Journal of the Korean Statistical Society (2021) 50:891–904
https://doi.org/10.1007/s42952-021-00131-7

Online ISSN 2005-2863
Print ISSN 1226-3192

RESEARCH ARTICLE

Seroprevalence of SARS-CoV-2 antibodies in South Korea

Kwangmin Lee[1] · **Seongil Jo**[2] · **Jaeyong Lee**[1]

Received: 28 January 2021 / Accepted: 10 May 2021 / Published online: 24 May 2021

Abstract

In 2020, Korea Disease Control and Prevention Agency reported three rounds of surveys on seroprevalence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) antibodies in South Korea. SARS-CoV-2 is the virus which inflicts the coronavirus disease 2019 (COVID-19). We analyze the seroprevalence surveys using a Bayesian method with an informative prior distribution on the seroprevalence parameter, and the sensitivity and specificity of the diagnostic test. We construct the informative prior of the sensitivity and specificity of the diagnostic test using the posterior distribution obtained from the clinical evaluation data. The constraint of the seroprevalence parameter induced from the known confirmed coronavirus 2019 cases can be imposed naturally in the proposed Bayesian model. We also prove that the confidence interval of the seroprevalence parameter based on the Rao's test can be the empty set, while the Bayesian method renders interval estimators with coverage probability close to the nominal level. As of the 30th of October 2020, the 95% credible interval of the estimated SARS-CoV-2 positive population does not exceed 318, 685, approximately 0.62% of the Korean population.

Keywords Seroprevalence · SARS-CoV-2 · Bayesian analysis · Informative prior

✉ Jaeyong Lee
leejyc@gmail.com

Kwangmin Lee
my1989@snu.ac.kr

Seongil Jo
joseongil@gmail.com

[1] Department of Statistics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea

[2] Department of Statistics, Inha University, 100 Inha-ro, Nam-Gu, Incheon, Republic of Korea

1 Introduction

In December 2019, the Chinese government reported a cluster of pneumonia patients of unknown cause in Wuhan, China. It was found that an unknown betacoronavirus causes the disease (Zhu et al., 2020). The Coronaviridae Study Group (CSG) of the International Committee on Taxonomy of Viruses named the virus as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), due to the similarity to SARS-CoV (Gorbalenya et al., 2020). The World Health Organization (WHO) also named the disease caused by SARS-CoV-2 as COVID-19, short for coronavirus disease 2019 (The World Health Organization, 2020a). As of January 10, 2021, over 90,000,000 people in the world are confirmed positive for COVID-19, and there are over 68,000 confirmed cases in South Korea.

Most statistical approaches use the number of confirmed cases to assess the spread of infectious diseases in a population. However, the number of confirmed cases does not include those that are infected but not detected. A seroprevalence survey can provide supplementary information. The seroprevalence is the number of people with antibodies to the virus in a population. The WHO (2020b) proposes to analyze seroprevalence surveys for the inference on the spread of a novel coronavirus. A seroprevalence survey reports result of diagnostic test which is a serological test to detect antibodies of SARS-CoV-2. For example, the third seroprevalence survey in Korea, given in Table 1, reports that 1379 randomly selected people are tested and among them 3 are tested positive. Based on this result, we can infer what percentage of the population has antibodies to the virus.

Seroprevalence surveys have been conducted in many countries, and the results are collected in Serotracker, a global seroprevalence dashboard (Arora et al., 2020). According to the recent update on December 12, 2020, Serotracker provides the survey results of 56 countries based on 491 studies.

The seroprevalence survey data can be analyzed under either the assumption that the diagnostic test used in the survey are 100% accurate or that the test is not. We will term these assumptions the *accuracy* and the *inaccuracy assumptions*, respectively. In words, under the accuracy assumption we assume that the test used in the survey is 100% accurate or equivalently the test has 100% sensitivity and specificity. On the other hand, under the inaccuracy assumption, the

Table 1 The result of the seroprevalence surveys in 2020 (Korea Disease Control and Prevention Agency, 2021)

Accouncement date	Collection period	Number of samples	Number of test-positive samples
9th of July	4.21 ~ 6.16	1500	0
11th of September	6.10 ~ 8.13	1440	1
23rd of November	8.14 ~ 10.31	1379	3

