulation that, at empirically relevant parameter values and sample sizes, the coverage probabilities for these intervals are close to their nominal level and are approximately equi-tailed. The second set of intervals are shown to contain the true parameter value with probability at least equal to the nominal level, but can be conservative in finite samples.

Keywords: Confidence Intervals, Novel Coronavirus, Serology Testing, Seroprevalence, Test Inversion

MSC 2020 Subject Classification: Primary 62F03, Secondary 62P10

*We acknowledge funding from the National Science Foundation under the Graduate Research Fellowship Program and under the grants MMS-1949845 and SES-1530661.

1. Introduction

Effective public health policy requires accurate measurement of the spread of infections diseases (Fauci et al., 2020; Peeling et al., 2020). Seroprevalence surveys, in which antibody tests are administered to samples of individuals from populations of interest, are a practical and widely applied strategy for assessing the progression of a pandemic (Krammer and Simon, 2020; Alter and Seder, 2020). However, antibody tests, which detect the presence of viral antibodies in blood samples, are imperfect.[1] Accounting for the variation in the results of seroprevalence surveys induced by this imperfection is important for informative assessment of the uncertainty in measurements of the spread of infectious diseases.

In this paper, we study the construction of confidence intervals in standard seroprevalence surveys. Given the public interest in explicit representations of disease incidence, our objective is to analyze the accuracy of various methods of constructing confidence intervals, so that results from empirical analyses can be reported with statistical precision. We demonstrate that some methods based on test inversion offer advantages relative to more standard confidence interval constructions in terms of the accuracy and validity of their coverage probabilities.

In a standard seroprevalence survey, the proportion of a population that has been infected with a disease is a smooth function of the parameters of three independent binomial trials. Although it may be expected that standard approaches to confidence interval construction are well suited for such a simple parametric problem, in Section 2 we demonstrate with simulation that standard Wald and percentile bootstrap confidence intervals have erratic coverage probabilities at empirically relevant parameter values and sample sizes when applied to this problem.

In fact, as documented in Brown et al. (2001), erratic coverage probabilities for confidence intervals constructed using standard methods surface even in the context of inference on a single binomial parameter. Bootstrap (and other) methods that are typically second-order correct in continuous problems may not achieve this accuracy in discrete problems.[2] Additionally, when a binomial random variable has parameter value near zero or one, even first-order approximations to its limiting distribution are not normal, while many standard methods for constructing confidence

[1] An early systematic review of the accuracy of SARS-CoV-2 antibody tests is given in Deeks et al. (2020), highlighting several methodological limitations. False positive and negative rates for five leading SARS-CoV-2 immunoassays are measured in Ainsworth et al. (2020). Estimates of false positive rates ranged from 0.1% to 1.1%. Estimates of false negative rates ranged from 0.9% to 7.3%.

[2] Usually, claims of second-order correctness are made based on Edgeworth expansions (Hall, 2013). However, in discrete settings, Cramér's condition, a necessary condition for the application of an Edgeworth expansion, fails, and second-order accuracy may not be achievable. For example, atoms in the binomial distribution based on n trials have order $n^{-1/2}$, so expansions to order n^{-1} must account for this discreteness.

intervals rely, either explicitly or implicitly, on a normal approximation holding. As the parameter of interest in a seroprevalence survey is a function of three binomial parameters, inference in this setting is more challenging than for a single binomial proportion.

To address the erratic coverage probabilities in standard confidence interval constructions, we consider several alternative approaches based on test inversion. A test inversion confidence interval for a parameter θ consists of the set of points θ_0 for which the null hypothesis $H(\theta_0) : \theta = \theta_0$ is not rejected. For parameters θ where the corresponding null hypothesis $H(\theta_0)$ is simple, the application of test inversion is straightforward. However, when the corresponding null hypothesis is composite, as in the case of seroprevalence, the application of test inversion is not immediate.

Thus, in Section 3 we explore the general problem of test inversion for parameters whose corresponding null hypotheses are composite. We consider both methods based on asymptotic or bootstrap approximation and methods with finite-sample guarantees. In the later case, a maximization of p-values over a nuisance parameter space is required, as in Berger and Boos (1994) and Silvapulle (1996). In practice, this maximization is carried out over a discrete grid. We provide a refinement to such an approximation that maintains the finite-sample coverage requirement. We take particular care in requiring that the confidence intervals that we develop behave well at both endpoints; that is, we require that they are equi-tailed.[3]

In Section 4, we apply these approaches to construct confidence intervals for seroprevalence. We consider several choices of test statistic, including maximum likelihood estimators and generalized likelihood ratio statistics. We demonstrate with simulation that the intervals based on asymptotic or bootstrap approximation have coverage probabilities that, at empirically relevant parameter values and sample sizes, are close to, but potentially below, the nominal level and are approximately equi-tailed. By contrast, the finite-sample valid construction results in longer intervals on average, but always have coverage probabilities that satisfy the coverage requirement.

We contextualize our analysis with data used to estimate seroprevalence at early stages of the 2019 SARS-Cov-2 pandemic. In particular, as a running example, we measure coverage probabilities and average interval lengths for each of the methods we consider at sample sizes and parameter values close to the estimates and sample sizes of Bendavid et al. (2020a) – a preprint posted on medRxiv on April 11$^{\text{th}}$, 2020.[4] This preprint estimates that the number of coronavirus cases in

[3] A $1-\alpha$ confidence interval is equi-tailed if the probabilities that the parameter exceeds the upper endpoint or is below the lower endpoint of the interval are both near or below $\alpha/2$. That is, an equi-tailed $1-\alpha$ confidence interval should be given by the set of points satisfying both an upper and a lower confidence bound, each at level $1-\alpha/2$.

[4] This preprint has subsequently been published as Bendavid et al. (2021).

Santa Clara County, California on April 3 - 4, 2020 was more than fifty times larger than the number of officially diagnosed cases, and as a result, received widespread coverage in the popular and scientific press (Kolata, 2020; Mallapaty, 2020). The methods and design of this study – including the reported confidence intervals – were questioned by many researchers (Eisen and Tibshirani, 2020), prompting the release of a revised draft on April 27th, 2020, which we refer to as Bendavid et al. (2020b), that integrated additional data.[5] Our analysis highlights statistical challenges in seroprevalence surveys at early stages of the spread of infectious diseases, when disease incidence is low and close to uncertain error rates of new diagnostic technologies.

This paper contributes to the literatures on inference in seroprevalence surveys (Rogan and Gladen, 1978; Hui and Walter, 1980; Walter and Irwig, 1988); see Jewell (2004) for a general introduction to epidemiological statistics. More broadly, we contribute to the large literature on test inversion. The classical duality between tests and confidence intervals is discussed in Chapter 3 of Lehmann and Romano (2005). Bootstrap approaches to confidence construction based on estimating nuisance parameters are developed in Efron (1981), DiCiccio and Romano (1990), and Carpenter (1999). Conservative approaches to confidence interval construction that maximize p-values over an appropriate nuisance parameter space are considered in Berger and Boos (1994) and Silvapulle (1996). For the problem considered in this paper, Toulis (2020) uses test inversion based on a particular choice of test statistic, though the resulting confidence interval is based on projection. Cai et al. (2020) is more closely related to one of the approaches we consider, and we discuss some important differences in Section 4.[6] Gelman and Carpenter (2020) take a Bayesian approach to the problem studied in this paper, and give a complementary analysis of uncertainty quantification in Bendavid et al. (2020a,b).

2. Standard Interval Constructions in Seroprevalence Surveys

A standard seroprevalence survey entails the collection of antibody test results from three samples of individuals of sizes n_1, n_2, and n_3, with $n = (n_1, n_2, n_3)^\top$. The first sample is selected at random from the population under study. All individuals in the second sample have not had the

[5] See Gelman (2020), Fithian (2020b), and Bennett and Steyvers (2020) for further discussion and analysis of Bendavid et al. (2020b). In particular, these articles highlight issues and propose alternative approaches for combining data measuring false positive rates from different samples.

[6] We became aware of Cai et al. (2020), which was posted on arXiv on November 29th, 2020, late in the preparation of this paper, on March 3rd, 2021.

disease of interest and all individuals in the third sample have had the disease of interest.[7] We let $X = (X_1, X_2, X_3)^\top$ denote the number of positive antibody test results in the corresponding samples. It is assumed that each X_i has a binomial distribution with success probability p_i and is independent of the other samples.[8] The quantities $1 - p_2$ and p_3 are referred to as the specificity and sensitivity of the test, respectively. We assume that the test has diagnostic value in the sense that $p_2 < p_3$, and so p_1 necessarily satisfies $p_2 \leq p_1 \leq p_3$.[9] Thus, the parameter $p = (p_1, p_2, p_3)^\top$ exists in the parameter space

$$\Omega = \left\{ p \in [0,1]^3 : p_2 \leq p_1 \leq p_3, p_2 < p_3 \right\}.$$

We consider confidence intervals for the probability π that an individual randomly selected from the population under study has had the disease. By the law of total of probability, $p_1 = p_2(1-\pi) + p_3\pi$, and so

$$\pi = \pi(p) = (p_1 - p_2) / (p_3 - p_2) \ . \tag{1}$$

We refer to π as seroprevalence. A natural estimate of π is given by $\check{\pi}_n = \pi(\check{p}_n)$, where $\check{p}_n = (\check{p}_{n,1}, \check{p}_{n,2}, \check{p}_{n,3})^\top$ and $\check{p}_{n,i} = X_i/n_i$ is the usual empirical frequency for group i. We let the maximum likelihood estimator (MLE) of p for the model $p \in \Omega$ be denoted by $\hat{p}_n$, where $\hat{p}_n = (\hat{p}_{n,1}, \hat{p}_{n,2}, \hat{p}_{n,3})^\top$.[10] Accordingly, the MLE of π for the model $p \in \Omega$ is given by $\hat{\pi}_n = \pi(\hat{p}_n)$.

The most obvious approach to constructing confidence intervals for π is to approximate the finite-sample distribution of $\hat{\pi}_n$ with its limiting normal distribution. The variance of the normal limiting distribution of $\hat{\pi}_n$ is given by

$$V_{\hat{\pi}_n}(p) = \frac{1}{(p_3 - p_2)^2}\sigma_1^2(p_1) + \frac{(p_1 - p_3)^2}{(p_3 - p_2)^4}\sigma_2^2(p_2) + \frac{(p_2 - p_1)^2}{(p_3 - p_2)^4}\sigma_3^2(p_3), \tag{2}$$

where $\sigma_i^2(p_i) = p_i(1-p_i)/n_i$. This leads to the standard Wald or delta method confidence interval $\left(\hat{\pi}_n \pm z_{1-\alpha/2}\sqrt{V_{\hat{\pi}_n}(\hat{p}_n)}\right)$, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal cumulative distribution function $\Phi(\cdot)$. This construction was used in Bendavid et al. (2020a).

[7] For example, in Bendavid et al. (2020a), the second sample was composed of blood samples taken before the COVID-19 epidemic and the third sample was composed of blood samples taken from patients who had recovered from confirmed cases of COVID-19.

[8] We assume the sample sizes are small relative to population size so that the difference between sampling with and without replacement is negligible.

[9] Note that some of the methods developed in Section 4 will not require $p_2 < p_3$.

[10] Typically, $\check{p}_n$ and $\hat{p}_n$ agree, with the exception occurring if $\check{p}_n \notin \Omega$.

As outlined in the introduction, the Wald interval may perform poorly in finite-samples due to discreteness of the data or the proximity of parameter values to the boundaries of their spaces. To address some of these issues, Bendavid et al. (2020b) apply the percentile bootstrap confidence interval developed in Efron (1981). A refinement of this interval construction, called the BC_a interval (Efron, 1987), is also applicable to this problem, with bias and acceleration constants estimated with the formula given in Efron (1987) and DiCiccio and Romano (1995).

