Estimation of Covid-19 Prevalence from Serology Tests: A Partial Identification Approach

Panagiotis (Panos) Toulis
panos.toulis@chicagobooth.edu

Econometrics and Statistics
University of Chicago, Booth School of Business

Serology tests for Covid-19

Estimating prevalence of Covid-19 gives crucial information on:

-lethality of disease; -policy design.

Serology tests detect antibodies in response to current or past Covid-19 infection. Currently in (hyper)active research phase.

Existing serology studies:

Germany: 14% prevalence in a hard-hit town.

Netherlands: 3.5% prevalence in sample of blood donors.

USA: 2-4% in Santa Clara study & LA county; 14% in NY State study.

Unfortunately, even the highest numbers are not high enough.

The “controversial” Santa Clara study

Several criticisms: high false positive rate (FPR), biased sample, etc.

Discussion got heated, mainly due to politics (lift or maintain quarantine?)

The statistical methods used to analyze the data (from both sides) are not the best. They rely on either normality approximations or bootstrap calculations.

Example argument 1: Study found 2/401 false positives in “calibration study” and 50/3330 positives in “main study”. Estimate of FPR is $\hat{p} = 2/401 = 0.5\%$ with 95% CI: $[-0.2\%, 1.2\%]$. This could imply $3330 \times 0.012 = 40$ false positives in main study out of 50 observed.

Example argument 2: Estimate $\hat{p}$ and $\hat{q}$ (true positive rate). Run parametric bootstrap to obtain prevalence estimates from $[0\%, 1.86\%]$ (truncated at zero).

A better approach: Partial Identification

Our data are:

$$\begin{aligned} S_0 &= \text{positives in calibration study} && (s_0 = 2); \\ S_1 &= \text{positives in main study} && (s_1 = 50). \end{aligned} \tag{1}$$

Unknown params.: $\theta = (p, q, \pi) =$ (FPR, TPR, prevalence), where

$$\pi = \#\text{true positives}/3330.$$

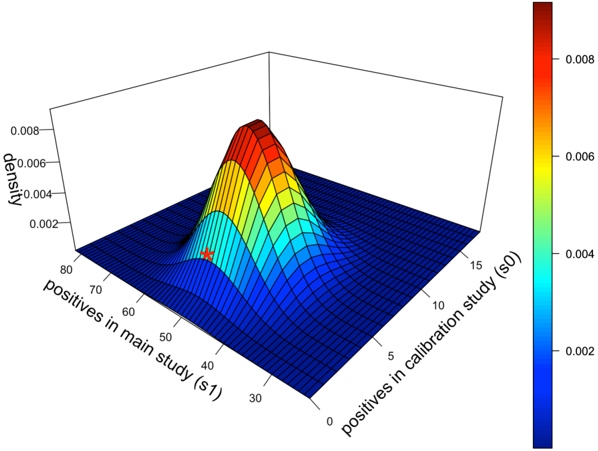
Key argument: for any given θ we can calculate *exactly* the density $f(S_0, S_1|\theta)$. Build confidence set as:

$$\widehat{\Theta} = \{\theta \in \Theta : f(s_0, s_1|\theta) > c_\theta\}.$$

Choose c_θ such that:

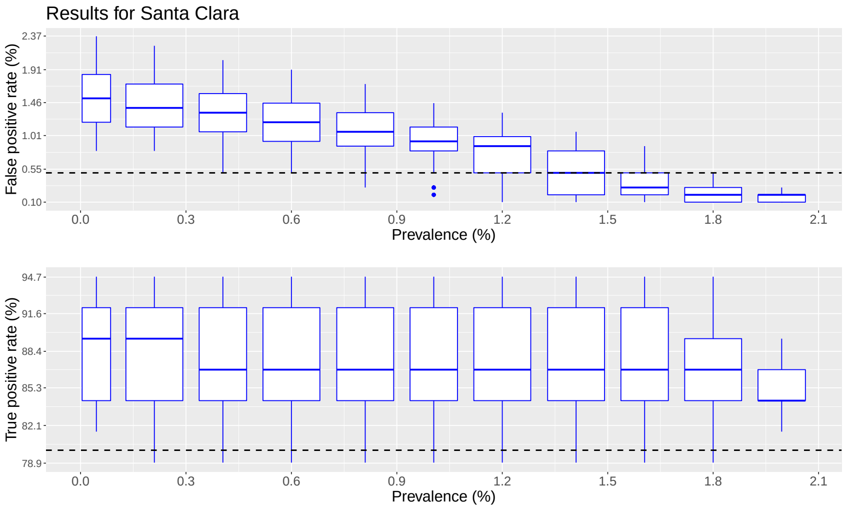
$$P(\theta_0 \in \widehat{\Theta}) \geq 1 - \alpha.$$

