Thoughts on the Bangladesh Mask Paper

I think as far as talking about t-tests, z-scores, and p-values and the effects of cluster randomization, the blog posts and paper by Ben do a better job than I'd hope to do. I'd use those as the gold standard at evaluating this paper.

Relative risk is more informative than effectiveness. – arg min blog

The cult of statistical significance and the Bangladesh Mask RCT. – arg min blog

What were the effects of the Bangladesh mask intervention? – arg min blog

https://arxiv.org/pdf/2112.01296.pdf

Basic Idea of the Bangladesh Study

The folks who did the Bangladesh study took about 300,000 people from about 600 villages, randomized them into groups that got no treatment (Control), surgical masks (Green or Blue masks), or cloth masks (Purple or Red). I'm not sure what the color change was measuring (cultural attitudes towards color?) or if that's what they used to identify compliance later. The randomization was on the village level. That is, an entire viliage was chosen to be in one of the groups, not individuals within the village (this will be important later).

For the next 9 weeks they had folks monitor mask compliance in these villages and perform two surveys (at the midpoint and end of study) to collect health data (ask if they had COVID symptoms as defined by the WHO). After the last survey, they asked anyone who had symptoms to do a blood test to look for COVID antibodies.

The goal was to compare the prevalence of antibodies in the population and deduce if mask wearing had an effect. The actual results are "does mask wearing have an effect on the fraction of population who have symptoms and subsequently test positve for COVID antibodies." It doesn't answer the question "does mask wearing reduce the prevalance of COVID in the population" which I think is the better thing to be looking at.

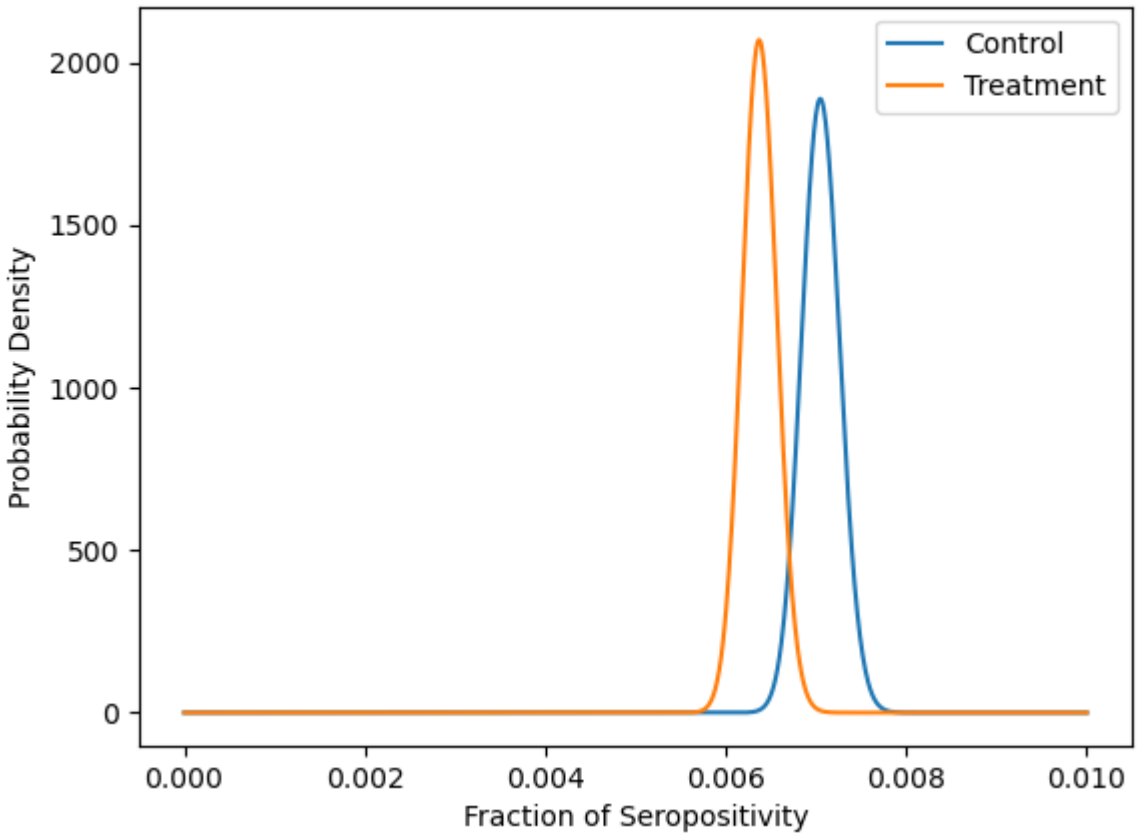
Basic Results

As I said, Ben does a much deeper dive into this in the above links, and it would be a waste of time for me to reproduce them here. I do want to draw some figures to help get an more intuitive idea

about what's happening here. There's no need to get into a t-test, z-score, p-value fight. Looking at probability distributions can give you a gut feeling to understand what's going on here.

Below I plot probability distributions based on a beta distribution. For binary results (postive or negative on COVID antibodies), the beta distribution is a natural choice and better than the Normal Gaussian model they use. In this first plot, we show the results of "all" of the data. That is the treatment group is anyone who got a mask and the control group are the "paired" villages that did not get any masks.