The Wald and bootstrap intervals are approximate. In contrast, it may be desirable to construct intervals that ensure coverage of at least $1-\alpha$ in finite samples, particularly if there are concerns that the finite-sample distribution of $\hat{\pi}_n$ is not well-approximated by a normal distribution. A simple, but crude, approach to constructing finite-sample valid confidence intervals is projection. In particular, suppose that $R_{1-\alpha}$ is a joint confidence region for p of nominal level $1-\alpha$. The projection method simply constructs the confidence interval $I_{1-\alpha} = \{\pi(p) : p \in R_{1-\alpha}\}$. The chance that $\pi \in I_{1-\alpha}$ is bounded below by the chance that $p \in R_{1-\alpha}$. Thus, if $R_{1-\alpha}$ has a guaranteed coverage of $1-\alpha$, then so does $I_{1-\alpha}$. For example, one possible choice of joint confidence region is the rectangle $R_{1-\alpha} = \prod_{i=1}^{3} I_{i,1-\gamma}$, where $I_{j,1-\gamma}$ is a nominal $1-\gamma$ confidence interval for p_j and γ is taken to satisfy $(1-\gamma)^3 = 1-\alpha$.[11] In this case, the computational cost of the projection interval $I_{1-\alpha}$ is minimal, as $\pi(p)$ is monotone increasing in p_1 and monotone decreasing in each of p_2 and p_3 as p varies on the parameter space Ω. Projection intervals are easy to implement, but are generally wide and conservative, in that the true coverage is often larger than the nominal level.

To assess the finite-sample performance of the delta method, bootstrap, and projection confidence intervals, we estimate their coverage probabilities and average lengths at parameterizations close to the sample size and estimates of Bendavid et al. (2020a). In this study, $n_1 = 3300$ participants were recruited for serologic testing for SARS-CoV-2 antibodies. The total number of positive tests was $X_1 = 50$. The authors use $n_2 = 401$ pre-COVID era blood samples to measure the specificity of their test, of which only $X_2 = 2$ samples tested positive. Similarly, the authors use $n_3 = 122$ blood samples from confirmed COVID-19 patients, of which $X_3 = 103$ samples

[11] In our implementation, we apply the standard Clopper and Pearson (1934) confidence intervals for p_j, as they have guaranteed coverage in finite-samples. Other choices exist, however, and in particular the intervals recommended in Brown et al. (2001) may perform well. Alternatively, the region $R_{1-\alpha}$ can be constructed by inverting likelihood ratio tests, but would incur a significantly larger computational cost. A related approach is developed in Toulis (2020).

	Interval	Ave. Length	Ave. Length vs. Delta Method	Coverage
Delta Method	[0.003,0.022]	0.0185	1.000	0.904
Percentile Bootstrap	[0.001,0.021]	0.0186	1.005	0.895
BC_α Bootstrap	[0.001,0.020]	0.0191	1.028	0.895
Projection	[0.001,0.028]	0.0270	1.4578	1.000

Table 1: Average Interval Length and Coverage of Nominal 95% Confidence Intervals for π

Notes: Table 1 reports the delta method, percentile bootstrap, BC_α bootstrap, and projection confidence intervals, at nominal level 95% computed on data from Bendavid et al. (2020a). Estimates of the average length and coverage for these intervals at sample sizes n and estimated values $\hat{p}_n$ from this study are also displayed. Estimates of average length and coverage are taken over 100,000 bootstrap replicates of X at the sample size n and the estimated parameters $\hat{p}_n$ from this study.

tested positive.[12] The realization of the MLE of π for these data is $\hat{\pi}_n = 0.012$.[13] Table 1 reports the nominal 95% confidence intervals constructed with the standard approaches discussed above.

For each parameter (e.g., p_1), we simulate replicates of X at each value of a grid around the estimated value of the parameter, holding the other five parameters fixed at their estimated values (e.g., $\hat{p}_{n,2}$, $\hat{p}_{n,3}$, n_1, n_2, n_3). For each method at each combination of parameter values and sample sizes, we compute the proportion of replicates for which the true value of π (i.e., the value of π associated with the parameterization) is below, contained in, or above the corresponding confidence interval with nominal coverage probability $\alpha = 0.05$.

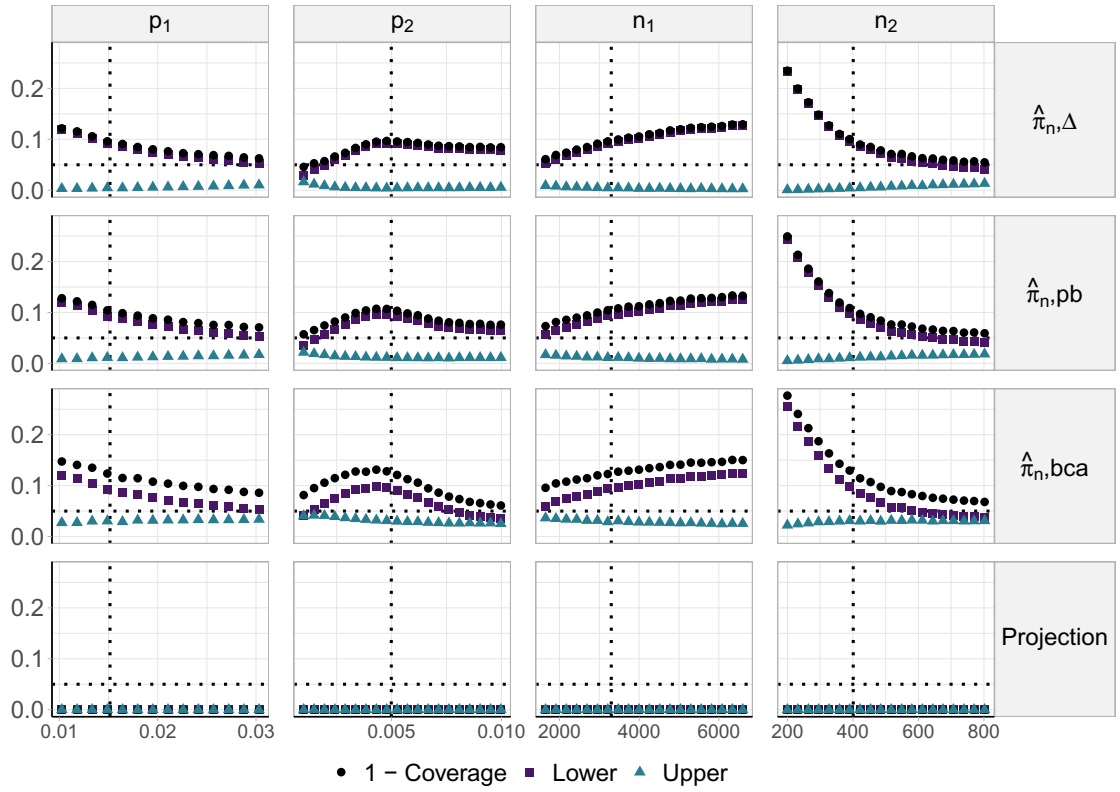
Figure 1 displays the results of this Monte Carlo experiment for each interval construction at parameter values around $\hat{p}_{n,1}$, $\hat{p}_{n,2}$, n_1, and n_2.[14] The black dots display one minus the proportion of replicates for which the realized confidence interval contains the true value of π, i.e., one minus the estimated coverage of the confidence interval. Additionally, Table 1 reports estimates of the coverage and average length of each interval taken over 100,000 bootstrap replicates at the sample

[12]The specificity and sensitivity samples combine data provided by the test manufacturer and additional tests run at the Stanford. We refer the reader to the statistical appendices of Bendavid et al. (2020a,b) for further details. In Bendavid et al. (2020b) it was revealed that there was an error in the recording of the sensitivity sample, i.e., that there were two fewer positive tests than reported. We adhere to the data as reported in Bendavid et al. (2020a).

[13]Bendavid et al. (2020a) report an alternative estimate of seroprevalence in which the demographics of their sample are weighted to match the overall demographics of Santa Clara County. We briefly discuss the application of the general methods developed in Section 3 to this setting in Section 5, and view further consideration as a useful extension. Gelman and Carpenter (2020) give a Bayesian approach that accommodates sample weights. In contemporaneous work, Cai et al. (2020) also address the case where are samples are reweighted according to population characteristics.

[14]There is little variation in the coverage probabilities for parameter values around p_3 and n_3, so we omit the results of this experiment for the sake of clarity.

Figure 1: Coverage Performance for Standard Interval Constructions



Notes: Figure 1 displays estimates of the coverage probabilities of the delta method $(\hat{\pi}_n, \Delta)$, percentile bootstrap $(\hat{\pi}_n, \text{pb})$, BC_α bootstrap $(\hat{\pi}_n, \text{bca})$, and projection intervals at parameter values close to the estimate $\hat{p}_n$ and sample size n of Bendavid et al. (2020a) as specified in Section 2. The nominal coverage probability is 0.95 and is denoted by the horizontal dotted line. The black dots denote one minus the proportion of replicates for which the true value of π falls in the realized confidence intervals, i.e., one minus the estimated coverage probability. The purple squares and blue triangles denote the proportion of replicates that fall below and above realized confidence intervals, respectively. The vertical dotted line denotes the estimated value of $\hat{p}_{n,1}$, $\hat{p}_{n,2}$ or sample size n_1, n_2 for Bendavid et al. (2020a).

sizes n and estimated values $\hat{p}_n$ from Bendavid et al. (2020a).

We find that the delta method and bootstrap intervals are quite liberal. In most cases the estimated coverage of a nominal 95% interval is below 90%. The estimated coverage decreases sharply as n_2 and p_1 become small and is not equi-tailed, in the sense that the proportions of replicates that fall below and above the confidence intervals are not approximately equal. By contrast, the projection method intervals are quite conservative. They are approximately 45% longer than the delta method intervals at sample sizes n and estimated values $\hat{p}_n$ from Bendavid et al. (2020a). These findings motivate the development of approximate and finite-sample valid alternative methods for constructing confidence intervals that have less erratic coverage probabilities.

3. Test Inversion

In this section, we consider both approximate and finite-sample valid approaches to the general problem of constructing test-inversion confidence intervals for parameters θ, where the corresponding null hypothesis $H(\theta_0) : \theta = \theta_0$ is composite. To this end, we require a more general notation. Suppose data X follows a general parametric model indexed by a parameter (θ, ϑ) in parameter space $\bar{\Omega}$. The parameter of interest θ is real-valued, and the nuisance parameter ϑ is finite dimensional. For a fixed value θ_0, the parameter space for the nuisance parameter ϑ is denoted by

$$\bar{\Omega}(\theta_0) = \{(\theta, \vartheta) \in \bar{\Omega} : \theta = \theta_0\} .$$

Observe that for the case of seroprevalence π, we have that $\theta = \pi$ and can assign $\vartheta = (p_1, p_3)$.

Test inversion reduces the problem of confidence interval construction for θ to the problem of testing $H(\theta_0) : \theta = \theta_0$ against $\theta > \theta_0$ and $\theta < \theta_0$. Consider a test of the null hypothesis $H(\theta_0) : \theta = \theta_0$ against the alternative $\theta > \theta_0$. A $1-\alpha/2$ lower confidence bound for θ constructed with test inversion is given by the infimum of the set of θ_0 such that $H(\theta_0)$ is not rejected at level $\alpha/2$ against the alternative $\theta > \theta_0$, which we denote by $L_{1-\alpha/2}$. A $1 - \alpha/2$ upper bound for θ, $U_{1-\alpha/2}$ may be constructed analogously by testing $H(\theta_0)$ against the alternative $\theta < \theta_0$. Thus, a $1 - \alpha$ confidence interval is given by $\left[L_{1-\alpha/2}, U_{1-\alpha/2}\right]$.

