Estimation through Sample Pooling under Constrained Testing Conditions for COVID-19

A White Paper for Health Researchers and Policy Makers

By James Rabe
For Stanford MS&E 433
Advised by Ashish Goel and Anup Malani[1]

Summary:

Obtaining an accurate estimate of the prevalence of Covid-19 is essential for policy makers. Yet in many developing countries essential testing capacity is stretched thin. This white paper draws on previous work on sample pooling and recommends a logistically simple strategy for estimating the prevalence of Covid-19 in a population. Both analytical derivation and simulations are used to confirm the accuracy of the estimator for the proposed sample pooling strategy. A full version of this paper is available on the Stanford MS&E 433 class website.

COVID-19 has spread alarmingly fast and wide during the early months of 2020. However, in many countries the extent of the spread is largely unknown due to shortages of testing capacity, hampering effective treatment and clouding policy decisions. It is therefore essential that governments and NGOs accurately and efficiently test populations for the disease. This problem is particularly acute in the developing world where testing personnel and resources are stretched thin (Biswas).

Current best practice for testing for a rare disease utilizes sample pooling, wherein biological samples are taken from a certain number of individuals and mixed together into one unit, called a pool. This pool is then tested for the disease, indirectly testing the group of individuals in the pool. Assuming that the disease is rare enough, most pools will be compromised of only non-infected individuals; if the pool tests negative, multiple individuals have efficiently been tested with a single test. If a pool tests positive, then at least one individual in the group is positive, and each individual that contributed to the pool is tested. The population infection rate is the estimated as the number of positive individuals over the total number of individuals covered in the pools. In the current health crisis sample pooling is being used for identifying in several countries, including Israel (HospiMedica International), Germany (Seifried, Cieseket, et al.), and the United States (Hogan, Sahoo, Pinsky).

Sample pooling has been utilized since World War II and since then variations have been proposed to make sample pooling more efficient. One variation, double pooling, decreases the number of total tests used by up to 30% through pooling two samples from each individual into different overlapping pools and testing only individuals whose samples were both in positive pools (Broder and Kumar). However, while many proposed variations of sample pooling present

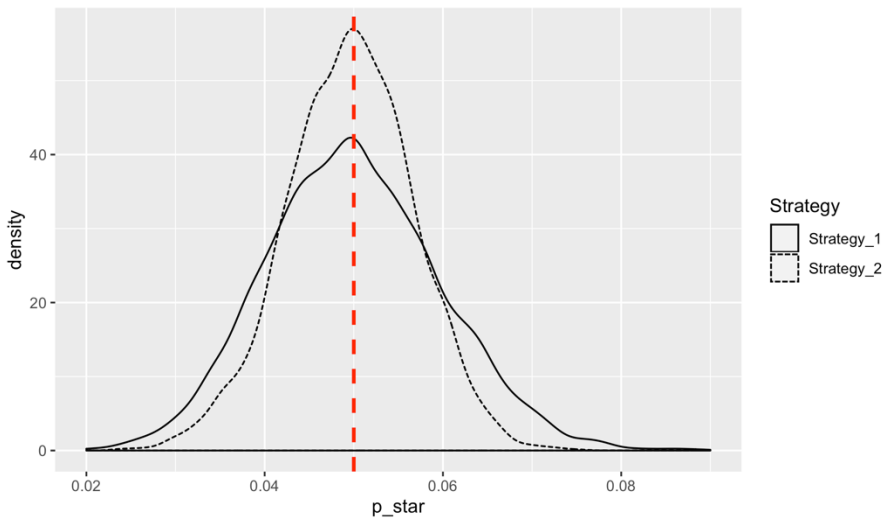
[1] We would like to acknowledge Professor Johan Ugander for introducing us to relevant sample pooling literature, and to Arjun Seshadri for developing a detailed theoretical analysis, which we hope to incorporate in future versions of this report.

mathematically elegant and efficient solutions, they also introduce logistical complications that can make the strategies unattractive. However, while traditionally sample pooling serves two goals (to identify infected individuals and to understand the underlying prevalence of disease in a population) variations that focus only estimating prevalence of a disease have been proposed which significantly reduce the logistical requirements of sample pooling while maintaining a high degree of accuracy (Sobel and Elashoff). If policy makers are concerned mainly with understanding the spread of a disease in their country, these modified sample pooling strategies are attractive and efficient tool.