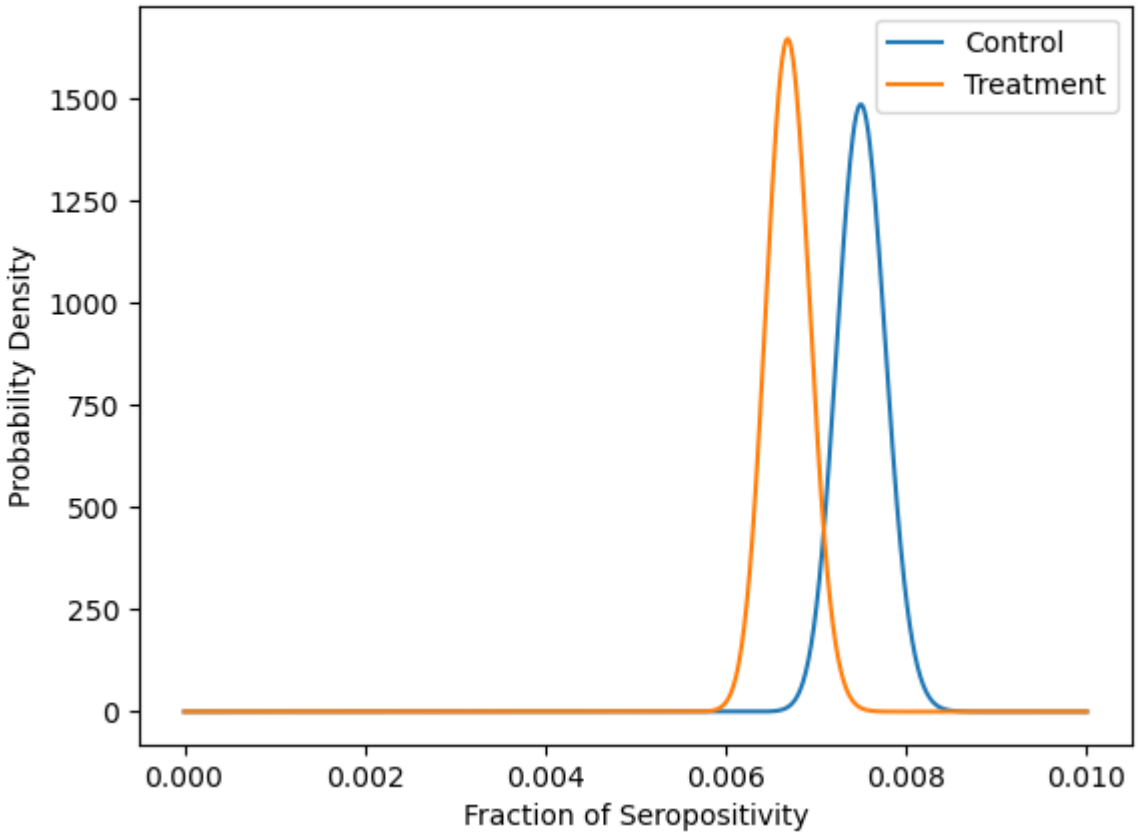
All Treatments



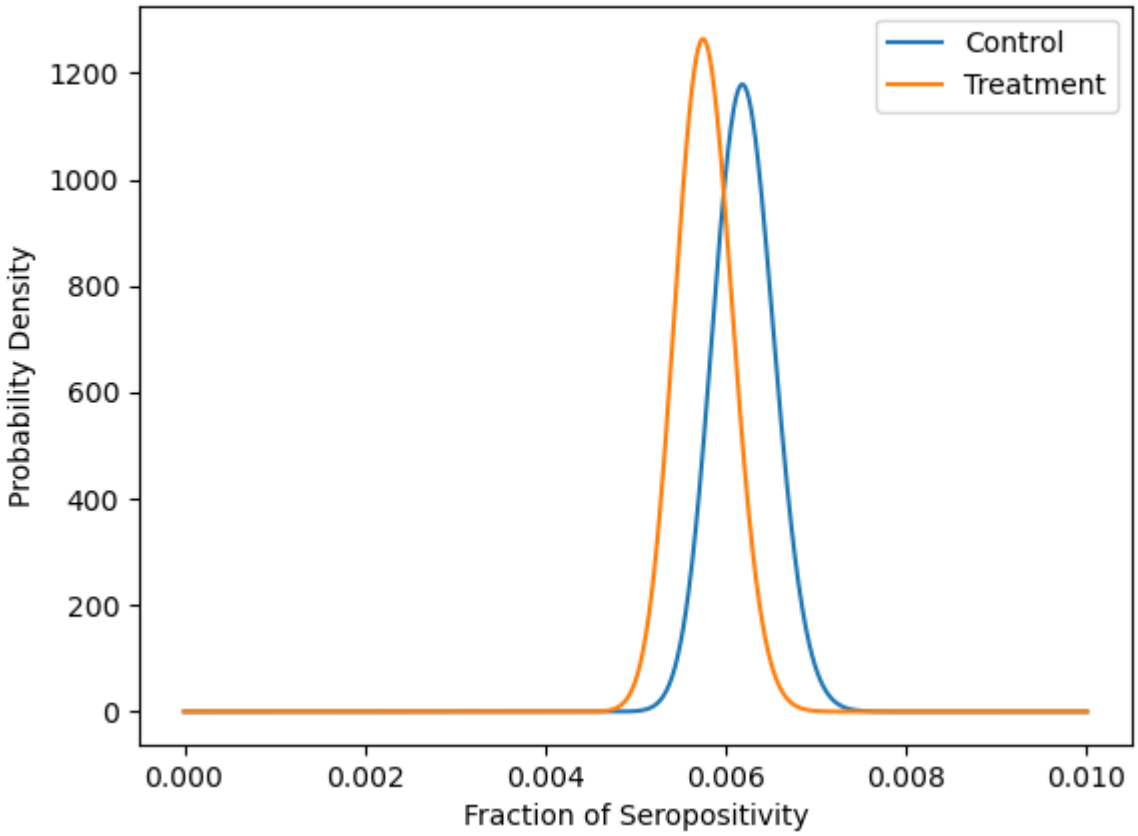
Looking at this, you can make the argument that there is a "statistical" difference since the distributions don't overlap much. But the actual difference is maybe 1 person in a 1000 (6 in a 1000 with a mask vs 7 in a 1000 without) that contracted COVID because the *entire village* was given masks.

We can also look at the same plot for specific treatments.