3.1 Simple Null Hypotheses

It is illustrative to assume that the nuisance parameter $\vartheta = \vartheta_0$ is known. In this case, the null hypothesis $H(\theta_0)$ is simple and one-sided tests that control the level at $\alpha/2$ are easily constructed. Consider the test that rejects $H(\theta_0)$ against $\theta > \theta_0$ for large values of the test statistic $T_n = T_n(X)$. The cumulative distribution function of T_n is given by $F_{n,\theta,\vartheta}(t) = \mathbb{P}_{\theta,\vartheta}\{T_n \leq t\}$. Additionally, define the related quantity $F_{n,\theta,\vartheta}(t^-) = \mathbb{P}_{\theta,\vartheta}\{T_n < t\}$, and let t_0 denote the observed value of T_n. The null probability that $T_n \geq t_0$ is given by

$$\hat{q}_{L,\theta_0,\vartheta_0} = 1 - F_{n,\theta_0,\vartheta_0}(t_0^-) , \tag{3}$$

and is a valid p-value in the sense that the test that rejects when this quantity is $\leq \alpha/2$ has size $\leq \alpha/2$; see, e.g., Lemma 3.3.1 in Lehmann and Romano (2005).[15] Thus, a $1 - \alpha/2$ confidence

[15]Throughout, we will denote various p-values by $\hat{q}$ rather than $\hat{p}$ because $\hat{p}$ is reserved for various estimates of binomial parameters.

set for θ includes all θ_0 such that $F_{n,\theta_0,\vartheta_0}(t_0^-) < 1-\alpha/2$. If $F_{n,\theta_0,\vartheta_0}(t_0^-)$ is continuous and strictly monotone decreasing in θ_0, then a $1-\alpha/2$ lower confidence bound θ_L may be obtained by solving

$$F_{n,\theta_L,\vartheta_0}(t_0^-) = 1-\alpha/2\ . \tag{4}$$

Similarly, an upper confidence bound $\hat{\theta}_U$ may be obtained by solving $F_{n,\theta_U,\vartheta_0}(t_0) = \alpha/2$.[16] Thus, $[\theta_L, \theta_U]$ is a $1-\alpha$ confidence interval for θ. For a single binomial parameter, this construction gives the classical Clopper and Pearson (1934) interval.

3.2 An Approximate Approach

If the nuisance parameter ϑ is unknown, then it may be approximated. In particular, if $\hat{\vartheta}(\theta_0)$ is the MLE for ϑ subject to the constraint $\theta = \theta_0$, then the infeasible p-value (3) can be replaced with

$$\hat{q}_{L,\theta_0,\hat{\vartheta}(\theta_0)} = 1 - F_{n,\theta_0,\hat{\vartheta}(\theta_0)}(t_0^-)\ , \tag{5}$$

where $F_{n,\theta_0,\hat{\vartheta}(\theta_0)}$ is approximated either analytically or with the parametric bootstrap.[17] Accordingly, the infeasible confidence interval $[\theta_L, \theta_U]$ is replaced with the feasible confidence interval $[\hat{\theta}_L, \hat{\theta}_U]$, where, the endpoints $\hat{\theta}_L$ and $\hat{\theta}_U$ are the values of θ_0 that satisfy

$$F_{n,\hat{\theta}_L,\hat{\vartheta}(\theta_0)}(t_0^-) = 1-\alpha/2 \quad \text{and} \quad F_{n,\hat{\theta}_U,\hat{\vartheta}(\theta_0)}(t_0) = \alpha/2\ , \tag{6}$$

respectively. In other words, either Wald or parametric bootstrap tests are constructed for each θ_0, where the distribution of the test statistic T_n is determined under the parameter $(\theta_0, \hat{\vartheta}(\theta_0))$. This approach was used in DiCiccio and Romano (1990) and DiCiccio and Romano (1995).

In this approximate approach, the family of distributions indexed by (θ, ϑ) has been reduced to an approximate least favorable one-dimensional family of distributions governed by the parameter $(\theta_0, \hat{\vartheta}(\theta_0))$ as θ_0 varies. This approach implicitly orthogonalizes the parameter of interest with respect to the nuisance parameter, so that the effect of estimating the nuisance parameter is negligible

[16] Even in the case that the distribution of T_n is discrete, the function $F_{\theta_0,\vartheta_0}(t_0^-)$ is typically continuous in θ_0 (as in the binomial case). If not, one could use the infimum over θ_0 such that $F_{\theta_0,\vartheta_0}(t_0^-) < 1-\alpha/2$ as a lower bound. Note that, in general, may wish to test at each endpoint of a reported confidence interval to determine whether it should be a closed or open interval. We simply take the conservative approach and report closed intervals.

[17] An alternative, related, approach proceeds by imposing and integrating out a potentially uninformative prior on the nuisance parameter, and then constructs test statistics from the resultant pseudo-likelihood function (see e.g., Severini (1999) and Datta and Mukerjee (2004)).

to second-order and then typically results in second-order accurate confidence intervals. See Cox and Reid (1987) for a discussion of the role of orthogonal parameterizations for inference about a scalar parameter in the presence of nuisance parameters.

3.3 An Infeasible Finite-Sample Approach

The quality of the coverage probability of the approximate intervals considered in the previous section will depend on the quality of the approximation of $\hat{\vartheta}(\theta_0)$ to the true value of the nuisance parameter ϑ_0 and the quality of the analytic or bootstrap approximation to the finite-sample distribution $F_{n,\theta_0,\hat{\vartheta}(\theta_0)}$. In situations in which qualities of these approximations are in doubt, e.g., due to discreteness or the proximity of true parameters to the boundary of their parameter spaces, intervals that ensure coverage of at least $1-\alpha$ in finite samples may be desirable. An infeasible approach to constructing such intervals proceeds by taking the supremum of the p-values over all possible values of the nuisance component ϑ, giving

$$\hat{q}_{L,\theta_0,\sup} = \sup_{(\theta_0,\vartheta)\in\bar{\Omega}(\theta_0)} \left(1 - F_{n,\theta_0,\vartheta}(t_0^-)\right) \ , \tag{7}$$

with finite-sample validity following from

$$\mathbb{P}_{\theta_0,\vartheta_0}\{\hat{q}_{L,\theta_0,\sup} \leq u\} \leq \mathbb{P}_{\theta_0,\vartheta_0}\{1 - F_{n,\theta_0,\vartheta_0}(T_n^-) \leq u\} \leq u.$$

Note that p-values of this form may be conservative, as the supremum over ϑ may be obtained at a value far from ϑ_0.[18] To address this issue, one can restrict the space of values for ϑ that are considered by first constructing a $1-\gamma$ confidence region for ϑ. Such an approach is considered in Berger and Boos (1994), Silvapulle (1996), and Romano et al. (2014). This refined approach proceeds as follows. Fix a small number γ and let $I_{1-\gamma}$ be a $1-\gamma$ confidence region for ϑ.[19]

[18] On the other hand, if the distribution of the test statistic does not vary much with ϑ, then these p-values will not be overly conservative. Therefore, it pays to choose a test statistic that is nearly pivotal, in the sense that its distribution does not depend heavily on ϑ.

[19] The difficulty of constructing such a region depends on the specific model. In the sequel, we focus attention on cases in which confidence regions $I_{1-\gamma}$ can be formed by taking the cartesian product of exact marginal intervals for each component of ϑ.

Consider the modified p-value defined by

$$\hat{q}_{L,\theta_0,I_{1-\gamma}} = \begin{cases} \sup_{\vartheta \in I_{1-\gamma}} \left(1 - F_{n,\theta_0,\vartheta}(t_0^-)\right) + \gamma & \text{if } \{\vartheta \in I_{1-\gamma} : (\theta_0, \vartheta) \in \bar{\Omega}(\theta_0)\} \neq \emptyset \\ \gamma & \text{otherwise .} \end{cases} \tag{8}$$

The p-value obtained by (8) is valid in finite samples; see Berger and Boos (1994) or Silvapulle (1996). Thus, the one-sided test that rejects when $\hat{q}_{L,\theta_0,I_{1-\gamma}} \leq \alpha/2$ leads to a $1 - \alpha/2$ lower confidence bound for θ.

3.4 A Feasible Finite-Sample Approach

The finite-sample valid approach considered in the previous section is infeasible, as it involves computing a supremum over an infinite set $\bar{\Omega}(\theta_0)$. A natural approximation to this approach is to approximate $\bar{\Omega}(\theta_0)$ with a finite discretization, so that the supremum is replaced by maximum over a finite set of values on a grid. In particular, if $G(\theta_0)$ denotes a finite grid over the space $\bar{\Omega}(\theta_0)$, then (7) can be approximated by

$$\hat{q}_{L,\theta_0,\max} = \max_{(\theta_0,\vartheta) \in G(\theta_0)} 1 - F_{n,\theta_0,\vartheta}(t_0^-) \, . \tag{9}$$

Similarly, the refinement given in (8) can be approximated by replacing $I_{1-\gamma}$ with $\hat{I}_{1-\gamma}$, where $\hat{I}_{1-\gamma}$ denotes a finite grid (or ϵ-net) approximating $I_{1-\gamma}$.[20]

We develop a modification to this construction that provably maintains finite-sample Type 1 error control for testing $H(\theta_0)$ by directly accounting for the approximation error induced by a finite discretization of $\bar{\Omega}(\theta_0)$. Towards this end, we require additional structure. Suppose now that the components of the data $X = (X_1, \ldots, X_k)^\top$ are independent, that the distribution of X_i depends on a parameter β_i, with $\beta = (\beta_1, \ldots, \beta_k) \in \Omega$, and that the family of distributions for X_i has a monotone likelihood ratio in X_i. As before, interest focuses on a real-valued parameter $\theta = f(\beta_1, \ldots, \beta_k)$, with the nuisance parameter ϑ given by $\beta_{-1} = (\beta_2, \ldots, \beta_k)$. That is, we assume that the model can be equivalently parameterized by $(\theta, \vartheta) \in \bar{\Omega}$ or through $\beta \in \Omega$, i.e., that the mapping from β to (θ, ϑ) is one-to-one. For a fixed value θ_0, the parameter space for β is given by $\Omega(\theta_0) = \{\beta \in \Omega : f(\beta) = \theta_0\}$.

As before, let $T_n = T_n(X_1, \ldots, X_k)$ be a test statistic for testing $H(\theta_0)$, with t_0 denoting its

[20] In contemporaneous work, Cai et al. (2020) use this construction to develop confidence intervals for seroprevalence that ensure finite-sample Type 1 error control up to the error induced by the finiteness of $G(\theta_0)$.

realized value. Assume that T_n is monotone with respect to each each component X_i.[21] Let $J_{n,\beta}(\cdot)$ denote the cumulative distribution function of T_n for the β- parametrization, so that $J_{n,\beta}(t) = \mathbb{P}_\beta\{T_n \leq t\}$, and let $\hat{\beta}(\theta_0)$ denote the MLE for β subject to the constraint that $\theta = \theta_0$. In this case, for example, we can represent the approximate p-value defined in Section 3.2 with $\hat{q}_{L,\theta_0,\hat{\vartheta}(\theta_0)} = 1 - J_{n,\hat{\beta}(\theta_0)}(t_0^-)$, and similarly for the other p-values previously introduced.