The column of the announcement date represents dates when KDCA reports the results of the surveys. The column of the collection period represents the periods during which the sets of samples are collected

sensitivity and specificity of the test can be less than 100%. Since the sensitivity and specificity of the test do not appear in the statistical model under the accuracy assumption, the statistical inference under the accuracy assumption is simpler than that under the inaccuracy assumption. Because of the simplicity, statistical models under the accuracy assumption are often employed in the real data analysis, especially when clinical evaluation data on the diagnostic test are not available. For example, under the accuracy assumption Song et al. (2020) and Noh et al. (2020) analyzed outpatient data sets in south-western Seoul and Daegu, respectively, and estimated the seroprevalence. Although the statistical models are simpler under the accuracy assumption, the estimates can be biased unless the assumption is met as pointed out in Diggle (2011). Under the inaccuracy assumption, Diggle (2011) proposed a corrected prevalence estimator and Silveira et al. (2020) constructed a confidence interval of the seroprevalence using a resampling method. In an analysis of a seroprevalence survey data of southern Brazil, Silveira et al. (2020) showed that confidence intervals can be $\{0\}$, which is hardly reliable. See Supplementary Table 2 in Silveira et al. (2020). We also prove that the confidence interval constructed from the Rao's test using the duality theorem (Bickel & Doksum, 2015) can be the empty set. These examples show that the frequentist confidence intervals of the seroprevalence under the inaccuracy assumption can be unreliable.

In this paper, we propose a Bayesian method under the inaccuracy assumption and apply the proposed method to the seroprevalence surveys of the South Korean population conducted in 2020 (Korea Disease Control and Prevention Agency, 2021). We use the posterior distribution obtained from the Bayesian model of the clinical evaluation data (Kohmer et al., 2020) as the informative prior distribution of the sensitivity and specificity on the diagnostic test.

The rest of the paper is organized as follows. In the next section, we describe the seroprevalence surveys of SARS-CoV-2 motivating this work and the diagnostic test for detection of SARS-CoV-2 antibodies. In Sect. 3, we conduct a frequentist analysis and discuss the phenomenon of empty confidence sets. In Sect. 4, we propose a Bayesian method for the seroprevalence survey. In Sect. 5, we compare that the proposed Bayesian method with two frequentist methods via simulation study and analyze the seroprevalence surveys of the South Korean population. We conclude the paper with a discussion section.

2 Seroepidemiological surveys and clinical evaluation of a serology test

2.1 Seroepidemiological surveys of SARS-CoV-2 in South Korea

Korea Disease Control and Prevention Agency (KDCA) conducted three rounds of seroprevalence surveys of SARS-CoV-2 for South Korean population in 2020. KDCA used the sets of samples collected in the Korea National Health and

Nutrition Examination Survey (KNHNES), which is a regular national survey to investigate the health and nutritional status of South Koreans since 1998 (Kwon et al., 2014), as the samples of the seroprevalance surveys, and perfomed the SARS-CoV-2 serological tests for the residual serum samples. Table 1 shows the summary of test results and the periods during which the samples are collected.

2.2 Clinical evaluation of plaque reduction neutralization test for SARS-CoV-2 antibodies

When KDCA performs a serology test for SARS-CoV-2, they use their in-house plaque reduction neutralization test (PRNT), a kind of serology test, which tests serum samples for their neutralization capacity against SARS-CoV-2. The statistical model we use for the seroprevalence survey data contains unknown sensitivity and specificity of the PRNT but KDCA does not provide clinical evaluation data for the sensitivity and specificity. We use the clinical evaluation data (Kohmer et al., 2020), summarized in Table 2, to construct informative prior as well as estimators of the unknown sensitivity and specificity.

3 Maximum likelihood estimator and a confidence interval

In this section, we specify a statistical model for seroprevalance surveys, and present the maximum likelihood estimator and a confidence interval of the seroprevalance. We assume that the sensitivity and specificity of serology test are fixed values for the maximum likelihood estimator and the confidence interval. Note that the sensitivity and specificity are the probabilities that the true positive has the positive test result and the ture negative has the negative test result, respectively.

We define *seroprevalence parameter*, θ, as the proportion of those who have antibodies against SARS-CoV-2 in the population. Let N be the sample size of the seroprevalence survey and X be the number of test-positive samples by the serology test used in the survey. Let p_+ and p_- denote the sensitivity and specificity of the serology test, respectively. We assume X is generated from the binomial distribution:

Table 2 The data of clinical evaluation of the PRNT by Kohmer et al. (2020)

		True state		
		Positive	Negative	Total
Test results of the PRNT	Positive	42	1	43
	Negative	3	34	37
	Total	45	35	80

The true state of a sample refers to whether the sample has the antibodies against SARS-CoV-2 in reality. This data set is used to construct informative prior as well as estimators of the unknown sensitivity and specificity of the PRNT

$$X \sim Binom\big(N, \theta p_+ + (1-\theta)(1-p_-)\big), \tag{1}$$

where $Binom(n, p)$ denotes the binomial distribution with parameters $n \in \mathbb{N}$ and $p \in [0, 1]$.

The binomial probability in (1), $\theta p_+ + (1-\theta)(1-p_-)$, is the probability that a subject gets a positive result from the serology test. When a subject gets a positive test result, two cases are possible, i.e., either the subject has the antibody and the test result is correct, or the subject does not have the antibody and the test result is incorrect. If subjects in the survey are randomly sampled, and the survey sampling is independent of the serology test error, then the probabilities of these two cases are θp_+ and $(1-\theta)(1-p_-)$, respectively. Thus, the probability of the positive test result is given as the sum of these probabilities.