Suppose $\theta_0 = (p, q, \pi) = (0.015, 1, 0)$. Then, $f(S_0, S_1|\theta_0)$ looks as follows:

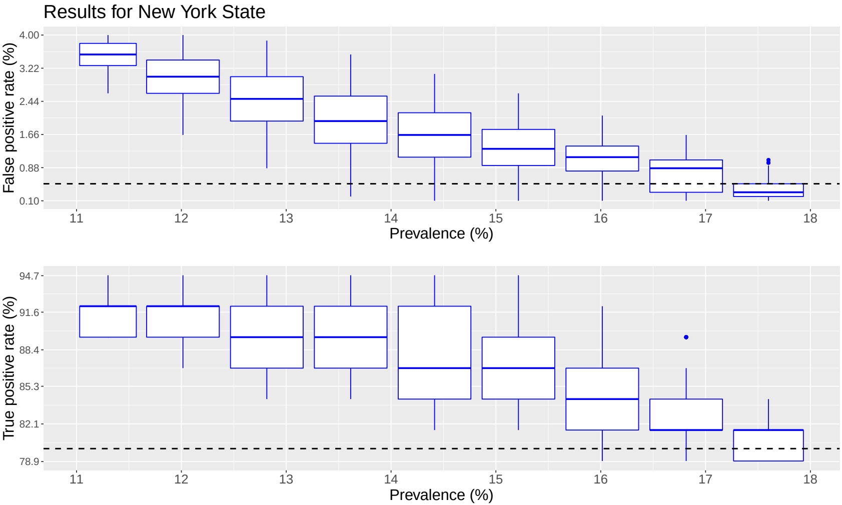


Santa Clara study. *Data: (s0, s1)=(2, 50)

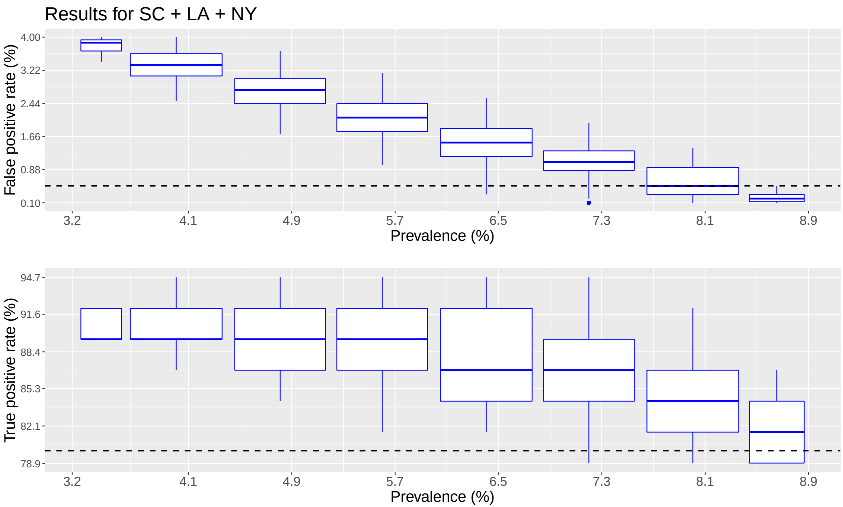
Check value of $f(s_0, s_1|\theta_0)$ to decide whether to include θ_0 in $\widehat{\Theta}$.



Here, we visualize (p, q, π) in $\widehat{\Theta}$; $\pi = 0\%$ is included; [0.6-1.8%] is more plausible; FPR is crucial for sharp identification; TPR is not.



In NY study, $\pi \in [11\%, 18\%]$; TPR is important here.



With all data combined, $\pi \in [3\%, 9\%]$.