We replace the supremum over $I_{1-\gamma}$ in (8) with a finite maximum while maintaining Type 1 control. Consider a partition of the values of ϑ in $I_{1-\gamma}$ into r regions $E_1, \ldots, E_r$. In our implementation, each region is given by a hyperrectangle of the form $\prod_{i=2}^{k}[\beta_i', \beta_i'']$, though this is not essential. For each region E_j, let $\underline{\beta_{-1}}(j) = (\underline{\beta_2}(j), \ldots, \underline{\beta_k}(j))$ be the vector giving the smallest value that all but the first component of β takes on in E_j; that is $\underline{\beta_i}(j) = \inf\{\beta_i : \beta_{-1} \in E_j\}$. Analogously, let $\bar{\beta}_{-1}(j)$ be the vector giving the largest value that all but the first component of β takes on in E_j. For a hyperrectangle E_j, clearly $\underline{\beta_i}(j) = \beta_i'$ and $\bar{\beta}_i(j) = \beta_i''$. Congruently, let

$$\underline{\beta_1}(j) = \inf\{\beta_1 : \beta \in \Omega(\theta_0), \beta_{-1} \in E_j\} \quad \text{and} \quad \bar{\beta}_1(j) = \sup\{\beta_1 : \beta \in \Omega(\theta_0), \beta_{-1} \in E_j\} \tag{10}$$

denote the smallest and largest values that the first component of β takes on $\Omega(\theta_0)$ for β_{-1} in E_j. If the infimum or supremum in (10) is over a non-empty set, then define $s_L(j) = J_{n,\bar{\beta}(j)}(t^-)$ and $s_U(j) = J_{n,\underline{\beta}(j)}(t)$, where $\bar{\beta}(j) = (\bar{\beta}_1(j), \bar{\beta}_{-1}(j))$ and $\underline{\beta}(j) = (\underline{\beta_1}(j), \underline{\beta_{-1}}(j))$. If there is no β in $\Omega(\theta_0)$ with $(\beta_2, \ldots, \beta_k)$ in E_j, then set $s_L(j) = 1$ and $s_U(j) = 0$, respectively.

We construct the p-values

$$\tilde{q}_{L,\theta_0,I_{1-\gamma}} = \max_{1\leq j\leq r}(1 - s_L(j)) + \gamma \quad \text{and} \quad \tilde{q}_{U,\theta_0,I_{1-\gamma}} = \max_{1\leq j\leq r} s_U(j) + \gamma \tag{11}$$

by taking the maximum over the adjusted p-values $1-s_L(j)$ and $s_U(j)$. This refinement is feasible and valid in finite samples.

Theorem 3.1 *Assume that the components of the data $X = (X_1, \ldots, X_k)^\top$ are independent, that each component X_i has distribution in a family having a monotone likelihood ratio, and that the statistic $T_n = T_n(X_1, \ldots, X_k)$ is monotone increasing with respect to each component X_i. Let $I_{1-\gamma}$ be a finite-sample valid $1-\gamma$ confidence region for $(\beta_2, \ldots, \beta_k)$. Then, the p-values $\tilde{q}_{L,\theta_0,I_{1-\gamma}}$*

[21] If T_n is monotone decreasing with respect to a particular component, say X_j, then X_j can be replaced by $-X_j$, whose family of distributions is then monotone increasing with respect to $-\beta_j$.

and $\tilde{q}_{U,\theta_0,I_{1-\gamma}}$ *are valid for testing* $H(\theta_0)$ *in the sense that, for any* $0 \leq u \leq 1$ *and any* ϑ*,*

$$\mathbb{P}_{\theta_0,\vartheta}\{\tilde{q}_{L,\theta_0,I_{1-\gamma}} \leq u\} \leq u \quad \textit{and} \quad \mathbb{P}_{\theta_0,\vartheta}\{\tilde{q}_{U,\theta_0,I_{1-\gamma}} \leq u\} \leq u\,.$$

PROOF OF THEOREM 3.1. First, note that $I_{1-\gamma}$ could be the whole space $\bar{\Omega}(\theta_0)$ by taking $\gamma = 0$. It follows from Lemma A.1 in Romano et al. (2011) (which is a simple generalization of Lemma 3.4.2 in Lehmann and Romano (2005)) that the family of distributions of T_n satisfies $J_{n,\beta'}(t) \leq J_{n,\beta}(t)$ for any t, $\beta = (\beta_1, \ldots, \beta_k)$, and $\beta' = (\beta'_1, \ldots, \beta'_k)$ with $\beta'_i \geq \beta_i$ for all $i \geq 1$. The same is true if t is replaced by t^-. Thus, we have that $J_{n,\bar{\beta}(j)}(t^-) \leq J_{n,\beta}(t^-)$ and $J_{n,\underline{\beta}(j)}(t) \geq J_{n,\beta}(t)$ for any β with $\theta = f(\beta) = \theta_0$ and $(\beta_2, \ldots, \beta_k) \in E_j$. Therefore, the p-value defined by (8) satisfies

$$\hat{q}_{L,\theta_0,I_{1-\gamma}} \leq \max_{1\leq j\leq r}(1 - s_L(j)) + \gamma = \tilde{q}_{L,\theta_0,I_{1-\gamma}}\,,$$

and similarly $\hat{q}_{U,\theta_0,I_{1-\gamma}} \leq \tilde{q}_{U,\theta_0,I_{1-\gamma}}$ for $\hat{q}_{U,\theta_0,I_{1-\gamma}}$ is defined analogously to (8) for tests of $H(\theta_0)$ against the alternative $\theta < \theta_0$. Since $\hat{q}_{L,\theta_0,I_{1-\gamma}}$ and $\hat{q}_{U,\theta_0,I_{1-\gamma}}$ are valid p-values, then so are any random variables that are stochastically larger. ∎

Thus, tests based on the p-values $\tilde{q}_{L,\theta_0,I_{1-\gamma}}$ and $\tilde{q}_{U,\theta_0,I_{1-\gamma}}$ may be used to test $H(\theta_0)$ and, through test inversion, yield finite-sample valid confidence bounds for θ.

4. Test-Inversion Inference for Seroprevalence

In this section, we apply the general methods considered in Section 3 to the problem of constructing approximate and finite-sample valid confidence intervals for seroprevalence.

4.1 Test Statistics

We begin by exhibiting a set of test statistics T_n applicable to our problem. Let $\hat{p}_n(\pi_0)$ denote the MLE for p restricted to $\Omega(\pi_0) = \{p \in \Omega : \pi = \pi_0\}$.[22] A natural choice for the test statistic T_n is the difference between $\hat{\pi}_n$ and π_0. This statistic can be Studentized with an estimate of its standard deviation, giving

$$\tilde{\pi}_n(\pi_0) = (\hat{\pi}_n - \pi_0) \Big/ \sqrt{V_{\hat{\pi}_n}(\hat{p}_n)}\,,$$

[22] Note that there is no explicit representation for $\hat{p}_n(\pi_0)$, but we may compute its value by solving a convex program.

where $V_{\hat{\pi}_n}(p)$ is given in (2). Alternatively, it can be Studentized with an estimate of its standard deviation under the constraint $\pi = \pi_0$, giving the test statistic

$$\tilde{\pi}_{n,C}(\pi_0) = (\hat{\pi}_n - \pi_0) \Big/ \sqrt{V_{\hat{\pi}_n}(\hat{p}_n(\pi_0))}.$$

Imposing the restriction $\pi = \pi_0$ explicitly in the estimate of the variance may provide a more accurate approximation to the variance of $\hat{\pi}_n$ under the null hypothesis. An analogous improvement has been established for the binomial case in Hall (1982).

Observe that as $p_1 = p_2(1-\pi) + p_3\pi$, we can rewrite the condition $\pi_0 = \pi$ as the linear restriction $b(\pi_0)^\top p = 0$ where $b(\pi_0) = (1, -(1-\pi_0), -\pi_0)^\top$. This observation suggests consideration of the linear test statistic $\hat{\phi}_n(\pi_0) = b(\pi_0)^\top \hat{p}_n$. Moreover, as the variance of $\hat{\phi}_n(\pi_0)$ is exactly equal to

$$V_{\hat{\phi}_n(\pi_0)}(p) = \sigma_1^2(p_1) + (1-\pi_0)^2 \sigma_2^2(p_2) + \pi_0^2 \sigma_3^2(p_3), \tag{12}$$

and can be estimated with the plug-in estimator $V_{\hat{\phi}_n(\pi_0)}(\hat{p}_n)$, the test statistic $\hat{\phi}_n(\pi_0)$ can be Studentized, giving the alternative test statistic

$$\tilde{\phi}_n(\pi_0) = \hat{\phi}_n(\pi_0) \Big/ \sqrt{V_{\hat{\phi}(\pi_0)}(\hat{p}_n)}.$$

In turn, we can Studentize $\hat{\phi}_n(\pi_0)$ with an estimate of its variance under the restriction $\pi_0 = \pi$, giving the statistic

$$\tilde{\phi}_{n,C}(\pi_0) = \hat{\phi}_n(\pi_0) \Big/ \sqrt{V_{\hat{\phi}(\pi_0)}(\hat{p}_n(\pi_0))}.$$

Observe that $\hat{\phi}_n(\pi_0)$ is well-defined if $p_2 = p_3$, and so $\hat{\phi}_n(\pi_0)$, $\tilde{\phi}_n(\pi_0)$, or $\tilde{\phi}_{n.C}(\pi_0)$ may be desirable choices in situations where p_2 is close to p_3.

Alternatively, we can use statistics based on the likelihood function. In particular, let $L(p \mid x)$ be the likelihood function, given by $L(p \mid x) = \prod_{1 \le i \le 3} \binom{n_i}{x_i} p_i^{x_i} (1-p_j)^{n_i - x_i}$. The generalized likelihood ratio test statistic for testing $H(\pi_0): \pi = \pi_0$ is given by

$$W_n = W_n(\pi_0) = 2 \cdot \sum_{1 \le j \le 3} \left(X_j \log\left(\frac{\hat{p}_{n,j}}{\hat{p}_{n,j}(\pi_0)} \right) + (n_j - X_j) \log\left(\frac{1-\hat{p}_{n,j}}{1-\hat{p}_{n,j}(\pi_0)} \right) \right). \tag{13}$$

Large values of W_n give evidence for both $\pi < \pi_0$ and $\pi > \pi_0$. To address this issue, we also

consider the signed square root likelihood ratio statistic for the restriction $\pi = \pi_0$, given by

$$R_n = R_n(\pi_0) = \text{sign}(\hat{\pi}_n - \pi_0) \cdot \sqrt{W_n(\pi_0)}.$$

Corrections to improve the accuracy of W_n based on its signed square root R_n have a long history; see Lawley (1956), Barndorff-Nielsen (1986), Fraser and Reid (1987), Jensen (1986), Jensen (1992), DiCiccio et al. (2001) and Lee and Young (2005). Frydenberg and Jensen (1989) consider the effect of discreteness on the efficacy of corrections to improve asymptotic approximations to the distribution of the likelihood ratio statistic. The statistic R_n can be re-centered and Studentized as

$$\tilde{R}_n = \left(R_n - m_n^R(\hat{p}_n(\pi_0))\right) / \sqrt{V_n^R(\hat{p}_n(\pi_0))},$$

where $m_n^R(p)$ and $V_n^R(p)$ denote the mean and variance of R_n under p, and in practice are computed with the bootstrap under $\hat{p}_n(\pi_0)$.

4.2 Approximate Intervals

We now outline the application of the approximate intervals developed in Section 3.2 to constructing confidence intervals for seroprevalence with the test statistics formulated in Section 4.1, and measure their performance in the Monte Carlo experiment developed in Section 2. Suppose that we are using the test statistic T_n with observed value t_0. Let $J_{n,p}^T(t) = \mathbb{P}_p\{T_n \leq t\}$ denote the distribution of the general statistic T_n under p, and also introduce the related quantity $J_{n,p}^T(t^-) = \mathbb{P}_p\{T_n < t\}$, which will be of use in computing p-values for tests of the null hypothesis $\pi = \pi_0$ against alternatives of the form $\pi > \pi_0$. The approximate test-inversion intervals are constructed by first computing, for each π_0, the p-values

$$\hat{q}_{L,\pi_0,\hat{p}_n(\pi_0)} = 1 - J_{n,\hat{p}_n(\pi_0)}^T(t_0^-) \quad \text{and} \quad \hat{q}_{U,\pi_0,\hat{p}_n(\pi_0)} = J_{n,\hat{p}_n(\pi_0)}^T(t_0).$$

The resultant interval with nominal coverage $1-\alpha$ then takes the form

$$\left\{\pi_0 : q_{L,\pi_0,\hat{p}_n(\pi_0)} \geq \alpha/2 \quad \text{and} \quad q_{U,\pi_0,\hat{p}_n(\pi_0)} \geq \alpha/2\right\}.$$

We begin by considering asymptotic approximations to the distribution $J_{n,\hat{p}_n(\pi_0)}^T$ for different test statistics.