When p_+ and p_- are given, the maximum likelihood estimator for θ is as follows. If $1 - p_- < p_+$, then

$$\hat{\theta}^{MLE} = \begin{cases} 0 & \text{if } X \le N(1-p_-), \\ 1 & \text{if } X \ge Np_+, \\ \frac{X/N-(1-p_-)}{p_++p_--1} & \text{if } N(1-p_-) < X < Np_+, \end{cases} \tag{2}$$

and if $p_+ < 1 - p_-$, then

$$\hat{\theta}^{MLE} = \begin{cases} 0 & \text{if } X \ge N(1-p_-), \\ \frac{X/N-(1-p_-)}{p_++p_--1} & \text{if } Np_+ < X < N(1-p_-). \end{cases}$$

Note if the number of test-positive samples is small or large enough, the maximum likelihood estimator can be 0 or 1. This means that nobody or everybody in the population has antibodies against SARS-CoV-2, which is hardly reliable.

We construct a confidence interval of θ from Rao's test (Rao, 1948) using the duality thoerem (Bickel & Doksum, 2015), and show that when X is too small or large, the confidence interval can be the empty set. Let $A(\theta_0) = [l_{\theta_0}, u_{\theta_0}]$ be the $100(1-\alpha)\%$ acceptance interval of the Rao's test under the null hypothesis $H_0 : \theta = \theta_0$. By the duality theorem $S(X) = \{\theta_0 \in [0,1] : X \in A(\theta_0)\}$ is a $100(1-\alpha)\%$ confidence interval for θ. Theorem 1 gives the acceptance interval, $A(\theta_0)$, and the condition that the confidence interval $S(X)$ is the empty set.

Theorem 1 *Consider model* (1).

(a) *The* $100(1-\alpha)\%$ *acceptance region of the test*

$$H_0 : \theta = \theta_0 vs H_1 : \theta \neq \theta_0$$

by the Rao's Score test is given as

$$[l_{\theta_0}, u_{\theta_0}] = [N\theta_0^* - \{N\chi^2_{0.05}(1)\theta_0^*(1-\theta_0^*)\}^{1/2}, N\theta_0^* + \{N\chi^2_{0.05}(1)\theta_0^*(1-\theta_0^*)\}^{1/2}],$$

where $\theta_0^* = \theta_0 p_+ + (1-\theta_0)(1-p_-)$ *and* $\chi_\alpha^2(1)$ *is* $100(1-\alpha)\%$ *quantile of chi-square distribution with* 1 *degree of freedom.*

(b) *If*

$$X < \inf_{\theta_0 \in [0,1]} l_{\theta_0} \, or X > \sup_{\theta_0 \in [0,1]} u_{\theta_0}, \tag{3}$$

the $100(1-\alpha)\%$ *confidence interval* $S(X) = \{\theta_0 \in [0,1] : X \in A(\theta_0)\}$ *is the empty set.*

Proof (a) Let $\theta^* = \theta p_+ + (1-\theta)(1-p_-)$ and

$$L(\theta) = L(\theta; N, X) = \binom{N}{X} \theta^X (1-\theta)^{N-X}.$$

The score statistics is given by

$$\begin{aligned}
&\left(\frac{d \log L(\theta)}{d\theta}\right)^2_{\theta=\theta_0} \left[E\left(-\frac{d^2 \log L(\theta)}{d\theta^2}\right)_{\theta=\theta_0}\right]^{-1} \\
&= \left(\frac{d \log L(\theta)}{d\theta^*}\right)^2_{\theta=\theta_0} \left[E\left(-\frac{d^2 \log L(\theta)}{d(\theta^*)^2}\right)_{\theta=\theta_0}\right]^{-1} \\
&= \frac{(X - N\theta_0^*)^2}{N\theta_0^*} + \frac{(N - X - N(1-\theta_0^*))^2}{N(1-\theta_0^*)} \\
&= \frac{(X - N\theta_0^*)^2}{N\theta_0^*(1-\theta_0^*)}.
\end{aligned}$$

Then, the $100(1-\alpha)\%$ acceptance interval is

$$\begin{aligned}
A(\theta_0) &= \left\{X : \frac{(X - N\theta_0^*)^2}{N\theta_0^*(1-\theta_0^*)} \le \chi_\alpha^2(1)\right\} \\
&= \{X : |X - N\theta_0^*| \le \{\chi_\alpha^2(1) N p_0^*(1-\theta_0^*)\}^{1/2}\},
\end{aligned}$$

which proves (a).