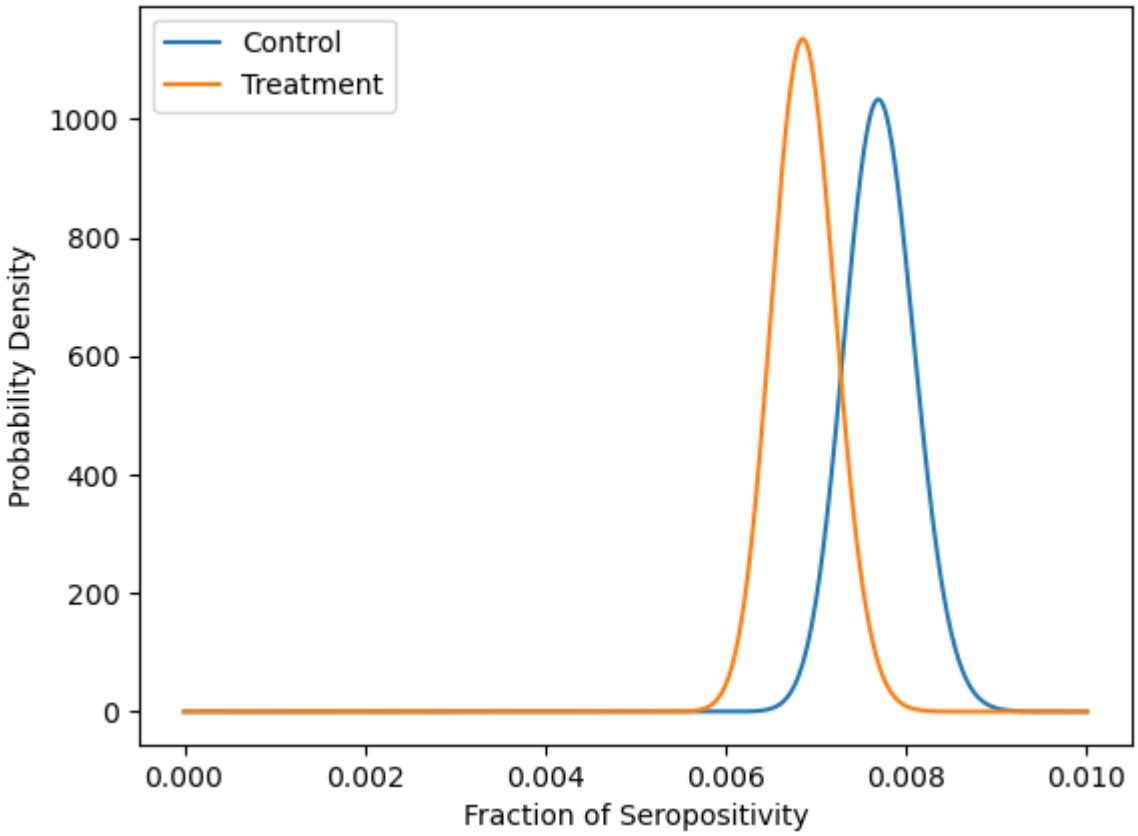
Treatment: All Surgical Masks



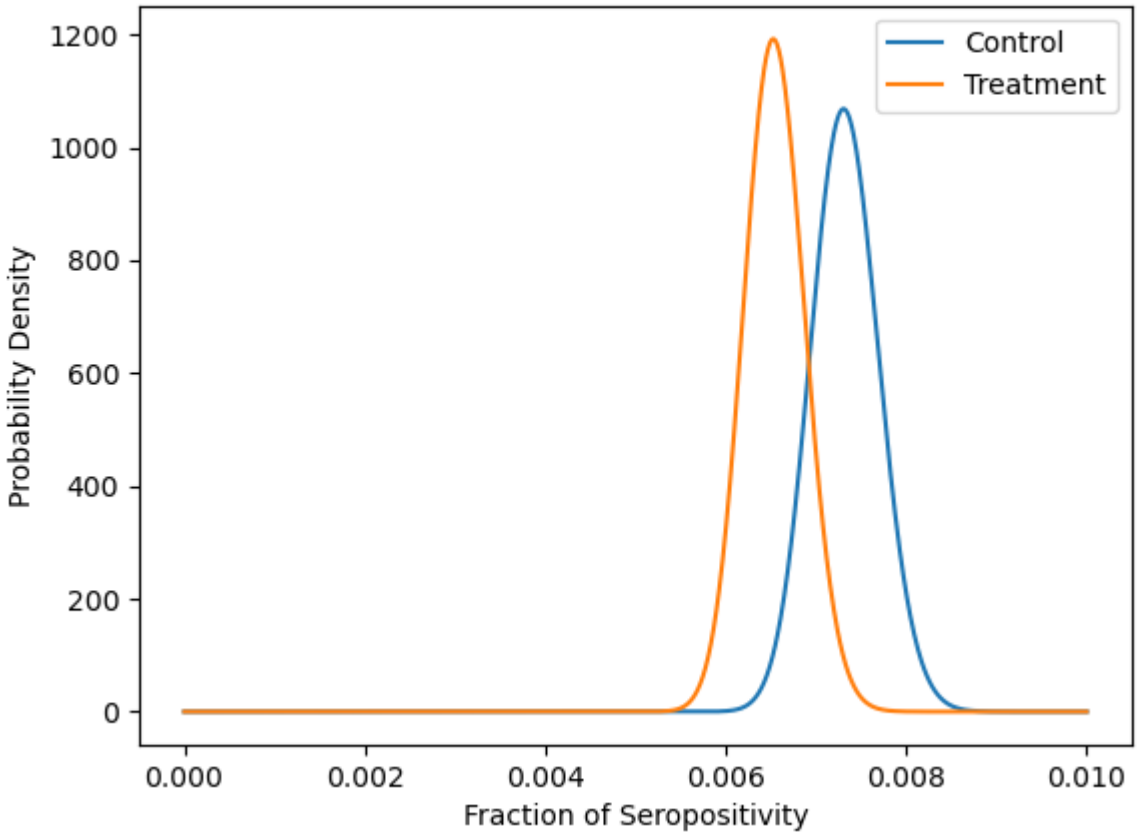
Treatment: All Cloth Masks



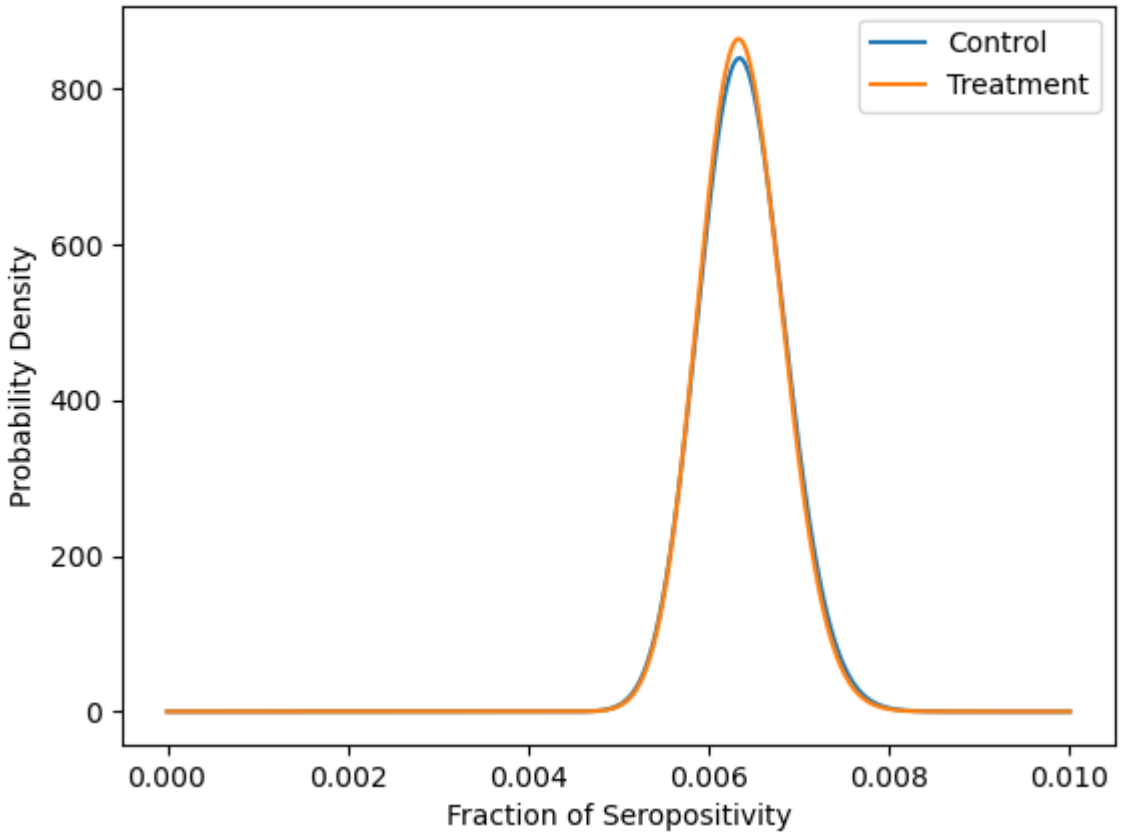
Treatment: Green Surgical Masks



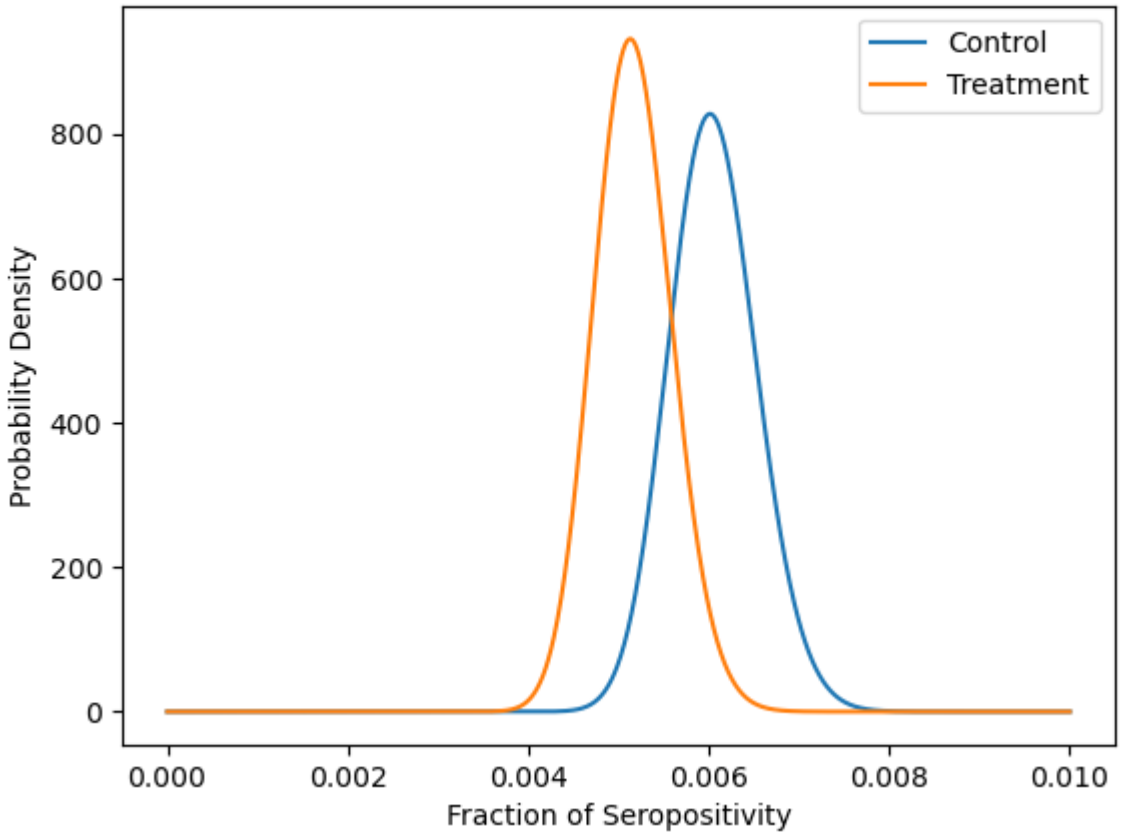
Treatment: Blue Surgical Masks



Treatment: Purple Cloth Masks



Treatment: Red Cloth Masks



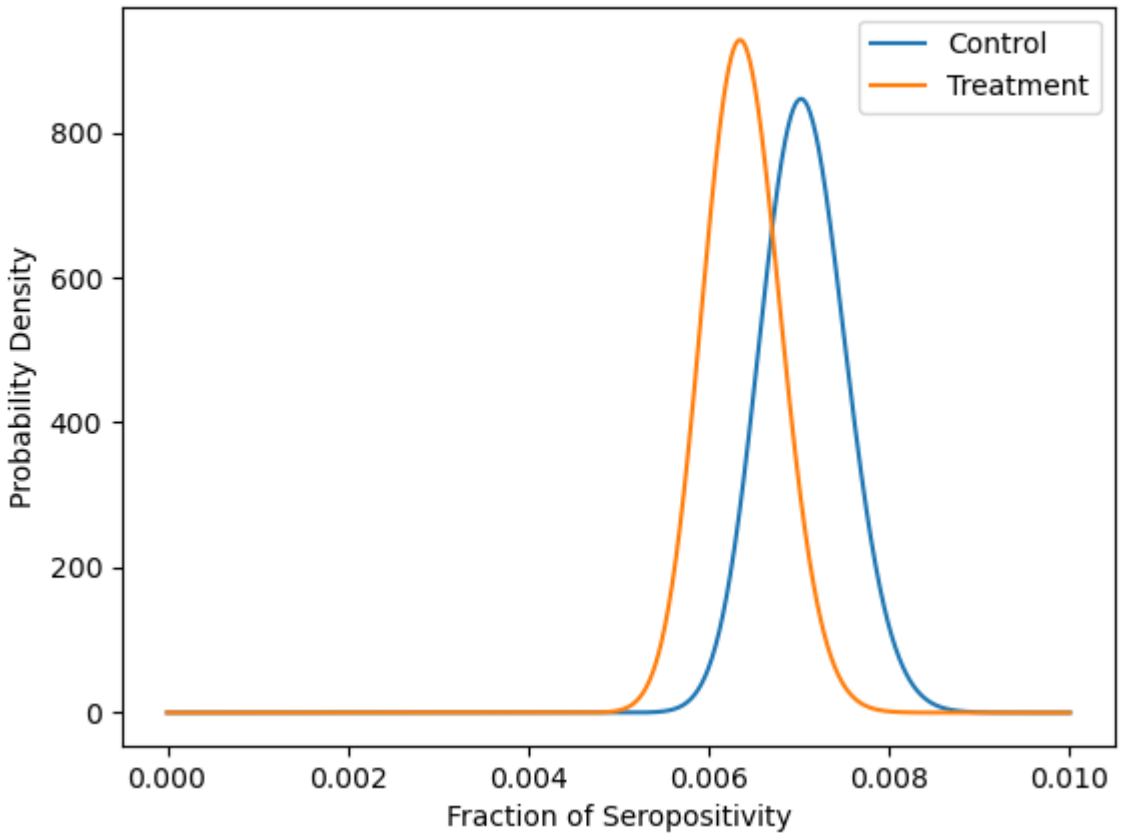
It appears in all subsets except purple that there is some (statistical) effect. But still, a negligible effect size. And for some reason, purple masks don't work. I'll keep that in mind next time I'm buying masks. Seriously though, my interpretation on color changes is that it gives us an idea of the underlying noise because the material should behave the same (unless there's some wierd cultural obsession with purple that I don't know about). Looking at the spread between mask colors gives you a better idea of how much uncerainty there is in the study that is not captured by the model.