Statistic	Interval	Ave. Length	Ave. Length vs. Delta Method	Coverage
$\tilde{\pi}_{n,C}$	[0.000,0.020]	0.0193	1.0404	0.963
$\tilde{\phi}_{n,C}$	[0.000,0.020]	0.0191	1.0302	0.963
W_n	[0.000,0.021]	0.0177	0.9563	0.927
$\tilde{R}_n$	[0.000,0.021]	0.0181	0.9758	0.950

Table 2: Average Interval Length and Coverage for Test-Inversion Nominal 95% Confidence Intervals for Seroprevalence Using Asymptotic Approximation

Notes: Table 2 reports the approximate test-inversion confidence intervals, constructed with an asymptotic approximation to test statistic null distributions, computed on data from Bendavid et al. (2020a). Estimates of the average length and coverage for these intervals at the n and estimate $\hat{p}_n$ from this study are also displayed. Estimates of average length and coverage are taken over 10,000 bootstrap replicates of X at the sample size n and the estimated parameters $\hat{p}_n$ from this study.

Observe that, under the null hypothesis $H(\pi_0)$ and provided that p is not on the boundary of Ω, the test statistics $\tilde{\pi}_{n,C}(\pi_0)$, $\tilde{\phi}_{n,C}(\pi_0)$, and $\tilde{R}_n$ are asymptotically $\mathcal{N}(0,1)$, and W_n is asymptotically χ^2_1. Thus, if we set T_n equal to any of the asymptotically normal statistics, we can approximate $J_{n,\hat{p}_n(\pi_0)}$ with a standard normal distribution. Likewise, we may apply a χ^2_1 approximation if we set T_n equal to W_n.[23]

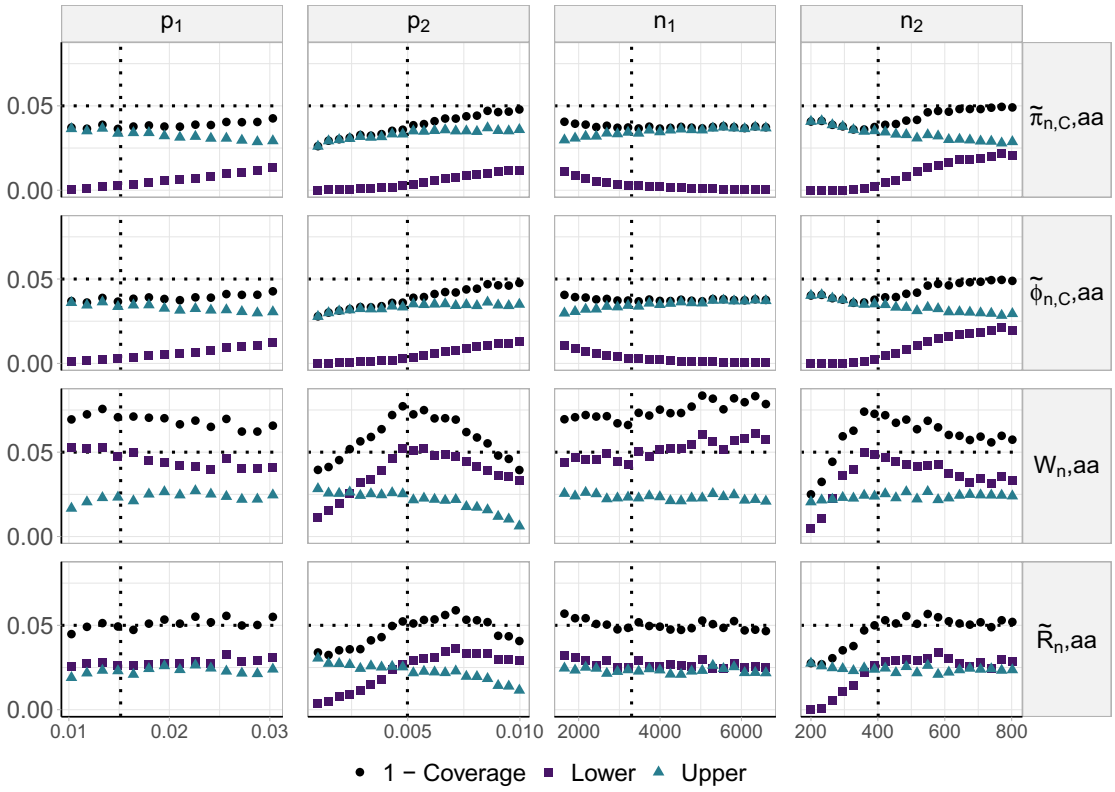
Table 2 reports realizations of these approximate confidence for the observed values from Bendavid et al. (2020a). Additionally, Table 2 presents estimates of coverage and average interval length taken over 10,000 bootstrap replicates computed at the n and estimate $\hat{p}_n$ from this study. Notably, each of these intervals now includes zero.[24] These interval constructions are roughly the same length, on average, as the delta method intervals, but have coverage probability significantly closer to the nominal level.

Figure 2 displays estimates of the coverage probabilities for these intervals in the Monte Carlo experiment developed in Section 2. Recall that the black dots display one minus the estimates of the coverage probabilities of the respective intervals, and that the purple squares and blue triangles display the proportion of the replicates in which the true value of π falls below and above the realized confidence interval, respectively. In contrast to the results for the standard methods displayed

[23]Observe that using normal approximations to $\hat{\pi}_n$ or $\tilde{\pi}_n$ is equivalent to constructing Wald intervals for these statistics. For that reason, we focus on statistics that make explicit use of the null hypothesis restriction $\pi = \pi_0$.

[24]If the null hypothesis $\pi = 0$ is of particular interest or concern, then there exists an exact uniformly most powerful unbiased level α test for the equivalent problem of testing $p_1 = p_2$ against $p_1 > p_2$. This is a conditional one-sided binomial test; see Section 4.5 of Lehmann and Romano (2005). Such a test does not exist for other values of π_0.

Figure 2: Coverage Performance for Test-Inversion Intervals Based on Asymptotic Approximation



Notes: Figure 2 displays estimates of the coverage probabilities of the approximate confidence intervals constructed with an asymptotic approximation to test statistic null distributions. The nominal coverage probability is 0.95 and is denoted by the horizontal dotted line. Estimates of the coverage for the interval constructed with the test statistic T_n are denoted by "T_n, aa" and are computed at parameter values close to the estimates $\hat{p}_n$ and sample size n of Bendavid et al. (2020a) as specified in Section 2. The black dots denote one minus the proportion of replicates for which the true value of π falls in the realized confidence intervals, i.e., one minus the estimated coverage probability. The purple squares and blue triangles denote the proportion of replicates that fall below and above realized confidence intervals, respectively. The vertical dotted line denotes the estimated value of $\hat{p}_{n,1}$, $\hat{p}_{n,2}$, or sample n_1, n_2 for Bendavid et al. (2020a).

in Figure 1, we estimate that the coverage probabilities for these methods are very close to the nominal value of 95% for most parameterizations. The intervals constructed with W_n and $\tilde{R}_n$ are the most equi-tailed.

Next, we refine this approach by directly computing $J_{n,\hat{p}_n(\pi_0)}$ with the bootstrap. This method is more accurate, but can be more computationally expensive. In particular, choosing test statistics that make use of the constrained MLE $\hat{p}_n(\pi_0)$ requires solving the associated convex program for each bootstrap replicate. As a result, for this case, we focus attention on test statistics that do not make use of $\hat{p}_n(\pi_0)$. Table 3 reports realizations of these approximate confidence intervals computed on data from Bendavid et al. (2020a), in addition to estimates of the coverage and average

Statistic	Interval	Ave. Length	Ave. Length vs. Delta Method	Coverage
$\hat{\pi}_n$	[0.000,0.021]	0.0189	1.0186	0.965
$\hat{\phi}_n$	[0.000,0.021]	0.0188	1.0133	0.962
$\tilde{\pi}_n$	[0.000,0.021]	0.0188	1.0150	0.953
$\tilde{\phi}_n$	[0.000,0.021]	0.0188	1.0141	0.953
W_n	[0.000,0.021]	0.0181	0.9788	0.946
R_n	[0.000,0.021]	0.0186	1.0035	0.951

Table 3: Average Interval Length and Coverage for Approximate Test-Inversion Nominal 95% Confidence Intervals Based on the Bootstrap

Notes: Table 3 reports the approximate test-inversion confidence intervals, constructed with a parametric bootstrap approximation to null distributions of test statistics, computed on data from Bendavid et al. (2020a). Estimates of the average length and coverage for these intervals at sample size n and estimate $\hat{p}_n$ from this study are also displayed. Estimates of average length and coverage are taken over 10,000 bootstrap replicates of X at the sample size n and estimate $\hat{p}_n$ from this study.

interval length at the n and estimate $\hat{p}_n$ for this study. Again, each of these intervals include zero, are roughly the same length as the delta-method intervals, and have coverage probability close to the nominal level.

Figure 3 displays estimates of the coverage probabilities in the same Monte Carlo experiment developed in Section 2. Again, these intervals have coverage close to the nominal value and are approximately equi-tailed. The interval constructed with R_n is most equi-tailed and appears to be the least sensitive to perturbations in p_2 and n_2.