(b) If

$$X < \inf_{\theta_0 \in [0,1]} l_{\theta_0} \text{ or } X > \sup_{\theta_0 \in [0,1]} u_{\theta_0},$$

then $X \notin A(\theta_0)$ for all $\theta_0 \in [0,1]$. It implies the confidence interval of X is the empty set. This completes the proof. □

The intuitive reason for the empty confidence interval is as follows. The set of sampling distributions for X is

$$\left\{Binom(N, \theta) : (1-p_-) \le \theta \le p_+\right\}.$$

When X/N is smaller (larger) than $1 - p_-$ (p_+), the probability that X is observed is small for every sampling distribution in the set. This makes test decisions be rejected for every null hypothesis. Thus, the extreme X implies p_- and p_+ are doubtful.

4 A Bayesian method with informative prior distributions

We propose a Bayesian method that avoids the empty confidence set problem. For the Bayesian analysis of model (1), we assign prior distributions on θ, p_+ and p_-. According to KNHNES design, the parameter θ refers to the seroprevalence in the population that includes those who have been confirmed to be tested positive for COVID-19 by the government. Thus, we need to assume θ is larger than the proportion of the confirmed cases, and we choose the following constrained prior distribution on parameter θ:

$$\pi(\theta) \propto (\theta)^{-1/2}(1-\theta)^{-1/2} I(\theta > \tilde{\theta}), \tag{4}$$

where $\pi(\theta)$ is the density function of the prior distribution on θ, and $\tilde{\theta}$ is the total number of confirmed cases divided by the number of the population. Note that the constrained prior distribution (4) is constructed by constraining Jeffereys prior distribution for binomial parameter (Yang and Berger, 1996).

To construct prior distributions on p_+ and p_-, we use the posterior distribution on the sensitivity and specificity obtained from a clinical evaluation of the serology test. In the clinical evaluation, we consider that the serology test is applied to samples of which the true states are known. The true state of a sample refers to whether the sample has the antibodies in reality. The data from the clinical evaluation is then represented as Table 3.

For the analysis of the clinical evaluation (Table 3), we specify a statistical model using the binomial distribution as

$$r_{++} \sim Binom(r_{\cdot+}, p_+), \tag{5}$$

$$r_{--} \sim Binom(r_{\cdot-}, p_-). \tag{6}$$

By applying the Jeffereys prior to the binomial parameters p_+ and p_-, we obtain the densitiy functions of posterior distributions, $\pi^*(p_+ \mid r_{++}, r_{\cdot+})$ and $\pi^*(p_- \mid r_{--}, r_{\cdot-})$, as

Table 3 Data format for clinical evaluation when the true states of samples are known

		True state		
		Positive	Negative	Total
Test result	Positive	r_{++}	r_{+-}	$r_{+\cdot}$
	Negative	r_{-+}	r_{--}	$r_{+\cdot}$
	Total	$r_{\cdot+}$	$r_{\cdot-}$	$r_{\cdot\cdot}$

$$\begin{aligned}\pi^*(p_+ \mid r_{++}, r_{\cdot+}) &\propto p^{Binom}(r_{++} \mid r_{\cdot+}, p_+)(p_+)^{1/2}(1-p_+)^{1/2},\\ \pi^*(p_- \mid r_{--}, r_{\cdot-}) &\propto p^{Binom}(r_{--} \mid r_{\cdot-}, p_-)(p_-)^{1/2}(1-p_-)^{1/2},\end{aligned} \tag{7}$$

where $p^{Binom}(\cdot \mid n, p)$ is the densitiy function of the binomial distribution $Binom(n, p)$ for $n \in \mathbb{N}$ and $p \in [0, 1]$. Note that the Jeffereys prior is a conjugate prior for the likelihood function $p^{Binom}(\cdot \mid n, p)$. Thus, the density function of the posterior distributions are calculated as

$$\begin{aligned}\pi^*(p_+ \mid r_{++}, r_{\cdot+}) &\propto (p_+)^{(r_{++}+1/2)}(1-p_+)^{(r_{\cdot+}-r_{++}+1/2)},\\ \pi^*(p_+ \mid r_{--}, r_{\cdot-}) &\propto (p_-)^{(r_{--}+1/2)}(1-p_-)^{(r_{\cdot-}-r_{--}+1/2)}.\end{aligned}$$

Finally, we use the posterior distributions to construct the informative prior distributions on p_+ and p_- of model (1). That is, we set

$$\begin{aligned}\pi(p_+) &\propto (p_+)^{(r_{++}+1/2)}(1-p_+)^{(r_{\cdot+}-r_{++}+1/2)},\\ \pi(p_-) &\propto (p_-)^{(r_{--}+1/2)}(1-p_-)^{(r_{\cdot-}-r_{--}+1/2)},\end{aligned}$$

where $\pi(p_+)$ and $\pi(p_-)$ are the density functions of the informative prior distributions.

