

Lab Assignment 6

STA 100 | A. Farris | Spring 2020

*A pdf copy of your assignment is due at the end of the day (11:59pm PT) Tuesday, May 26. Submission of the pdf will be through Canvas. Please put in the effort to make it look reasonably professional – you’re encouraged to use R Markdown. Note that specific tasks for you are **highlighted**. You are free to work in groups, but you must submit your own writeup, and run your own code.*

The Stanford seroprevalence study

As of mid-May 2020, a key scientific question with regard to the COVID-19 pandemic is the infection fatality rate (IFR) of the disease. Unlike the case fatality rate (CFR), which measures fatalities as a fraction of detected cases, the IFR measures the proportion of all infections, **whether detected or not**, that result in a fatality.

The IFR measures the cost in lives of the disease as a proportion of infections, and is therefore crucial in measuring the cost of policies that are intended to slow or stop the disease. As policymakers come under political and financial pressure to relax social distancing measures, an understanding of the cost associated with such easing is of paramount importance.

An important quantity in an understanding of the IFR is what is essentially its **denominator**, the number of infections that have taken place. Because in part of the dearth of reliable and systematic testing in most countries, there is substantial uncertainty surrounding this number. Furthermore, this number is of independent interest to public health authorities and policymakers, as it contains valuable information with respect to understanding how far the pandemic is from attaining some form of herd immunity.

In early April, a team of Stanford researchers began to publicize the results of a seroprevalence study that they had performed, in which they estimated the true number of infections that had taken place in Santa Clara County on the basis of a sample of people administered antibody tests. Their results, they reported, suggested that the number of infections was much higher than previously suspected, suggesting the IFR of the disease to be **lower** than previously suspected. The team raised eyebrows among their scientific peers for publicizing their results¹ before they had undergone peer review, and even before they had posted a preprint² of the details of their analysis for comment by the scientific community.

Their preprint attracted substantial interest, both because of the importance of the topic and the publicity efforts on the part of the authors. Quickly, the authors came under criticism from their peers for their methods. In this lab project, we will explore one aspect of this criticism briefly.

Accounting for specificity

In interpreting the results of the antibody tests, it is of course important to factor in the **false positive rate** of the test.

The authors report the results of a validation study by the antibody test’s manufacturer, in which they applied the antibody tests to 371 blood samples predating the Covid-19 pandemic. They report that 369 of these tested negative, suggesting an estimated specificity of $\frac{369}{371} \approx 0.9946$, and an estimated false positive rate of $\frac{2}{371} \approx 0.0054$.

¹Stanford’s Jay Bhattacharya appears on a political talk show to publicize his results, Fox News, April 14, 2020: <https://www.youtube.com/watch?v=6NjCitwKJSQ>

²Bendavid, Eran, et al. *COVID-19 Antibody Seroprevalence in Santa Clara County, California*. MedRxiv Preprint (April 17, 2020) <https://www.medrxiv.org/content/10.1101/2020.04.14.20062463v1>

From the Stanford study, the authors report that out of 3,330 patients, they found 50 to test positive, for an estimated prevalence (i.e. population proportion of infected people) of $\frac{50}{3330} \approx 0.015$. However, this estimate does not in itself account for the possibility of false positives.

Let's put their estimate to the test. Do we have conclusive evidence that the Stanford study found *any* true positives? The authors are certainly claiming more than this, but this is at least a first hurdle in their burden of proof.

One way to think about this is that we want to carry out a hypothesis test of the Null hypothesis that the population proportion with false positives in the manufacturer study was actually the same as that in the Stanford study.

We have available to us a good test for evaluating whether an unknown proportion is equal to a given value, the Binomial Test. Can we use the Binomial test to evaluate whether the proportions of positive test results were the same in the manufacturers' as in the Stanford study?

No, we can't, because we don't know the true proportion in either case! In both cases, we have a sample quantity, and do not know the population quantity.

Fortunately, we have another tool that will do the trick. Suppose that the true proportion of positives is the same for patients in the Stanford study as it is for those in either of the two studies. We can then say that, for a randomly selected patient from both studies put together, the event that they test positive is independent of the event that they participated in the Stanford study. Therefore, if we organize the results into a contingency table:

	Manufacturer	Stanford
Negative	369	3280
Positive	2	50

then a test for association between these two factors suffices.

With the data in the contingency table above, carry out a test for independence between testing positive and participation in the Stanford study, using significance level 0.05. To do so, state the Null and alternative hypotheses and report the p-value. Do you find sufficient evidence to conclude that the Stanford study shows a higher chance of testing positive than the manufacturer's test did? Interpret this result briefly in your own words. Is your result consistent with or in contradiction to the Stanford study's stated results?

Postscript

The authors changed their work to accommodate some criticisms in a revision of their preprint on April 30³. They revised downward their estimates of the number infected slightly, but their scientific colleagues were not entirely satisfied by the revisions⁴⁵. At time of writing, the peer review process continues.

³Revised preprint, April 30, 2020: <https://www.medrxiv.org/content/10.1101/2020.04.14.20062463v2>

⁴https://stanford.zoom.us/rec/play/uclscce-r_TM3E4WduQSDCvZ6W47oeP6s1CNKrqcIzxuyB3YFOWagNONHYOOmQfdem9liCP2Uxclk_hhI?continueMode=true&_x_zm_rtaid=APFnyzyqRNuVVHvZI9ccEA.1589747689765.7e22018ae11ba4104bca3921248ece82&_x_zm_rtaid=41 , link via <https://bmir.stanford.edu/education/colloquia.html>

⁵Expanding on Trevor Hastie's comments in the Zoom meeting above: <https://www.stat.berkeley.edu/~wfithian/overdispersionSimple.html> .

Appendix: R Script

```
library(knitr)
SeroprevalenceTable <- matrix(c(369,2,3280,50),2,2)
rownames(SeroprevalenceTable) <- c("Negative","Positive")
colnames(SeroprevalenceTable) <- c("Manufacturer", "Stanford")
kable(SeroprevalenceTable)
```