Cluster Randomization Issues

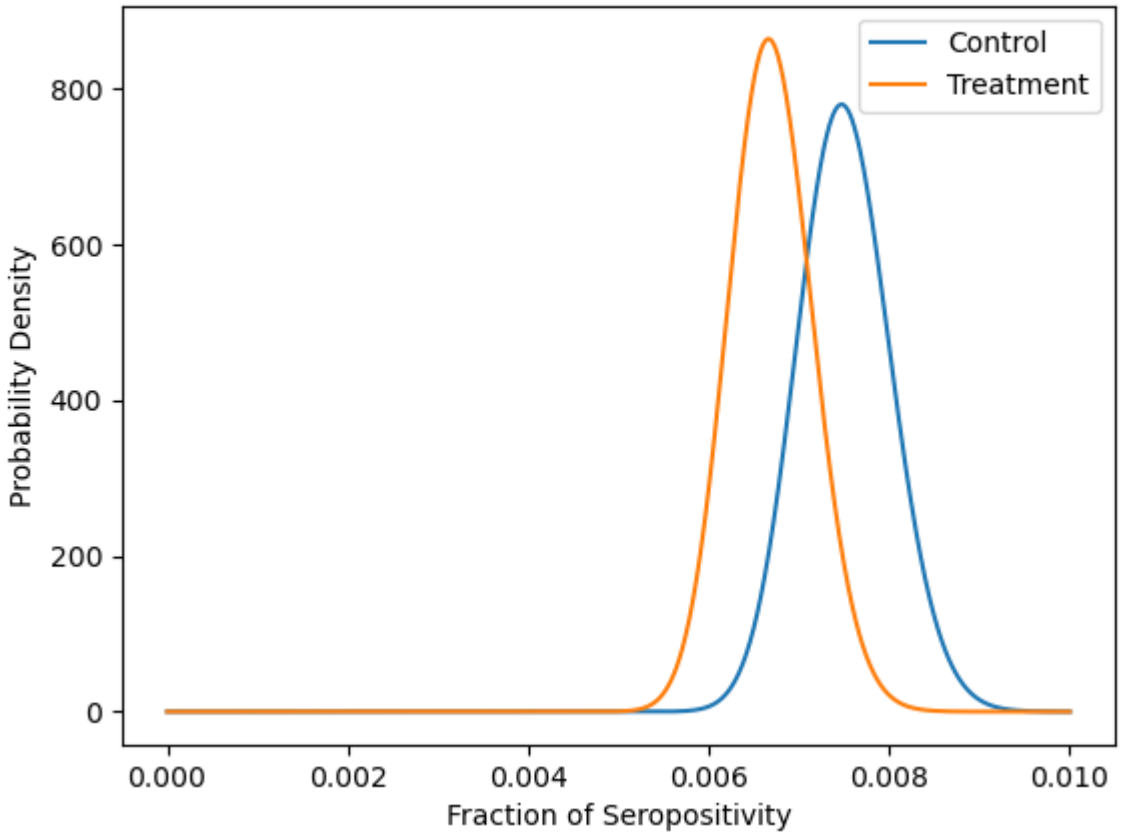
As Ben talks about in his blog post, there are issues with the analysis because of cluster randomization. That is, rather than randomizing the treatment among individuals, they randomize the treatment among villages. That makes sense because you need everyone in the village doing something to measure an effect. But because of that clustering, the people in the village are not independent: there is a correlation within a village. We account for that with the *Design Effect (DE)* that reduces the sample size. Ben describes how this works in his blog post, and we follow the same thing here by including the DE with the beta distribution. This has the effect of spreading out the

distribution, making the results less statistically significant. I have a hard time looking at any of these plots and saying that the treatment is statistically significant.