The posterior samples from the posterior are obtained by STAN (Carpenter et al., 2017) and the STAN code for the posterior is given in the supplementary material.

5 Numerical studies

5.1 Simulation study

In this subsection, we compare the proposed Bayesian method with two frequentist methods with accuracy and inaccuracy assumptions. The frequentist method with inaccuracy assumption uses the maximum likelihood estimator and the confidence interval given in (2) and Theorem 1, respectively. For the sensitivity and specificity in (2), we plug-in the maximum likelihood estimator from the generated clinical evaluation data. The frequentist method with accuracy assumption considers the statistical model

$$X \sim Binom(N, \theta),$$

instead of model (1). The frequentist method with accuracy assumption uses X/N as a point estimator for θ, and constructs a confidence interval of θ from the approach introduced in Clopper and Pearson (1934). Note that the frequentist method with accuracy assumption does not consider serology test error and assumes serology test is 100% accurate.

For the simulation study, we generate the seroprevalence survey data from the distributions (1) and the clinical evaluation data from (5) and (6). We fix

sample sizes in these distributions as $N = 1500$, $r_{\cdot+} = 45$ and $r_{\cdot-} = 35$ based on Tables 1 and 2. We set $(p_+, p_-) = (0.80, 0.80)$, (0.95, 0.95) and (0.99, 0.99) since 95% confidence intervals of p_+ and p_- based on Table 2 are included in [0.80, 1] and estimates of them are close to (0.95, 0.95). For parameter θ, we assume $\theta = r\tilde{\theta}$, where $\tilde{\theta}$ is the cumulative number of confirmed cases, and consider $\tilde{\theta} \in \{0.1\%, 0.2\%, \ldots, 1\%\}$ and $r = 4$ to describe the early stage of the outbreak. If the number of the confirmed cases is 0.1% of the population and the people with antibodies are 0.4% of population, $\tilde{\theta} = 0.1\%$ and $r = 4$. In the supplementary material, we include the simulation results with $r = 2$ and 8, and $(p_+, p_-) = (0.85, 0.85)$ and (0.9, 0.9).

We assess the performance of the proposed Bayesian method and two frequentist methods using the mean squared error and the coverage probability given as

$$\text{mean squared error} = \frac{1}{S}\sum_{i=1}^{S}(\hat{\theta}_i - \theta)^2,$$

$$\text{coverage probability} = \frac{1}{S}\sum_{i=1}^{S} I(l_i < \theta < u_i),$$

where S is the number of repetitions, θ is the true seroprevalence, $\hat{\theta}_i$ is point estimators, and $[l_i, u_i]$ denotes interval estimators. We use the 95% confidence intervals for two frequentist methods, and the 95% highest posterior density interval for the Bayesian method. The posterior mean is used as the point estimator of the Bayesian method. In this simulation study, we set $S = 100$.

Figure 1 shows the mean squared errors for the proposed Bayesian method and two frequent methods. In all cases, the Bayesian method has the smallest mean squared errors. Of the two frequentist methods, the frequentist method with inaccuracy method is generally better, but when $(p_+, p_-) = (0.99, 0.99)$ and the accuracy

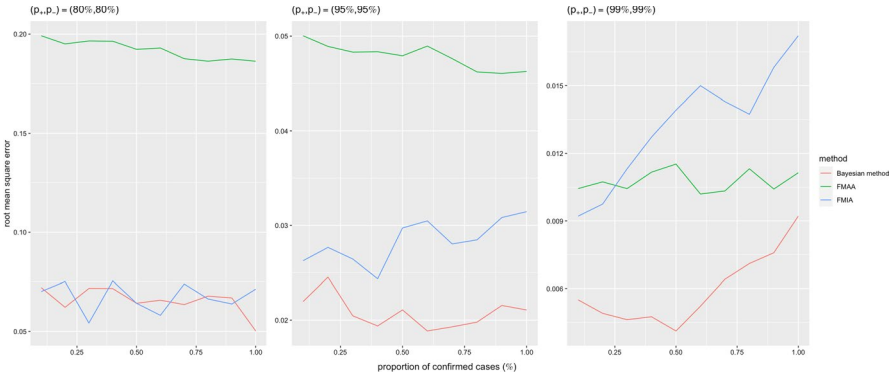

Fig. 1 The graphs represent the root mean square errors of the Bayesian method and the frequentist methods with accuracy or inaccuray assumptions when $(p_+, p_-) = (0.80, 0.80)$, (0.95, 0.95) and (0.99, 0.99). The root mean square error is in y-axis and the proportion of the confirmed cases $\tilde{\theta} \in \{0.1\%, 0.2\%, \ldots, 1\%\}$ is in x-axis. "FMAA" and "FMIA" refer to the frequentist methods with accuracy and inaccuracy assumptions, respectively

