Statistical Appendix

In this appendix, we describe our statistical methods in more detail. Section 1 describes our approach to calculating population weights. Section 2 describes our approach to adjusting our population prevalence estimate for the sensitivity and specificity properties of the LFA test kit we are using. Finally, Section 3 describes our approach to incorporating three separate sources of uncertainty in our prevalence estimates: sampling variability, error in the sensitivity estimate, and error in the specificity estimate.

Population weighting

In all but the unadjusted prevalence results, we reweight our sample to reflect the sex, race/ethnicity, and zip code of residence distribution of Santa Clara County. We derive these weights from the 2018 American Community Survey, from which we derived an estimate of the population of each zip code, as well as the race/ethnicity and sex distribution of county residents. We applied the county-wide sex distribution to each race/ethnicity group in each zip code to estimate the number of people within each zip-race-sex group. For example, zip code 95037 has a total of 51,652 residents, of which 50.6% are female, 49.7% are white, 7.7% are Asian, 33.0% are Hispanic, and 9.5% are other. We applied the female proportion to each race category in the zip code to obtain the number of residents in each zip-race-sex group.

Let $Epop_{zip,race,sex}$ be the number of people in each zip-race-sex cell produced by this calculation. Let $SmpSz_{zip,race,sex}$ be the number of people in our sample population in each cell. We need to up-weight cells that are underrepresented in our population relative to their frequency in Santa Clara County, and down-weight cells that are overrepresented. We can accomplish this by weighting proportional to the following ratio:

$$\tilde{\theta}_{zip,race,sex} = \frac{Epop_{zip,race,sex}}{SmpSz_{zip,race,sex}}$$

We renormalize so that the sum of our weights equals the size of the SCC sample, N. Define

$$S \equiv \sum_{zip,race,sex} \tilde{\theta}_{zip,race,sex}.$$

Our final sample weights are:

$$\theta_{zip,race,sex} = \frac{N}{S}\tilde{\theta}_{zip,race,sex}$$

This is identical to the formula we present in the Methods:

$$weight_{zsr} = \frac{\frac{N^C_{zsr}}{N^C_{total}}}{\frac{N^S_{zsr}}{N^S_{total}}}$$

Adjusting the prevalence estimate for test kit accuracy

Our main goal is to derive an estimate of the population prevalence of specific COVID-19 antibody seroconversion in Santa Clara County. However, we observe a noisy signal of antibody presence because the test kit we use has both type 1 and type 2 errors. In this section, we describe our approach to adjusting for these errors. To spare notation, we do not incorporate the sample weighting process we describe

above. Introducing sample weighting would complicate our notation, but would not change the approach. The analytic weights were used in the results shown in Table 2.

Let $\pi = P(COVID +)$ represent the population prevalence of antibodies to COVID-19, and let $q = P(TEST +)$ be the proportion of participants who test positive in our sample (this latter quantity measured using our sample weights). Note: we consider $TEST +$ as any band on the test kit indicating the presence of IgG or IgM antibodies or both.

Let $r = P(TEST + \mid COVID +)$ be the sensitivity of the test and let $s = P(TEST - \mid COVID -)$ be the specificity of the test. Let $z = P(COVID + \mid TEST +)$ be the positive predictive value of the test, and $y = P(COVID + \mid TEST -)$ be the (one minus) the negative predictive value of the test.

By Bayes rule,

$$P(COVID + \mid TEST +) = \frac{P(TEST + \mid COVID +)P(COVID +)}{P(TEST + \mid COVID +)P(COVID +) + P(TEST + \mid COVID -)P(COVID -)}$$

and

$$P(COVID + \mid TEST -) = \frac{P(TEST - \mid COVID +)P(COVID +)}{P(TEST - \mid COVID +)P(COVID +) + P(TEST - \mid COVID -)P(COVID -)}.$$

Rewriting these in our notation, we have:

$$z = \frac{r\pi}{r\pi + (1-s)(1-\pi)}, \text{and}$$
$$y = \frac{(1-r)\pi}{(1-r)\pi + s(1-\pi)}.$$

By the definition of conditional probability, we also have:

$$P(COVID +) = P(COVID + \mid TEST +)P(TEST +) + P(COVID + \mid TEST -)P(TEST -),$$

Or in our notation:

$$\pi = zq + y(1-q).$$

If we plug in our expressions for y and z and simplify, we have a quadratic expression in π.

$$1 = \frac{rq}{r\pi + (1-s)(1-\pi)} + \frac{(1-q)(1-r)}{(1-r)\pi + s(1-\pi)}.$$

We solve for π as a function of the sample prevalence, sensitivity, and specificity:

$$\pi = \frac{q + s - 1}{r + s - 1}.$$

There is one important caveat to this formula: it only holds as long as (one minus) the specificity of the test is higher than the sample prevalence. If it is lower, all the observed positives in the sample could be due to false-positive test results, and we cannot exclude zero prevalence as a possibility. As long as the specificity is high relative to the sample prevalence, this expression allows us to recover population prevalence from sample prevalence, despite using a noisy test.