In this short paper, we investigate an efficient sample pooling strategy that does not require retesting. This strategy proposes pooling individuals' biological samples and testing each pool without retesting individuals. Instead the Maximum Likelihood Estimator (MLE) is used to estimate the population infection rate based solely upon the pool size, the total number of pools tested, and the number of pools that tested positive. This strategy could be an improvement above current practices as it may use fewer tests to arrive at certain variance thresholds and, critically, would obviate the logistical burden on health workers of tracking down and re-sampling individuals from infected pools. The derivation of the MLE is briefly outlined at the end, more details are in the full version of the report.

To confirm that the proposed strategy of not retesting individuals in favor of using the MLE to estimate disease prevalence presents an accurate estimation of disease prevalence, we simulated five thousand populations, each generated randomly with a disease rate of 5%. We then tested traditional sample pooling (strategy 1) against the proposed sample pooling without individual resting (strategy 2) on each population using the same number of tests for each strategy. The resulting graph (Figure 1) indicates that strategy 2 provides a lower variance estimation of the true population infection rate (indicated as the vertical dashed line). Additionally, over the five thousand trials, strategy 2 estimated the true population infection rate better than strategy 1 in 62.7% of trials.

Figure 1: Density of P_star Estimations for Strategies 1 and 2



Sensitivity analysis was performed, varying true infection rate, pool size, and number of tests used. The result was that strategy 2 performed better than strategy 1 in all cases where the population was random, and samples were random (i.e. cases were not clustered together in some pools and sparser in others). One caveat is that both strategies can suffer from small sample bias. This means that in both strategies, enough

tests need to be used so that the probability of seeing multiple positive individuals is large. Simulation revealed that when the true disease prevalence is less than 5% at least two hundred tests must be used and when the true disease prevalence is less than 1% at least one thousand tests must be used.

Thus, our analysis indicates that sampling without retesting and instead using the MLE to estimate disease prevalence in a population is at least as accurate as traditional sample pooling. Combined with the logistical ease of use (no retesting of individuals), we find that this proposed strategy presents an attractive way for policy makers to estimate the spread of a disease in a population.

The MLE can be derived from a likelihood function where:

p be the true infection rate in the population
j be the pool size used
k be the number of pools tested
m be the number of pools that tested positive

$$Likelihood\ (p, k, j, m) = ((1-p)^j)^{k-m} * (1-(1-p)^j)^m$$

Resulting in the Maximum Likelihood Estimator of:

$$p = 1 - \sqrt[j]{\frac{k-m}{k}}$$

Works Cited

Biswas, Soutik. "Coronavirus: Why Is India Testing so Little?" BBC News, BBC, 20 Mar. 2020, www.bbc.com/news/world-asia-india-51922204.

Broder, Andrei Z, and Ravi Kumar. A Note on Double Pooling Tests (Preliminary Version). Google Mountain View, CA, arxiv.org/pdf/2004.01684.pdf.

MILTON SOBEL, R. M. ELASHOFF, Group testing with a new goal, estimation, Biometrika, Volume 62, Issue 1, April 1975, Pages 181–193, https://doi.org/10.1093/biomet/62.1.181

Erhard Seifried, Sandra Ciesek, and et al. Pool testing of SARS-CoV-2 samples increases test capacity. https://www.medica.de/de/News/Redaktionelle_News/ Pool-Testen_von_SARS-CoV-2_Proben_erhoht_Testkapazitat.

Hogan CA, Sahoo MK, Pinsky BA. Sample Pooling as a Strategy to Detect Community Transmission of SARS-CoV-2. JAMA. Published online April 06, 2020. doi:10.1001/jama.2020.5445

HospiMedica International. "Israeli Researchers Introduce Pooling Method for COVID-19 Testing of Over 60 Patients Simultaneously." Hospimedica.com, HospiMedica International, 24 Mar. 2020, www.hospimedica.com/coronavirus/articles/294781273/israeli-researchers-introduce-pooling-method-for-covid-19-testing-of-over-60-patients-simultaneously.html.