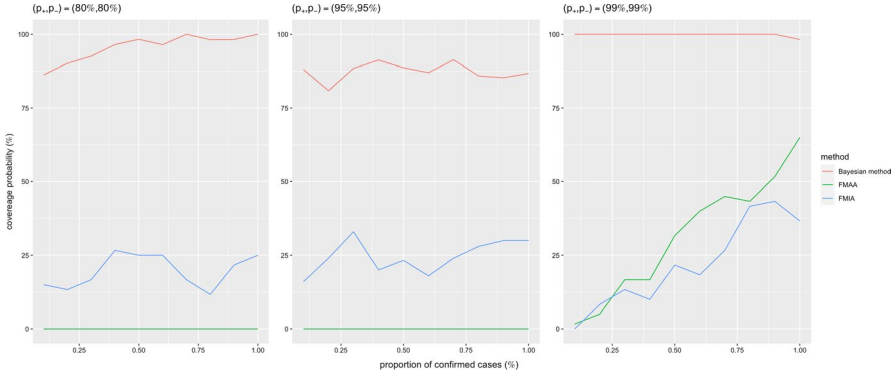

Fig. 2 The graphs represent the coverage probabilities of 95% interval estimators of the Bayesian method and the frequentist methods with accuracy or inaccuray assumptions when $(p_+, p_-) = (0.80, 0.80)$, (0.95, 0.95) and (0.99, 0.99). The coverage probability is in y-axis and the proportion of the confirmed cases $\tilde{\theta} \in \{0.1\%, 0.2\%, \ldots, 1\%\}$ is in x-axis. "FMAA" and "FMIA" refer to the frequentist methods with accuracy and inaccuracy assumptions, respectively

assumption is almost met, the frequentist method with accuracy assumption gets better.

Figure 2 shows the coverage probabilities for the proposed Bayesian method and two frequent methods. In all cases, the Bayesian method has the coverage probabilities close to the nominal 95% coverages, while those of the frequentist methods are not even close to the nominal coverage probability. Even as a frequentist method, the proposed Bayesian method renders superior interval estimators.

5.2 Seroprevalence in South Korea

In this subsection, we apply the proposed Bayesian method and the frequentist methods to the three rounds of surveys given in Table 1.

Under the inaccuracy assumption we show all the maximum likelihood estimators are zero and the confidence intervals are the empty set. We assume the fixed (p_+, p_-) to be (42/45, 34/35), which is calculated from the clinical evaluation data (Table 2) and formula $(r_{++}/r_{\cdot+}, r_{--}/r_{\cdot-})$ according to the notation in Table 3. Based on Eq. (2), all the maximum likelihood estimatiors are zero, since values of $N(1 - p_-)$ are 42.9, 41.1 and 39.4 which are all larger than the observed Xs. The confidence intervals are the empty set since values of $\inf_{\theta \in [0,1]} l_\theta$ are 30.2, 28.8 and 27.3 which satisfy condition (3) in Theorem 1.

We analyze the survey data (Table 1) using the proposed Bayesian method and compare the result with that by the frequentist method with accuracy assumption. Let θ_1, θ_2 and θ_3 be the seroprevalance parameters for each survey.

We construct a constrained prior distribution on $(\theta_1, \theta_2, \theta_3)$. As Eq. (4) we assign conditions that each θ_i is larger than the percentage of the confirmed cases at the survey date. Additionally, we add constraint $I(\theta_1 \le \theta_2 \le \theta_3)$ since the seroprevalence is monotonely increasing over time. The prior distribution with the constraint is

$$\pi(\theta_1, \theta_2, \theta_3) \propto \prod_{i=1}^{3} \theta_i^{-1/2}(1-\theta_i)^{-1/2} I(\theta_i > \tilde{\theta}_i) I(\theta_1 \le \theta_2 \le \theta_3),$$

where $\tilde{\theta}_i$ is the proportion of the cumulative confirmed cases at the last date in the collection period of the ith set of samples for $i \in \{1, 2, 3\}$. We construct informative prior distributions on p_+ and p_- using the clinical evaluation of the PRNT for SARS-CoV-2 performed by Kohmer et al. (2020). By applying the clinical evaluation data (Table 2) to Eq. (7), we obtain the informative prior distributions as

$$\begin{aligned} p_+ &\sim Beta(42.5, 3.5), \\ p_- &\sim Beta(34.5, 1.5), \end{aligned}$$

where $Beta(\alpha, \beta)$ denotes the beta distribution with the density function of