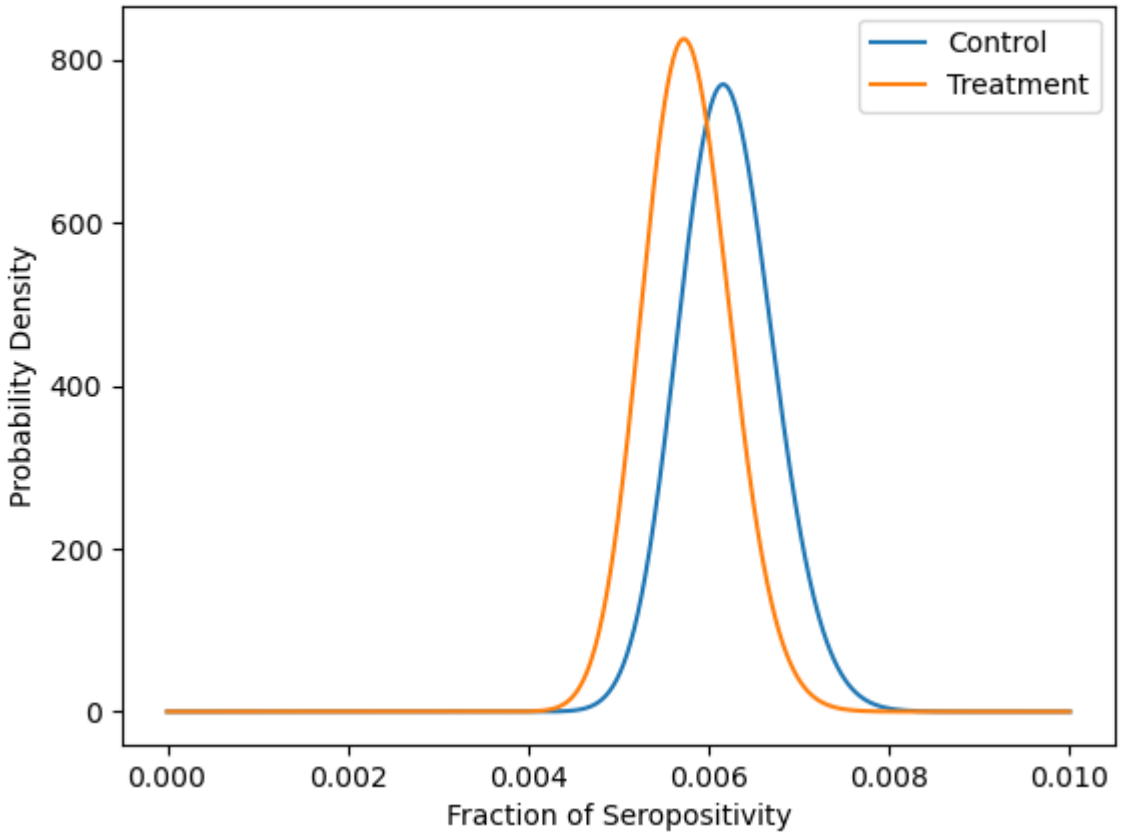
All Treatments



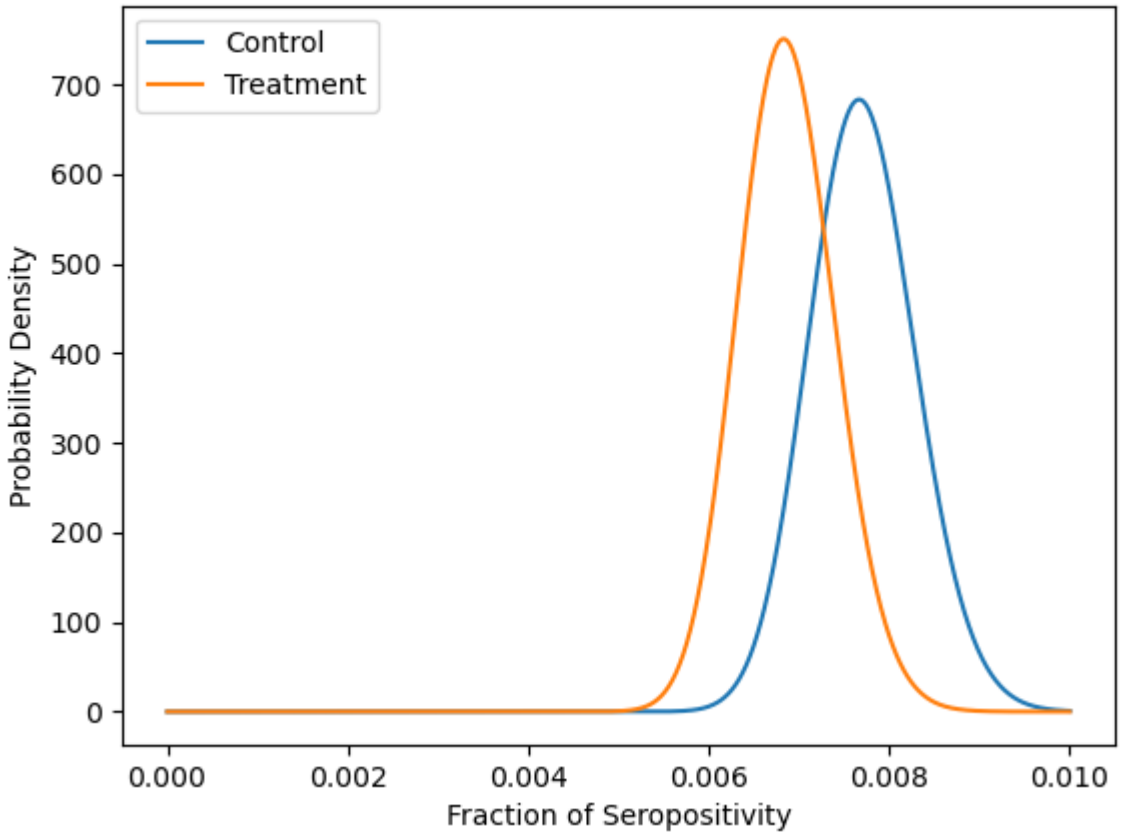
Treatment: All Surgical Masks



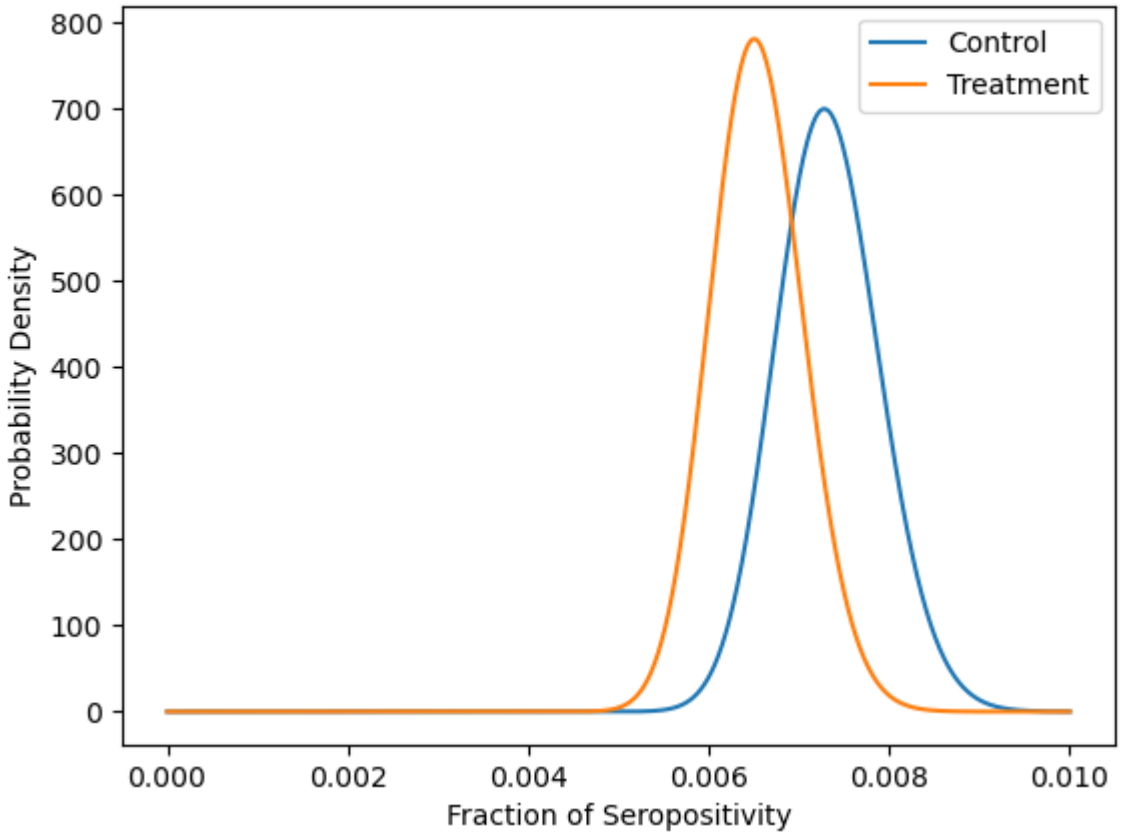
Treatment: All Cloth Masks



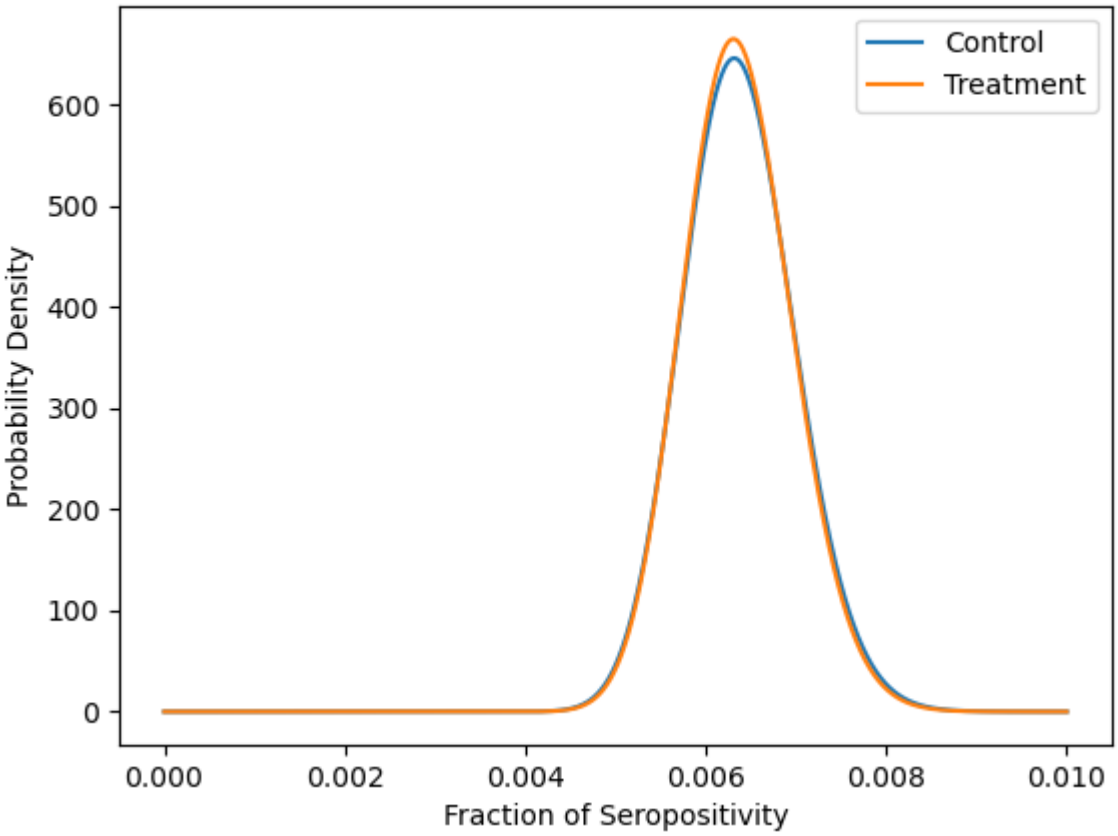
Treatment: Green Surgical Masks



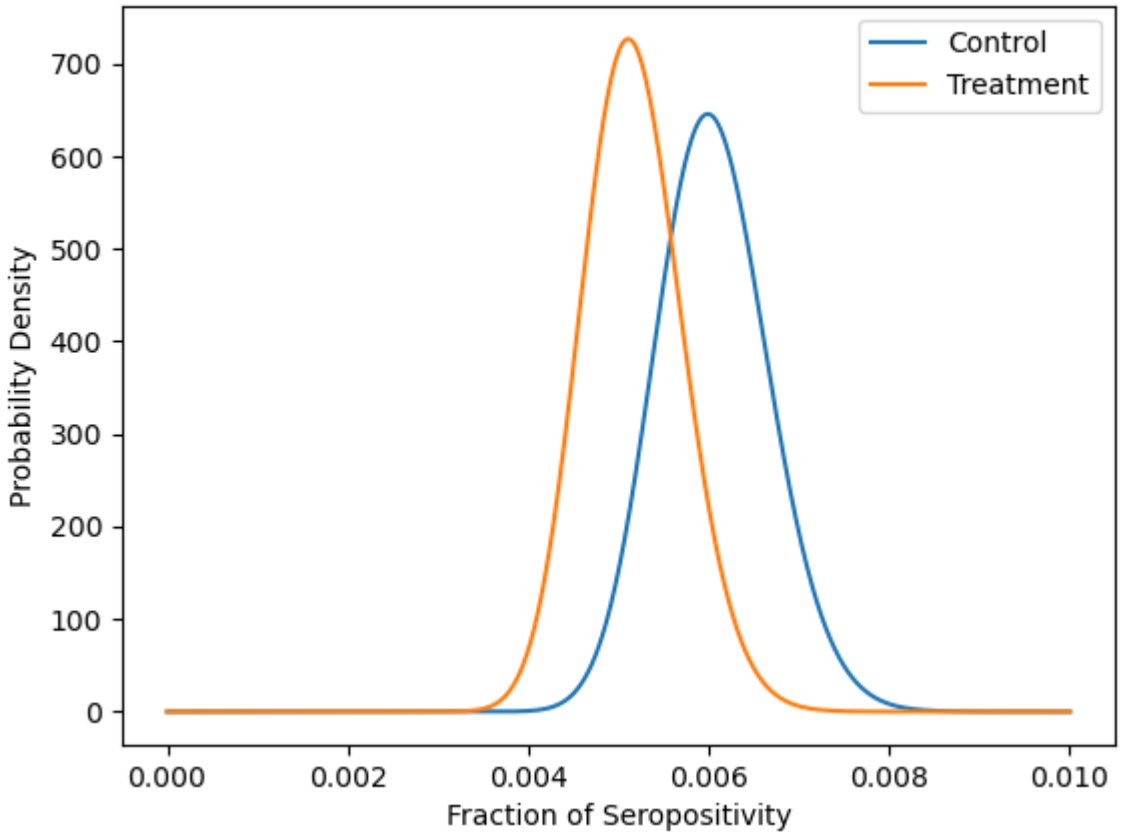
Treatment: Blue Surgical Masks



Treatment: Purple Cloth Masks



Treatment: Red Cloth Masks



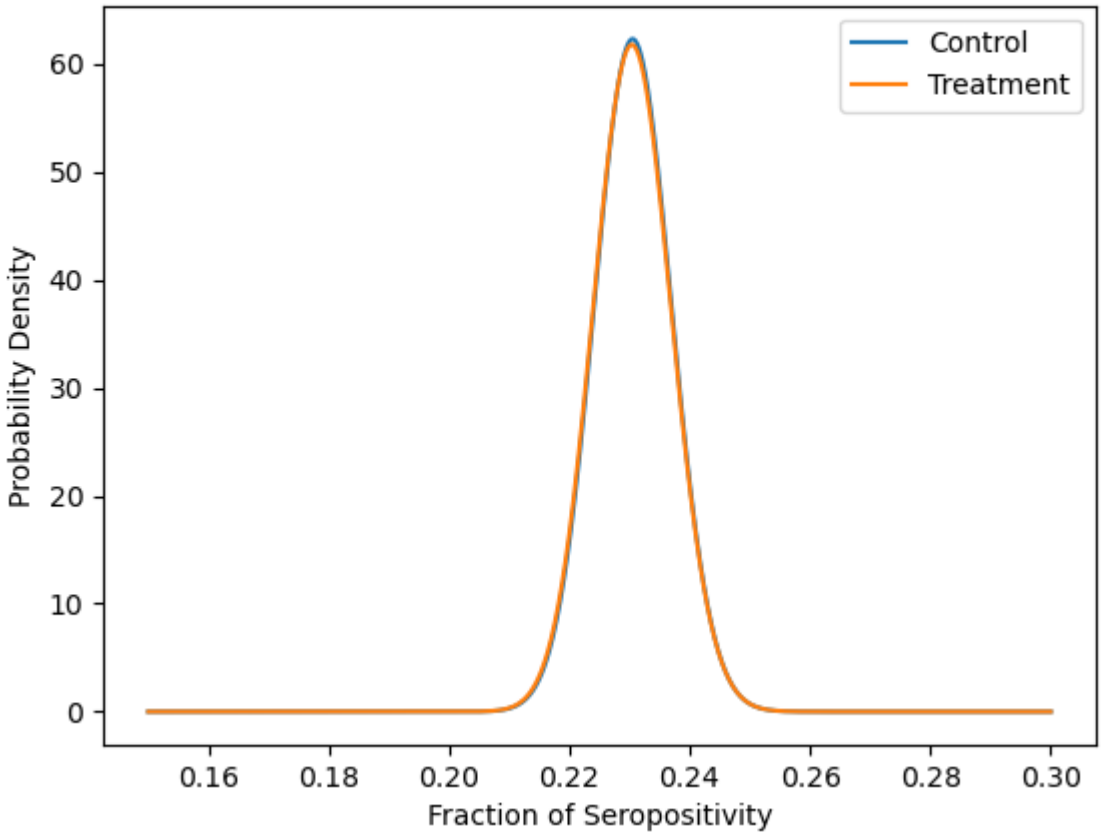
Sample Size Issues

I have a hard time with the statement: we enrolled 300,000 people into our study and then tested about 10,000 of them for COVID antibodies and then use the fraction of the 10,000 that were postive to extrapolate back to the population. Maybe it's all valid but if we only tested 10,000 people then the variance we use for our t-, z-, and p- values should be based off of 10,000 subjects, not 300,000 subjects (which necessarily makes our variance much bigger). Going all the way back to 300,000 subjects means that we're assuming zero seroprevalance everywhere else and that just doesn't feel right.