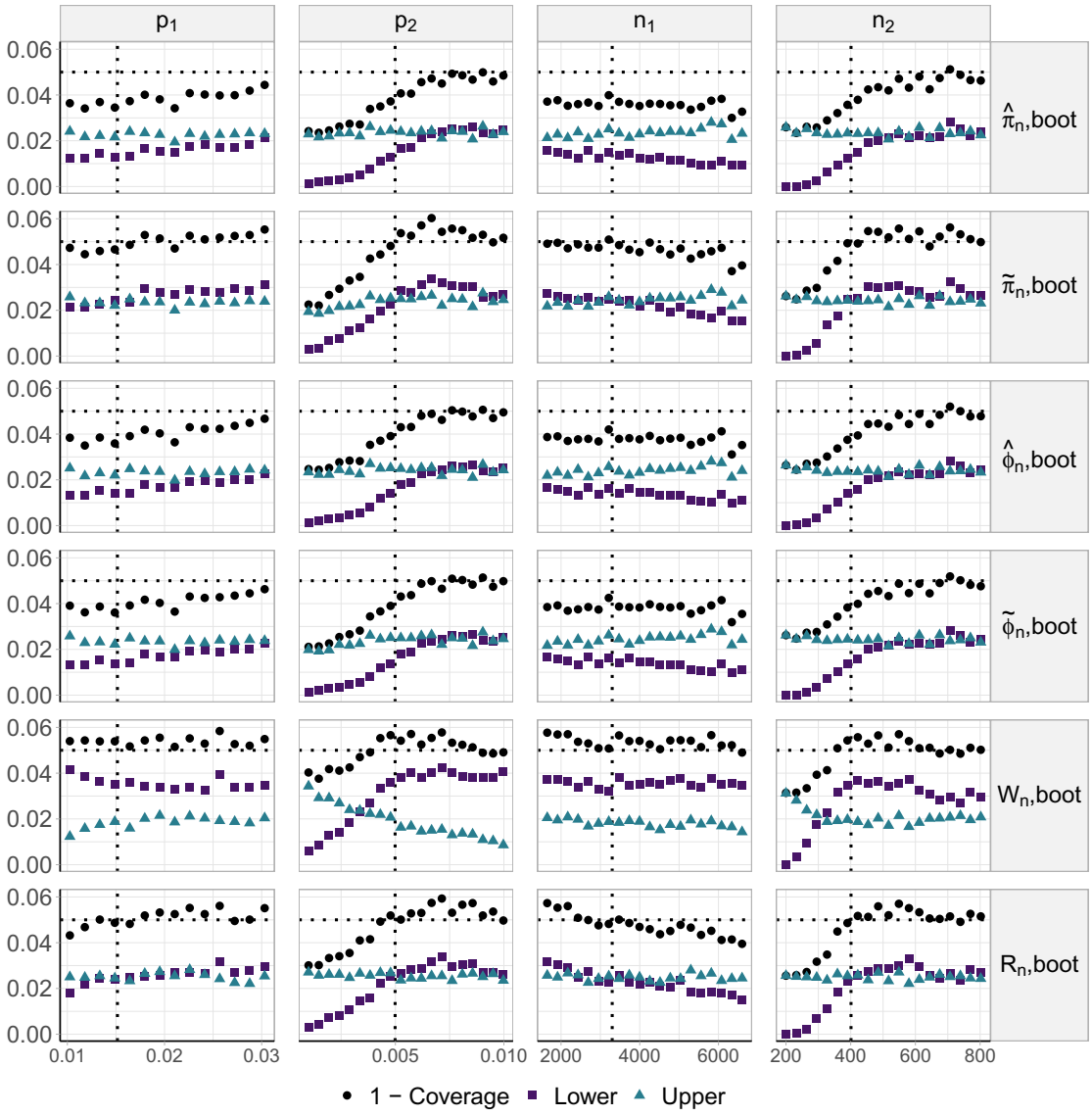
4.3 Finite-Sample Valid Intervals

We now turn to the application of the finite-sample valid intervals discussed in Section 3.4. We focus our development on the test statistic $\hat{\phi}_n(\pi_0)$ as it is linear, and therefore monotone with respect to each sample X_i, as is required.[25]

To begin, we partition the parameter space Ω into a parameter of interest and a nuisance component. Recall that finite-sample exact intervals formed by maximizing p-values over a nuisance space will perform best if the distribution of the chosen test statistic does not vary much with the

[25]One may also consider the test statistic $\hat{\pi}_n$, as $\pi(\cdot)$ is monotone with respect to each component as long as $p \in \Omega$.

Figure 3: Coverage Performance for Test-Inversion Intervals Based on Bootstrap Approximation



Notes: Figure 3 displays estimates of the coverage probabilities of the approximate confidence intervals constructed with the bootstrap. The nominal coverage probability is 0.95 and is denoted by the horizontal dotted line. Estimates of the coverage for the interval constructed with the test statistic T_n are denoted by "T_n, boot" and are computed at parameter values close to the estimates $\hat{p}_n$ and sample size n of Bendavid et al. (2020a) as specified in Section 2. The black dots denote one minus the proportion of replicates for which the true value of π fall in the realized confidence intervals, i.e., one minus the estimated coverage probability. The purple squares and blue triangles denote the proportion of replicates that fall below and above realized confidence intervals, respectively. The vertical dotted line denotes the estimate $\hat{p}_{n,1}$, $\hat{p}_{n,2}$ or sample size n_1, n_2 for Bendavid et al. (2020a).

nuisance parameter. For small values of π_0, the variance of $\hat{\phi}_n(\pi_0)$ is insensitive to changes in p_3, as the variance $\sigma_3^2(p_3)$ enters into (12) linearly and scaled by π_0^2. Additionally, for sample sizes comparable to the measurements taken in Bendavid et al. (2020a), where n_1 is much larger than n_2, the variance of $\hat{\phi}_n(\pi_0)$ will be less sensitive to changes in p_1 than to changes in p_2. Thus, we set the nuisance component $\vartheta = (p_1, p_3)$, giving the parameterization (π, ϑ).[26]

To fix ideas, consider Figure 4, which displays a heat-map of the 0.975 quantile of $\hat{\phi}_n(\pi_0)$ under different values of the nuisance parameter ϑ, with the sample size and the null hypothesis restriction π_0 equal to the sample size and estimated prevalence $\hat{\pi}_n$ from Bendavid et al. (2020a). The square black dot denotes the constrained MLE, $\hat{\vartheta}(\pi_0) = (\hat{p}_{n,1}(\pi_0), \hat{p}_{n,3}(\pi_0))$. The black line exhibits the boundary of the parameter space $\bar{\Omega}(\pi_0)$.

Recall that in the constructions of approximate intervals considered in Section 4.2, a point π_0 is excluded from a confidence interval with nominal coverage 0.95 if the observed value of the chosen test statistic T_n exceeds or falls below the 0.975 or 0.025 quantiles of the statistic's finite-sample distribution at the constrained MLE, $\hat{p}_n(\pi_0)$. However, as illustrated in Figure 4, the 0.975 quantile of the bootstrap distribution of $\hat{\phi}_n(\pi_0)$ has considerable variation with the nuisance parameter ϑ. As a result, these approximate intervals will not exactly control the coverage probability in finite samples, as the event that ϑ differs from $\hat{\vartheta}(\pi_0)$ occurs with positive probability.

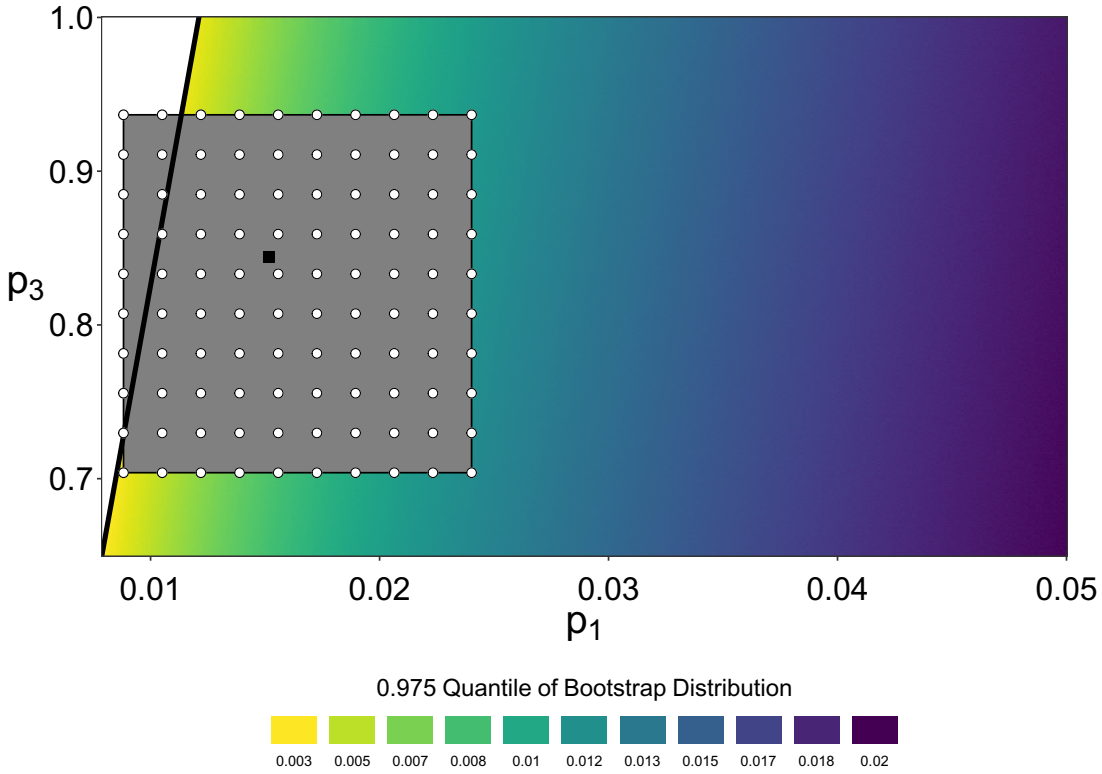
In turn, comparing the realized value of a test statistic to quantiles of the statistic's finite-sample distribution at every value of the nuisance component ϑ is both infeasible, as the space of ϑ is infinite, and impractical, as it would lead to extremely conservative intervals. In fact, we can see that in Figure 4, the 0.975 quantile of the bootstrap distribution of $\hat{\phi}_n(\pi_0)$ is approximately four times as large at $p_1 = 0.05$ than at $p_1 = \hat{p}_{n,1}(\pi_0)$.

Thus, the finite-sample approach developed in Section 3.4 begins by constructing a $1 - \gamma$ confidence region for ϑ and forming a finite grid over this space. The initial confidence region $I_{1-\gamma}$ is illustrated in Figure 4 by the greyed rectangle, and a 10×10 grid over this space is illustrated by the grid of white dots. The confidence region $I_{1-\gamma}$ is formed by taking the Cartesian product of

[26] We note that for different sample sizes, it may be attractive to set the nuisance component $\vartheta = (p_2, p_3)$. For example, Bendavid et al. (2020b) – the April 27th draft of Bendavid et al. (2020a) – includes larger sensitivity and specificity samples n_2 and n_3. In particular, the specificity sample n_2 was increased from 401 to 3324. These additional data were are aggregated over several samples taken at different times and locations. Gelman (2020), Fithian (2020b), and Bennett and Steyvers (2020) highlight issues with this aggregation.

The choice $\vartheta = (p_2, p_3)$ also has computation advantages. In particular, by the identity $p_1 = p_2(1 - \pi) + p_3\pi$, for any value of seroprevalence π_0 and any values of p_2 and p_3 satisfying $p_2 < p_3$, there is a value of p_1 that satisfies $p_2 \leq p_1 \leq p_3$ such that $\pi_0 = (p_1 - p_2)/(p_3 - p_2)$. That is, any value of (p_2, p_3) corresponds to a unique value of p_1 consistent with a given value of π_0 and satisfying the a priori restrictions on the parameter space Ω.

Figure 4: Bootstrap Quantiles and Initial Nuisance Parameter Confidence Region



Notes: Figure 4 displays a heat-map of the 0.975 quantile of $\hat{\phi}_n(\pi_0)$ under different parameter values (π_0, ϑ), where the sample size and null hypothesis restriction π_0 equal to the sample size and estimated prevalence $\hat{\pi}_n$ from Bendavid et al. (2020a). The black line exhibits the boundary of the parameter space $\bar{\Omega}(\pi_0)$. The black square dot denotes the constrained MLE $\hat{\vartheta}(\pi_0) = (\hat{p}_{n,1}(\pi_0), \hat{p}_{n,3}(\pi_0))$. With $\gamma = 0.001$, the grey rectangle denotes a $1 - \gamma$ confidence region for ϑ constructed by taking the cartesian product of two $\sqrt{1-\gamma}$ level confidence regions for p_1 and p_3 each constructed with the method of Clopper and Pearson (1934). The white dots denote a 10×10 grid over this space.

two $\sqrt{1-\gamma}$ level confidence regions for p_1 and p_3, each constructed by using the exact intervals of Clopper and Pearson (1934). For the purposes of this figure, we set $\gamma = 0.001$. This grid partitions the values of ϑ in $I_{1-\gamma}$ into $r = 81$ rectangles, which we enumerate $E_1, \ldots, E_r$. Define, for $i = 1, 3$, the extreme points $\underline{p_i}(j) = \inf\{p_i : (p_1, p_3)^\top \in E_j\}$ and $\overline{p_i}(j) = \sup\{p_i : (p_1, p_3)^\top \in E_j\}$ as well as

$$\underline{p_2}(j) = \inf\{p_2 : p \in \Omega(\pi_0), (p_1, p_3)^\top \in E_j\} \quad \text{and} \tag{14}$$
$$\overline{p_2}(j) = \sup\{p_2 : p \in \Omega(\pi_0), (p_1, p_3)^\top \in E_j\} .$$

As the test statistic $\hat{\phi}_n(\pi_0)$ is monotone increasing in X_1 and monotone decreasing in X_2 and X_3,

define $p_L(j) = \left(\overline{p_1}(j), \underline{p_2}(j), \underline{p_3}(j)\right)$ and $p_U(j) = \left(\underline{p_1}(j), \overline{p_2}(j), \overline{p_3}(j)\right)$ as well as

$$s_L(j) = J^{\hat{\phi}(\pi_0)}_{n,p_L(j)}(t_0^-) \quad \text{and} \quad s_U(j) = J^{\hat{\phi}(\pi_0)}_{n,p_U(j)}(t_0)$$

where $s_L(j)$ and $s_L(j)$ are set equal to 1 and 0, respectively, if the infimum or supremum in (14) are taken over the empty set. Thus, by Theorem 3.1 we can construct the finite-sample valid p-values

$$\tilde{q}_{L,\pi_0,I_{1-\gamma}} = \max_{1\leq j\leq r} \left(1 - s_L(j)\right) + \gamma \quad \text{and} \quad \tilde{q}_{U,\pi_0,I_{1-\gamma}} = \max_{1\leq j\leq r} \left(s_U(j)\right) + \gamma$$

for testing the null hypothesis $\pi = \pi_0$. Hence, the resultant finite-sample valid interval with nominal coverage $1-\alpha$ takes the form $\left\{\pi_0 : \tilde{q}_{L,\pi_0,I_{1-\gamma}} \geq \alpha/2 \quad \text{and} \quad \tilde{q}_{U,\pi_0,I_{1-\gamma}} \geq \alpha/2\right\}$.