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx}.$$

Collecting the prior distributions and three rounds of seroprevalance survey results, we construct the generative model as

$$\begin{aligned} X_1 \mid \theta_1, p_+, p_- &\sim Binom(N_1, \theta_1 p_+ + (1-\theta_1)(1-p_-)), \\ X_2 \mid \theta_2, p_+, p_- &\sim Binom(N_2, \theta_2 p_+ + (1-\theta_2)(1-p_-)), \\ X_3 \mid \theta_3, p_+, p_- &\sim Binom(N_3, \theta_3 p_+ + (1-\theta_3)(1-p_-)), \\ p_+ &\sim Beta(42.5, 3.5), \\ p_- &\sim Beta(34.5, 1.5), \\ \pi(\theta_1, \theta_2, \theta_3) &\propto \prod_{i=1}^{3} (\theta_i)^{-1/2}(1-\theta_i)^{-1/2} I(\theta_i \ge \tilde{\theta}_i) I(\theta_1 \le \theta_2 \le \theta_3), \end{aligned}$$

where (N_i, X_i) is the pair of the number of samples and the number of test-positive samples of ith seroprevalance survey for $i \in \{1, 2, 3\}$.

For inference, we generate posterior samples using Markov chain Monte Carlo (MCMC) sampling method. Specifically, we generate 4000 posterior samples through running 4 Markov chains with different initial values, where each

Table 4 Summary statistics of posterior distributions of the population who has antibodies against SARS-CoV-2 for the three rounds of the seroprevalance surveys

Date	Cumulative confirmed cases	Posterior mean	The 95% credible interval
16th of June	12198	26014.9	[12531.4, 63146.7]
13th of July	14873	55712.6	[18402.4, 137279.3]
31st of October	26635	133755.8	[29025.2, 318684.6]

The date column represents the last dates of the collection period of each survey. The column of cumulative confirmed cases represents the cumulative numbers of confirmed cases on the corresponding dates

chain has 1000 samples after a burn-in period of 1000 samples. We implement the MCMC algorithm with Stan (Carpenter et al., 2017). We extract the posterior samples of θ_1, θ_2 and θ_3, and multiply the number of the population in 2020, 51, 829, 023 (Ministry of the Interior and Safety, 2021), to the parameters. We then give the summary statistics of the multiplied posterior samples in Table 4.

According to Table 4, the ratio of the posterior mean to the confirmed cases ranges from 3.1 to 4.5, which represents the proportion of the total number of the infected cases divided by the total number of the detected cases.

Finally, we compare the result of the proposed Bayesian method with the cumulative number of confirmed cases and the result of the frequentist method with the accuracy assumption. As in the proposed Bayesian method, we multipy the number of the population to the point estimator and the confidence interval. The comparison is then represented in Fig. 3.

Figure 3 shows that the lower bounds of interval estimation by the Bayesian method are larger than the number of confirmed cases as expected, but the other does not satisfy the inequality condition. Each upper bound of the interval estimations by the Bayesian method is smaller than the corresponding one obtained by the frequentist method with the accuracy assumption. Under the inaccuracy assumption, the Bayesian method considers that test-positive cases may include false-negative cases, which is critical when the test-positive number is small enough. Thus, the Bayesian method makes the upper bounds shrink.

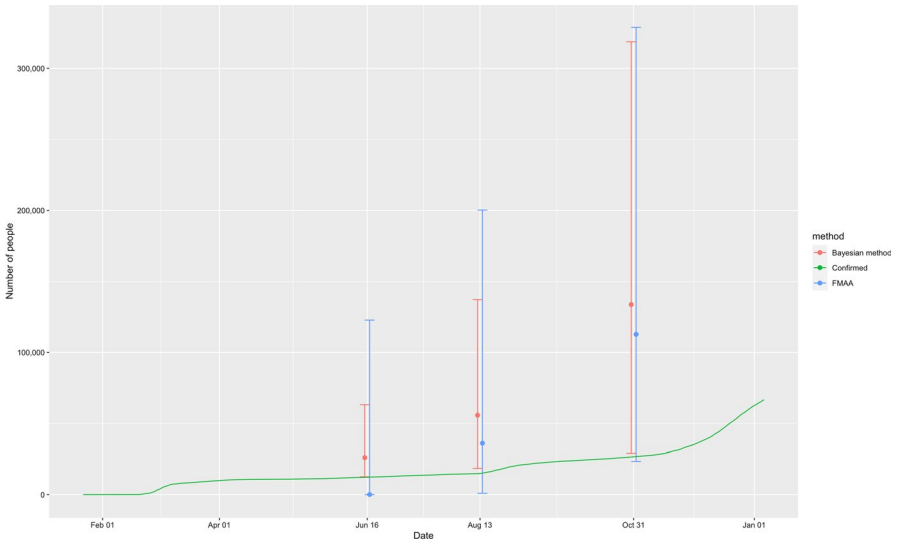

Fig. 3 Dots and error bars denoted by "Bayesian method" represent the posterior mean and 95% credible intervals of the multiplied posterior distributions on seroprevalance parameters by the proposed Bayesian method. Dots and error bars denoted by "FMAA" represent the multiplied point estimators and the multiplied 95% confidence intervals of the results of the frequentist method with the accuracy assumption. The line graph denoted by "Confirmed" represents the daily cumulative confirmed cases