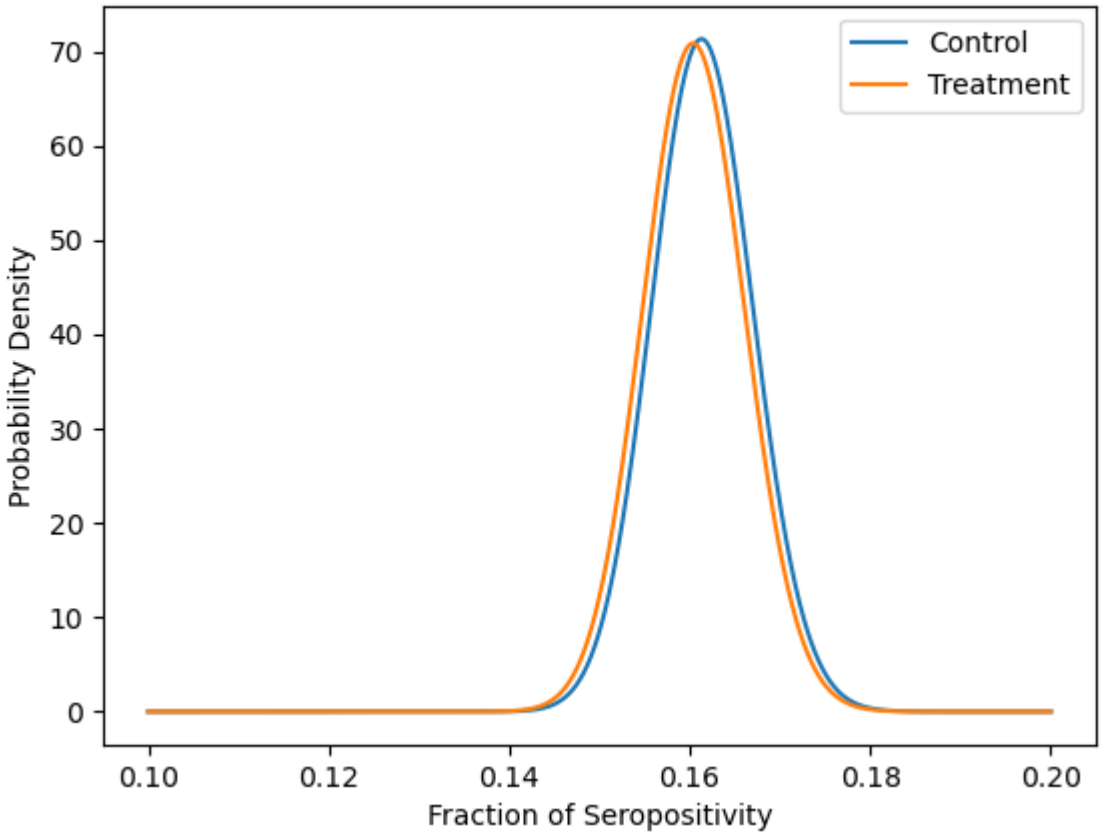
The authors discuss this in their paper by limiting their study to "the effect of mask wearing on symptomatic seroprevalance" but that feels like a first-year grad student cop out. So below here are the same plots, but with the number of subjects based off the number of people who were tested, not the entire population (also includes the DE issue from above)

Note, that the x-axis here is fraction of *symptomatic people who were tested* who were postive, not fraction of the population who was positivite. I'm also not super happy with this because they only tested symptomatic people, so this would be biased if masks strongly prevented any symptoms (but we don't see that in the original study).