This approach is closely related to the method developed in Cai et al. (2020), though there are some differences. Roughly, Cai et al. (2020) compute p-values for test of the null hypothesis $\pi = \pi_0$ with the parametric bootstrap using the particular choice of test statistic $\tilde{\pi}_n$ at each point of a grid spanning a confidence region for the nuisance parameter. Their construction begins by constructing a joint confidence region for all three parameters, while our approach proceeds from a smaller initial region for just p_1 and p_3. We make an additional correction for a grid approximation to the nuisance space, which allows us to ensure finite-sample validity. Their construction does not guarantee that the resulting intervals are equi-tailed.

Table 4 reports realizations of these finite-sample valid intervals for several values of γ, in addition to the projection intervals discussed in Section 2, for Bendavid et al. (2020a).[27] The table also reports estimates of the coverage and average interval length at the estimated values of $\hat{p}_n$ for this study. The cost of ensuring finite-sample valid coverage is large. The realized intervals are roughly 40% wider on average than intervals constructed with the delta method, and the coverage is very close to one. Figure 5 displays estimates of the coverage probabilities for the finite-sample valid intervals as well as the projection intervals in the same Monte Carlo experiment developed in Section 2.[28] Again, the coverage is very close to one at small sample sizes. However, the finite-sample valid intervals outperform the projection intervals. The difference is most salient in the measurements of coverage.

The additional costs associated with correction of the approximation error induced by a finite

[27] These results are insensitive to small changes in the grid size g.

[28] Note that the proportion of Monte Carlo replicates for which the true value of π falls below the realized intervals is very close to zero at most parameter values, and so dots denoting one minus the estimated coverage and the proportion of Monte Carlo replicates for which the true value of π falls above the realized intervals are approximately overlaid.

Method	γ	Interval	Ave. Length	Ave. Length vs. Delta Method	Coverage
Exact	0.0001	[0.000,0.028]	0.0283	1.5311	0.999
	0.0010	[0.000,0.027]	0.0269	1.4538	0.998
	0.0100	[0.000,0.026]	0.0259	1.3979	0.998
Projection		[0.001,0.028]	0.0270	1.4578	1.000

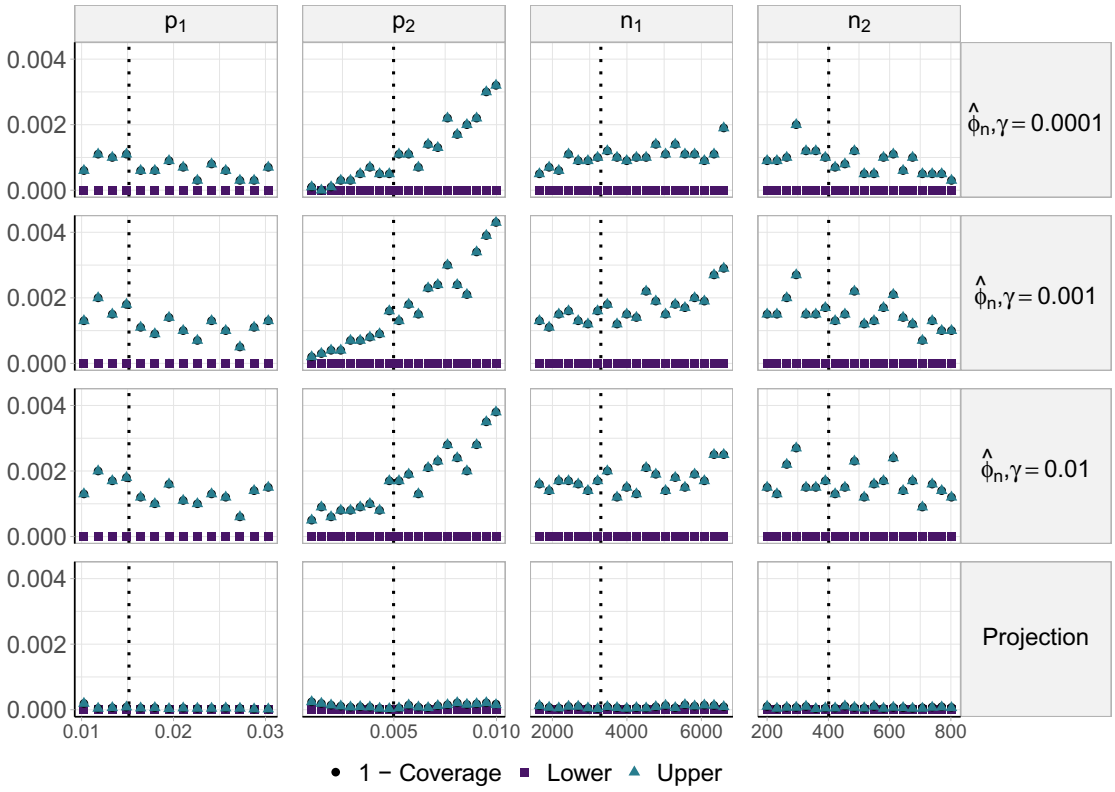
Table 4: Average Interval Length and Coverage for Finite-Sample Valid Test-Inversion and Projection Nominal 95% Intervals

Notes: Table 4 reports the finite-sample valid test-inversion and projection confidence intervals computed on data from Bendavid et al. (2020a). Estimates of the average length and coverage for these intervals at sample size n and estimate $\hat{p}_n$ from this study are also displayed. Estimates of average length and coverage are taken over 10,000 bootstrap replicates of X at the sample size n and the estimate $\hat{p}_n$ from this study.

discretization of the nuisance space are not overly burdensome. In particular, consider the test inversion intervals for π constructed with the p-values $\hat{q}_{L,\pi_0,\hat{I}_{1-\gamma,g}}$ and $\hat{q}_{U,\pi_0,\hat{I}_{1-\gamma,g}}$, where the former p-value is defined in (8), the latter is defined analogously for upper confidence bounds, and $\hat{I}_{1-\gamma,g}$ is a $g \times g$ grid over the initial confidence region $I_{1-\gamma}$. That is, $\hat{I}_{1-\gamma,g}$ denotes the white dots in Figure 4, where in that case $g = 10$. The realized value for these intervals with $g = 10$ and $\gamma = 10^{-2}$ for data from Bendavid et al. (2020a) are $[0.000, 0.025]$. For these values of g and γ, this interval construction has an average length of 0.0249, which is 34.85% longer than the delta-method interval on average, i.e., they are 3.5% shorter than the finite-sample valid intervals considered in this section.

There are several facets of the finite-sample valid confidence intervals considered in this section that could potentially be improved. These include the choice of the nuisance parameter ϑ that leads to an initial confidence region $I_{1-\gamma}$. Additionally, an initial confidence region may be constructed by taking the product of appropriate one-sided bounds, respectively, rather than using a single joint confidence region for both lower and upper bounds. This change should save roughly $\gamma/2$ in over-coverage. It may be more desirable to use a Studentized test statistic, as its distribution may vary even less within the initial confidence region. However, a nontrivial modification of the correction for the approximation error induced by a finite discretization of the nuisance space is required, since monotonicity may be violated. Lastly, finer grids over the nuisance space may be applied to further reduce the length of intervals.

Figure 5: Coverage Performance for Finite-Sample Valid Test-Inversion and Projection Intervals



Notes: Figure 5 displays estimates of the coverage probabilities of the finite-sample valid test inversion and projection confidence intervals. The nominal coverage probability is 0.95. Estimates of the coverage are computed at parameter values close to the estimates $\hat{p}_n$ and sample size n from Bendavid et al. (2020a) as specified in Section 2. The black dots denote one minus the proportion of replicates for which the true value of π falls in the realized confidence intervals, i.e., one minus the estimated coverage probability. The purple squares and blue triangles denote the proportion of replicates that fall below and above realized confidence intervals, respectively. Note that, in the case of this figure, the black dots and blue triangles are approximately overlaid.The vertical dotted line denotes the estimated value of $\hat{p}_{n,1}$, $\hat{p}_{n,2}$, or sample size n_1, n_2 from Bendavid et al. (2020a).

5. Conclusion

We demonstrate that standard methods for constructing confidence intervals in basic seroprevalence surveys derived from the delta method, the percentile bootstrap, and the BC_a bootstrap have coverage probabilities that behave erratically and are not consistently near the nominal level at empirically relevant sample sizes and parameter values. By contrast, we show that methods that combine test inversion with the parametric bootstrap lead to stable coverage probabilities that are close to the nominal level across a variety of statistics. Specifically, statistics that are properly Studentized or based on the generalized likelihood ratio statistic exhibit superior performance. Test

inversion based on the signed square root generalized likelihood ratio statistic gives the best overall performance in terms of stability and validity over a range of empirically relevant parameterizations and sample sizes. On the other hand, if one desires methods with guaranteed coverage in finite-samples, then we have provided an alternative construction with finite-sample validity, at the cost of coverage above the nominal level and longer intervals on average.

Our reanalysis of the uncertainty in estimates of the seroprevalence of SARS-CoV-2 antibodies in Santa Clara County, California on April 3-4, 2020 suggests that the data collected in Bendavid et al. (2020a) are insufficient to rule out small population proportions of SARS-CoV-2 antibodies. However, it is important to note that the maintained assumption – that the sample of antibody tests administered to the population of interest is collected at random – is likely not to hold in many applications. In particular, in the case of Bendavid et al. (2020a), it stands to reason that populations that differed in their likelihood of exposure to COVID-19 differed in their likelihood of volunteering for antibody testing. In an attempt to account for these selection effects, Bendavid et al. (2020a) apply post-stratification weighting by zip code, sex, and race to match population weights measured with the 2018 American Community Survey. With the use of this re-weighting, estimates of seroprevalence and corresponding confidence intervals appear to increase by approximately one to two percent.[29] Consequently, the interpretation of the results of the Bendavid et al. (2020a) seroprevalence survey appears to be contingent on the form of population weighting applied. This sensitivity highlights the fundamental importance of high quality data collection in survey design, and supports a view of seroprevalence surveys as an important input into, but not a final answer for, assessments of the progression of early stages of infectious diseases.