6 Discussion

In this article, we have proposed a Bayesian method with informative prior for the seroprevalence surveys in South Korea, which uses the clinical evaluation result of a serology test, the PRNT, to construct informative prior for the sensitivity and specificity of the serology test. We have compared the proposed method with the two frequentist methods with accuracy and inaccuracy assumptions. With the accuracy assumption, the serology test is 100% correct while with the inaccuracy assumption it is not. The main advantages of the proposed method are two folds. First, the method allows the constrained parameter space, which has an obvious lower bound as the proportion of the cumulative confirmed cases. Second, when we consider the inaccuracy assumption, the method provides interval estimates whose frquentist coverage probabilities are close to the nominal level while the frequentist method with inaccuracy assumption gives empty confidence interval.

The results in this paper has also a limitation. Each set of samples in the seroprevalence survey does not cover all the regions in South Korea. In the first survey announced on the 9th of July, the survey samples do not include those from the populations of several major cities such as Daegu, Daejeon, and Sejong. Daegu particularly was the city of the first mass outbreak in South Korea. The other surveys also do not cover all the cities. The second survey samples do not include those from Ulsan, Busan, Jeonnam, and Jeju, and for the third survey, Gwangju and Jeju are not covered.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42952-021-00131-7.

Acknowledgements Seongil Jo was supported by INHA UNIVERSITY Research Grant, and Jaeyong Lee was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (Nos. 2018R1A2A3074973 and 2020R1A4A1018207).

References

Arora, R. K., Joseph, A., Van Wyk, J., Rocco, S., Atmaja, A., May, E., et al. (2020). SeroTracker: a global SARS-CoV-2 seroprevalence dashboard. *Infectious Diseases: The Lancet*, *21*(4), E75–E76.

Bickel, P. J., & Doksum, K. A. (2015). *Mathematical statistics: Basic ideas and selected topics* (Vol. 117). CRC Press.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., et al. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1), 1–32.

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

Diggle, P. J. (2011). Estimating prevalence using an imperfect test. *Epidemiology Research International, 2011*, 608719.

Gorbalenya, A., Baker, S., Baric, R., de Groot, R., Drosten, C., Gulyaeva, A., et al. (2020). The species severe acute respiratory syndrome related coronavirus: Classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology, 5,* 536–544.

Kohmer, N., Westhaus, S., Rühl, C., Ciesek, S., & Rabenau, H. F. (2020). Brief clinical evaluation of six high-throughput SARS-CoV-2 IgG antibody assays. *Journal of Clinical Virology, 129,* 104480.

Korea Disease Control and Prevention Agency (2021). http://www.kdca.go.kr/. Accessed 7 Jan 2021

Kwon, S., Kim, Y., Jang, M., Kim, Y., Kim, K., Choi, S., et al. (2014). Data resource profile: The Korea National Health and Nutrition Examination Survey (KNHANES). *International Journal of Epidemiology, 43*(1), 67–77.

Ministry of the Interior and Safety (2021). *Demographics of Resident registration*. www.mois.go.kr. Accessed 7 Jan 2021.

Noh, J. Y., Seo, Y. B., Yoon, J. G., Seong, H., Hyun, H., Lee, J., et al. (2020). Seroprevalence of anti-SARS- CoV-2 antibodies among outpatients in southwestern Seoul, Korea. *Journal of Korean medical science*, *35*(33), e311

Rao, C. R. (1948). *Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation, Mathematical Proceedings of the Cambridge Philosophical Society* (Vol. 44, pp. 50–57). Cambridge University Press.

Silveira, M. F., Barros, A. J., Horta, B. L., Pellanda, L. C., Victora, G. D., Dellagostin, O. A., et al. (2020). Population-based surveys of antibodies against SARS-CoV-2 in Southern Brazil. *Nature Medicine, 26*(8), 1196–1199.

Song, S.-K., Lee, D.-H., Nam, J.-H., Kim, K.-T., Do, J.-S., Kang, D.-W., et al. (2020). IgG seroprevalence of COVID-19 among individuals without a history of the coronavirus disease infection in Daegu, Korea. *Journal of Korean medical science*, *35*(29), e269.

The World Health Organization. (2020a). *Novel coronavirus (2019-nCoV): situation report, 22*. The World Health Organization.

The World Health Organization. (2020b). *Population-based age-stratified seroepidemiological investigation protocol for coronavirus 2019 (COVID-19) infection, 26 May 2020*. Technical report: World Health Organization.

Yang, R., & Berger, J. O. (1996). *A catalog of noninformative priors. Institute of Statistics and Decision Sciences*. Duke University.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., et al. (2020). A novel coronavirus from patients with pneumonia in China, 2019. *New England Journal of Medicine*, *382*(8), 727–733.