Delta method approach to variance measurement

In this section, we derive an estimate of the variance of our population prevalence estimate that accounts for the error in the test kit sensitivity and specificity numbers, as well as sampling variation. We retain our notation from the previous section, and start with the formula we derived that relates all our variables of interest:

$$\pi = \frac{q + s - 1}{r + s - 1}.$$

We use a first order Taylor series approximation to derive the delta method approximation for $Var(\pi)$.

$$\pi \approx \left.\frac{\partial \pi}{\partial q}\right|_{(q_0, r_0, s_0)} (q - q_0) + \left.\frac{\partial \pi}{\partial r}\right|_{(q_0, r_0, s_0)} (r - r_0) + \left.\frac{\partial \pi}{\partial s}\right|_{(q_0, r_0, s_0)} (s - s_0).$$

The three relevant derivatives are as follows:

$$\left.\frac{\partial \pi}{\partial q}\right|_{(q_0, r_0, s_0)} = \frac{1}{r_0 + s_0 - 1},$$

$$\left.\frac{\partial \pi}{\partial r}\right|_{(q_0, r_0, s_0)} = \frac{q_0 + s_0 - 1}{(r_0 + s_0 - 1)^2}, \text{ and}$$

$$\left.\frac{\partial \pi}{\partial s}\right|_{(q_0, r_0, s_0)} = \frac{r_0 - q_0}{(r_0 + s_0 - 1)^2}.$$

Since q, s, and r are all derived from independent samples:

$$Var(\pi) \approx \left(\frac{1}{r_0 + s_0 - 1}\right)^2 Var(q) + \left(\frac{q_0 + s_0 - 1}{(r_0 + s_0 - 1)^2}\right)^2 Var(r) + \left(\frac{r_0 - q_0}{(r_0 + s_0 - 1)^2}\right)^2 Var(s).$$

Thus, the delta method estimate for $Var(\pi)$ is a weighted sum of three variance terms: sampling variability, sampling error in sensitivity, and sampling error is specificity.

In the main text, we explore three alternative assumptions about the sampling error in sensitivity and specificity. For each of these scenarios, we estimate the variance in sensitivity and specificity by the standard formulas for a binomial outcome: $Var(r) = \hat{r}(1 - \hat{r})$ and $Var(s) = \hat{s}(1 - \hat{s})$, where $\hat{r}$ and $\hat{s}$ are the estimated values of sensitivity and specificity that pertain to each scenario.

- In the first scenario, we estimate these quantities based upon the numbers provided by the manufacturer of the test kits. For sensitivity, the manufacturer reported 78 positive test readings for 85 samples (from Chinese blood samples) known to have specific IgM antibodies to the receptor-binding domain (RBD) spike on the SARS-nCOV2 virus. They reported 75 positive test readings for 75 of the samples with specific IgG antibodies to the same RBD spike. We adopt a conservative estimate of sensitivity equal to $\hat{r} = \frac{78}{85} \approx 91.8\%$. The manufacturer reports specificity based on an experiment using their kit on a sample of 371 known negative blood samples collected from before the epidemic, and 369 were tested negative. This implies a specificity of $\hat{s} = \frac{369}{371} \approx 99.5\%$.
- In the second scenario, we estimate these quantities based on tests run locally at Stanford University. We identified serum from 37 patients who had RT-PCR-confirmed cases of COVID-19 and either IgG or IgM on a locally-developed ELISA assay; of these, 25 tested positive with the test kit, implying a sensitivity of $\hat{r} = \frac{25}{37} \approx 67.6\%$. We also identify serum from 30 patients drawn from before the COVID-19 epidemic, and all 30 tested negative, implying a specificity of

$\hat{s} = \frac{30}{30} = 100\%$. In this case, since the standard formula implies $Var(s) = 0$, we instead estimate $Var(s)$ from a simple transformation of the width of the Clopper-Pearson exact confidence interval.

- In the third scenario, we estimate these quantities by combining the manufacturer tests with our local tests in a simple additive fashion. Under these assumptions, the sensitivity estimate is $\hat{r} = \frac{103}{122} \approx 84.4\%$ and the specificity estimate Is $\hat{s} = \frac{399}{401} \approx 99.5\%$.

To provide one concrete example, we have the following values under the first scenario:

$q_0 = 0.028$ $\quad$ $var(q_0) = 0.027$
$r_0 = 0.918$ $\quad$ $var(r_0) = 0.076$
$s_0 = 0.995$ $\quad$ $var(s_0) = 0.005$

$\left(\frac{1}{r_0+s_0-1}\right)^2 = 1.202$
$\left(\frac{q_0+s_0-1}{(r_0+s_0-1)^2}\right)^2 = 7.423 \times 10^{-4}$
$\left(\frac{r_0-q_0}{(r_0+s_0-1)^2}\right)^2 = 1.143$

Combining all of those values, we get $Var(\pi) = 0.039$, and $SE(\pi) = 0.0034$
We use this procedure throughout.