The methods discussed in this paper are applicable to, but not tailored for, post-stratification weighted estimation. If there are S strata of the population of interest, with p_1^i denoting the probability that a randomly selected individual in the ith stratum tests positive, then seroprevalence in the ith stratum is $\pi_i = (p_1^i - p_2)/(p_3 - p_2)$. If the ith stratum gets the known weight w_i, then the overall seroprevalence is $\pi = \sum_i w_i \pi_i$, which is a function of $S + 2$ binomial parameters. However, for even moderately large values of S, the finite-sample valid intervals developed in Section 3.4 will be computationally expensive due to the need to compute p-values for each point in

[29]The demographic data necessary for the replication of this result are not available. Bendavid et al. (2020a) report a weighted seroprevalence estimate and confidence interval – purportedly constructed with the delta method – of 0.0249 and (0.0201,0.0349), respectively. However, Fithian (2020a) argues convincingly that there were coding errors made in the computation of these intervals. Cai et al. (2020) report weighted percentile bootstrap confidence intervals of (0.0110,0.0372), where we note that there are small differences in the specificity and sensitivity estimates that they use relative to the data studied in this article.

a discretization of an $S + 1$ dimensional first-stage confidence region.[30] The computational cost of the approximate intervals considered in Section 3.2 will not be higher than for the unweighted problem. We view further consideration of confidence interval constructions that are well suited for post-stratification weighting as a useful direction for further research.[31]

In many applied contexts it is likely valuable – both for estimation and uncertainty quantification – to incorporate other forms of information made available in the collection of samples and test characteristics. For example, in Bendavid et al. (2020b) a larger specificity sample is constructed by aggregating several samples from different populations. As indicated in Fithian (2020b), there is evidence that there is greater variation in estimates of specificity across these samples than would be expected if each of the test results in these samples were independent and identically distributed. Gelman and Carpenter (2020) propose a hierarchical approach to accounting for this over-dispersion, suggesting a model in which the specificity parameters of the tests implemented in each sample – including the sample taken from the population of interest – are drawn from a pre-specified parametric distribution. As the process generating the specificity samples might tend to be different from the process generating the population of interest (e.g., specificity samples may be drawn from hospital patients local to test manufacturers), we would advocate for an approach in which specificity was modeled as a function of relevant population characteristics. Gelman and Carpenter (2020) also highlight the possibility of incorporating individual-level symptom data; we second this suggestion and view it as a useful direction for further research.

The methods presented in Section 3 apply quite generally to the construction of confidence intervals for real-valued parameters θ. A subclass of problems can be described as follows. Suppose $X_1, \ldots, X_k$ are independent, with X_i distributed as a binomial with parameters n_i and p_i. It is desired to construct a confidence interval for some parameter $\theta = f(p_1, \ldots, p_k)$. In this case, the family of distributions of X_i has monotone likelihood ratio in X_i. The application of Theorem 3.1 requires specification of a nuisance parameter ϑ, construction of a confidence interval for ϑ, and verification that the chosen test statistic is monotone with respect to each of its components. The latter is straightforward when the test statistic is given by $T_n = f(\check{p}_1, \ldots, \check{p}_n)$, where $\check{p}_i = X_i/n_i$. Some important special cases include differences of proportions, measures of relative risk reduc-

[30] In this case, adapting the finite-sample valid procedures proposed in this article for use with asymptotic approximations to the distribution of the signed square root likelihood ratio statistic (see e.g., Brazzale et al. (2007); Jensen (1986, 1992)) may significantly facilitate computation.

[31] Both Gelman and Carpenter (2020) and Cai et al. (2020) propose approaches to this problem, with the former taking a Bayesian perspective.

tion, and odds ratios.[32] Agresti and Min (2002) and Fagerland et al. (2015) also consider similar confidence interval constructions for these three parameters that involve minimizing p-values over a nuisance parameter space, but do not account for the discretization required.

[32]Uniformly most accurate unbiased confidence bounds exist only for the odds ratio based on classical constructions, but optimality considerations fail for the other parameters; see e.g., Problem 5.29 of Lehmann and Romano (2005)

References

Agresti, A. and Min, Y. (2002). Unconditional small-sample confidence intervals for the odds ratio. *Biostatistics*, 3(3):379–386.

Ainsworth, M., Andersson, M., Auckland, K., Baillie, J. K., Barnes, E., Beer, S., Beveridge, A., Bibi, S., Blackwell, L., Borak, M., et al. (2020). Performance characteristics of five immunoassays for sars-cov-2: a head-to-head benchmark comparison. *The Lancet Infectious Diseases*, 20(12):1390–1400.

Alter, G. and Seder, R. (2020). The power of antibody-based surveillance. *The New England Journal of Medicine*.

Barndorff-Nielsen, O. (1986). Inference on full and partial parameters based on the standardized signed log likelihood ratio. *Biometrika*, 73:307–322.

Bendavid, E., Mulaney, B., Sood, N., Shah, S., Bromley-Dulfano, R., Lai, C., Weissberg, Z., Saavedra-Walker, R., Tedrow, J., Bogan, A., Kupiec, T., Eichner, D., Gupta, R., Ioannidis, J. P. A., and Bhattacharya, J. (2021). COVID-19 antibody seroprevalence in Santa Clara County, California. *International Journal of Epidemiology*. dyab010.

Bendavid, E., Mulaney, B., Sood, N., Shah, S., Ling, E., Bromley-Dulfano, R., Lai, C., Weissberg, Z., Saavedra, R., Tedrow, J., Tversky, D., Bogan, T. K., Eichner, D., Gupta, R., Ioannidis, J., and Bhattacharya, J. (2020a). Covid-19 antibody seroprevalence in santa clara county, california. *MedRxiv*, April 11.

Bendavid, E., Mulaney, B., Sood, N., Shah, S., Ling, E., Bromley-Dulfano, R., Lai, C., Weissberg, Z., Saavedra, R., Tedrow, J., Tversky, D., Bogan, T. K., Eichner, D., Gupta, R., Ioannidis, J., and Bhattacharya, J. (2020b). Covid-19 antibody seroprevalence in santa clara county, california. *MedRxiv*, April 27.

Bennett, S. T. and Steyvers, M. (2020). Estimating covid-19 antibody seroprevalence in santa clara county, california. a re-analysis of bendavid et al. *medRxiv*.

Berger, R. and Boos, D. (1994). P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89(427):1012–1016.

Brazzale, A. R., Davison, A. C., Reid, N., et al. (2007). *Applied asymptotics: case studies in small-sample statistics*, volume 23. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, New York.

Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, pages 101–117.

Cai, B., Ioannidis, J., Bendavid, E., and Tian, L. (2020). Exact inference for disease prevalence based on a test with unknown specificity and sensitivity. *arXiv preprint arXiv:2011.14423*.

Carpenter, J. (1999). Test inversion bootstrap confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1):159–172.

Clopper, C. J. and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413.

Cox, D. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society, B*, 49:1–39.

Datta, G. S. and Mukerjee, R. (2004). *Probability matching priors: higher order asymptotics*, volume 178. Lecture Notes In Statistics, 178, Berlin: Springer-Verlag.

Deeks, J. J., Dinnes, J., Takwoingi, Y., Davenport, C., Spijker, R., Taylor-Phillips, S., Adriano, A., Beese, S., Dretzke, J., di Ruffano, L. F., et al. (2020). Antibody tests for identification of current and past infection with sars-cov-2. *Cochrane Database of Systematic Reviews*, (6).

DiCiccio, T., Martin, M., and Stern, S. (2001). Simple and accurate one-sided inference from signed roots of likelihood ratios. *The Canadian Journal of Statistics*, 29(1):67–76.

DiCiccio, T. J. and Romano, J. P. (1990). Nonparametric confidence limits by resampling methods and least favorable families. *Internat. Statist. Rev.*, 58:59–76.

DiCiccio, T. J. and Romano, J. P. (1995). On bootstrap procedures for second-order accurate confidence limits in parametric models. *Statistica Sinica*, pages 141–160.

Efron, B. (1981). Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics*, 9(2):139–158.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185.

Eisen, M. B. and Tibshirani, R. (2020). How to identify flawed research before it becomes dangerous. *The New York Times*, 21.

Fagerland, M. W., Lydersen, S., and Laake, P. (2015). Recommended confidence intervals for two independent binomial proportions. *Statistical methods in medical research*, 24(2):224–254.

Fauci, A. S., Lane, H. C., and Redfield, R. R. (2020). Covid-19—navigating the uncharted. *The New England Journal of Medicine*, 382(13):1268.

Fithian, W. (2020a). "I am grateful to moderator @jonc101x and speaker @jsross119 for allowing me to make a brief statement at the very interesting Stanford BMIR seminar yesterday (43:00 mark). A lightly edited text version follows:". *Tweet*, @wfithian, 22 May, 10:02 A.M.

Fithian, W. (2020b). Statistical comment on the revision of bendavid et al. *https://www.stat.berkeley.edu/wfithian/overdispersionSimple.html*.

Fraser, D. A. S. and Reid, N. (1987). On conditional inference for a real parameter: a differential approach on the sample space. *Biometrika*, 75:251–264.

Frydenberg, M. and Jensen, J. L. (1989). Is the 'improved likelihood ratio statistic' really improved in the discrete case? *Biometrika*, 76:655–661.

Gelman, A. (2020). Concerns with that stanford study of coronavirus prevalence. *Statistical Modeling, Causal Inference, and Social Science blog*.

Gelman, A. and Carpenter, B. (2020). Bayesian analysis of tests with unknown specificity and sensitivity. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(5):1269–1283.

Hall, P. (1982). Improving the normal approximation when constructing one-sided confidence intervals for binomial or poisson parameters. *Biometrika*, 69(3):647–652.

Hall, P. (2013). *The Bootstrap and Edgeworth Expansion*. Springer, New York.

Hui, S. L. and Walter, S. D. (1980). Estimating the error rates of diagnostic tests. *Biometrics*, pages 167–171.

Jensen, J. L. (1986). Similar tests and the standardized log ratio statistic. *Biometrika*, 73:567–572.

Jensen, J. L. (1992). The modified signed likelihood statistic and saddlepoint approximations. *Biometrika*, 79:693–703.

Jewell, N. P. (2004). *Statistics for Epidemiology*. CRC Press, Boca Raton, Florida.

Kolata, G. (2020). Coronavirus infections may not be uncommon, tests suggest. *The New York Times*.

Krammer, F. and Simon, V. (2020). Serology assays to manage covid-19. *Science*, 368(6495):1060–1061.

Lawley, D. N. (1956). A general method for approximating to the distribution of the likelihood ratio criteria. *Biometrika*, 43:295–303.

Lee, S. and Young, G. (2005). Parametric bootstrapping with nuisance parameters. *Statistics and Probability Letters*, 71:143–153.

Lehmann, E. and Romano, J. (2005). *Testing Statistical Hypotheses*. Springer, New York.

Mallapaty, S. (2020). Antibody tests suggest that coronavirus infections vastly exceed official counts. *Nature (Lond.)*.

Peeling, R. W., Wedderburn, C. J., Garcia, P. J., Boeras, D., Fongwen, N., Nkengasong, J., Sall, A., Tanuri, A., and Heymann, D. L. (2020). Serology testing in the covid-19 pandemic response. *The Lancet Infectious Diseases*.

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

Romano, J., Shaikh, A., and Wolf, M. (2011). Consonance and the closure method in multiple testing, article 12. *International Journal of Biostatistics*, 7(1).

Romano, J., Shaikh, A., and Wolf, M. (2014). A practical two-step method for testing moment inequalities. *Econometrica*, 82(5):1979–2002.

Severini, T. A. (1999). On the relationship between bayesian and non-bayesian elimination of nuisance parameters. *Statistica Sinica*, pages 713–724.

Silvapulle, M. (1996). A test in the presence of nuisance parameters. *Journal of the American Statistical Association*, 91(436):1690–1693.

Toulis, P. (2020). Estimation of covid-19 prevalence from serology tests: A partial identification approach. *University of Chicago, Becker Friedman Institute for Economics Working Paper*, (2020-54).

Walter, S. D. and Irwig, L. M. (1988). Estimation of test error rates, disease prevalence and relative risk from misclassified data: a review. *Journal of clinical epidemiology*, 41(9):923–937.