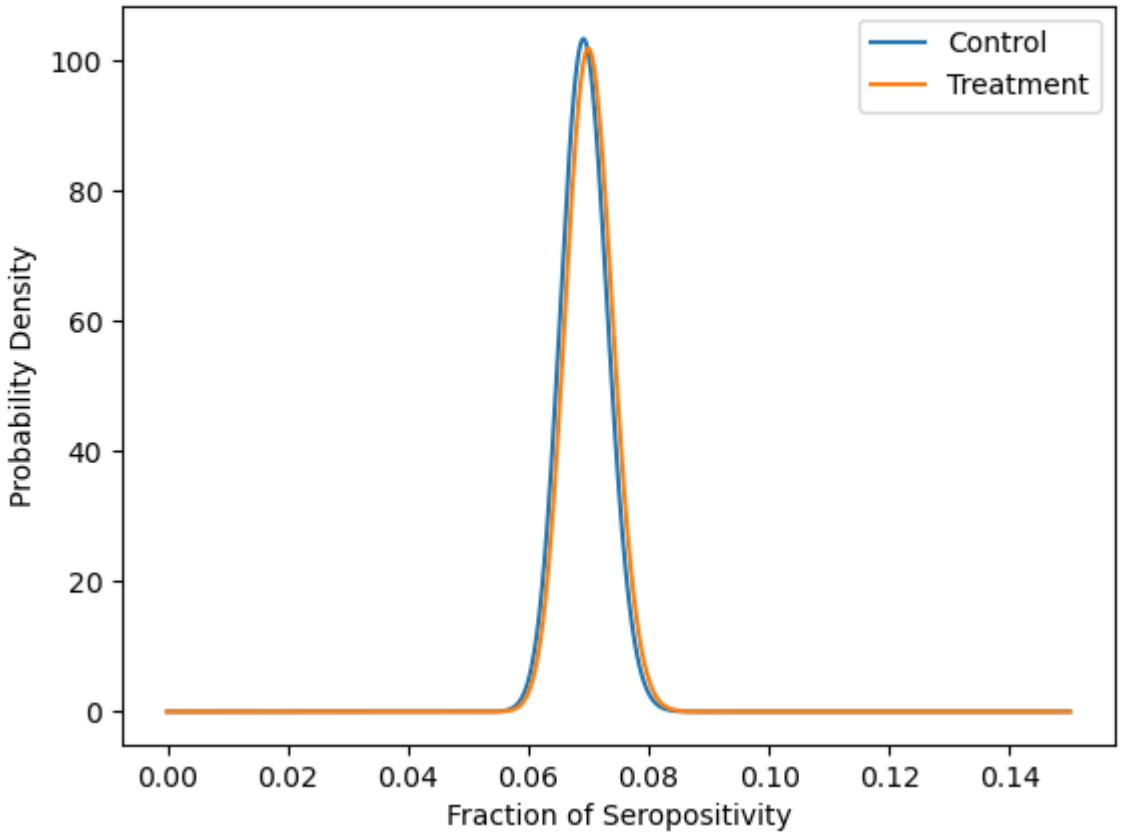
All Treatments



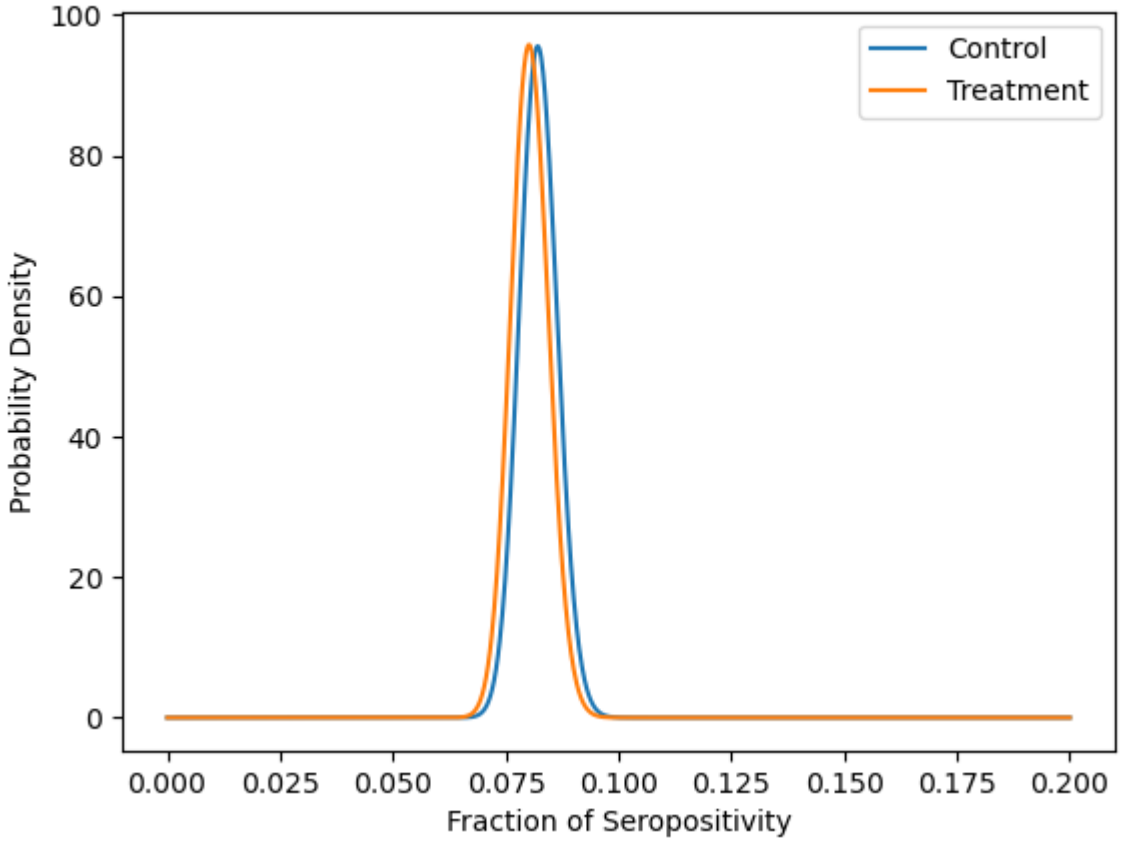
Treatment: All Surgical Masks



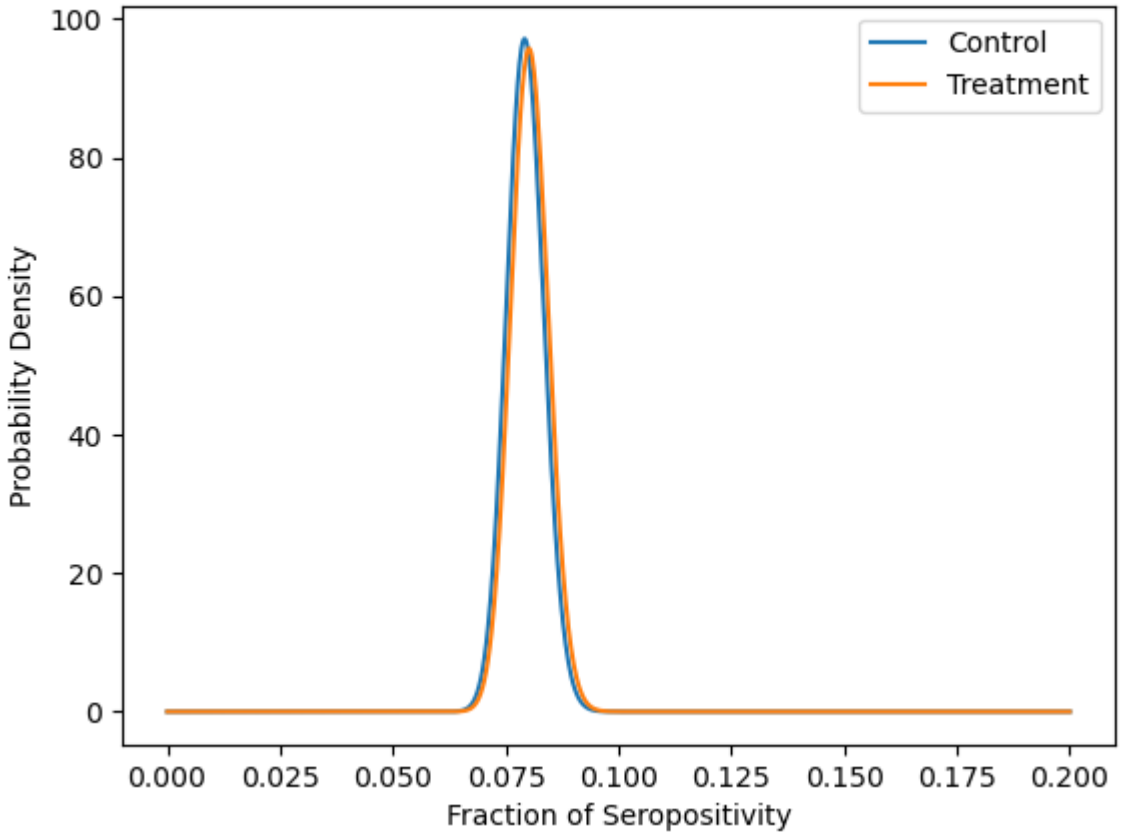
Treatment: All Cloth Masks



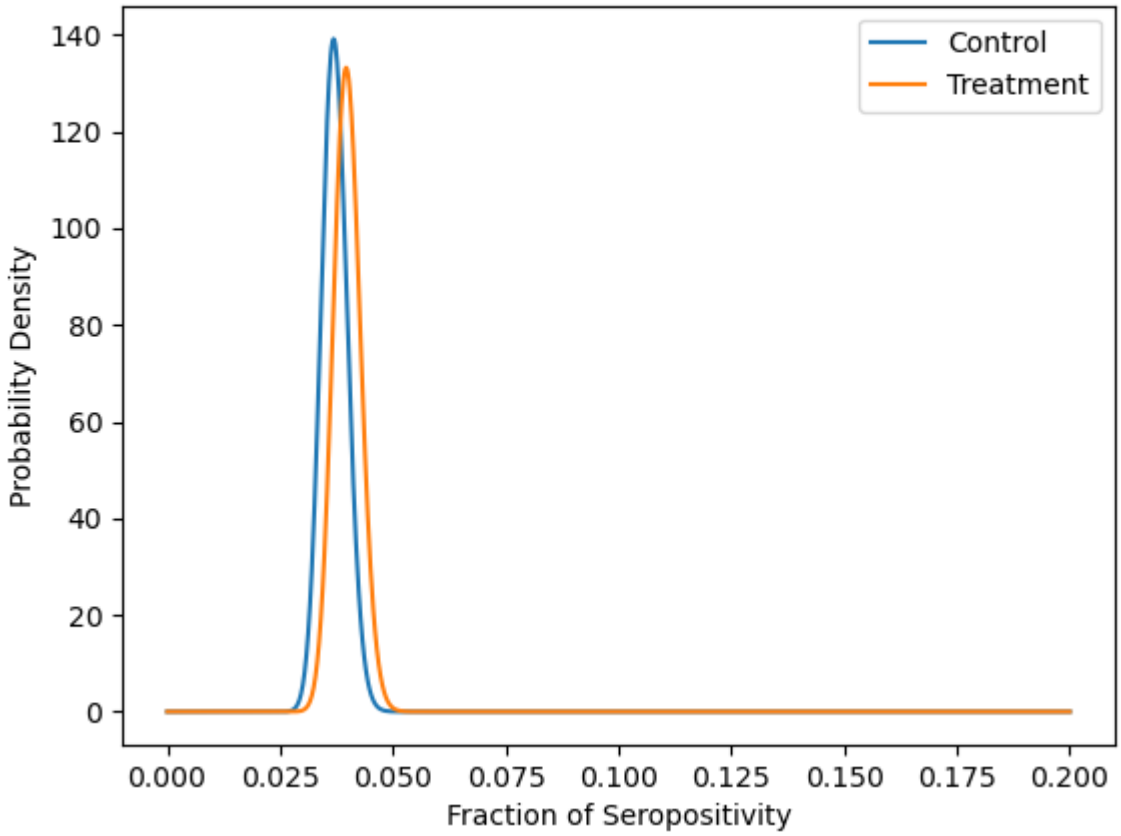
Treatment: Green Surgical Masks



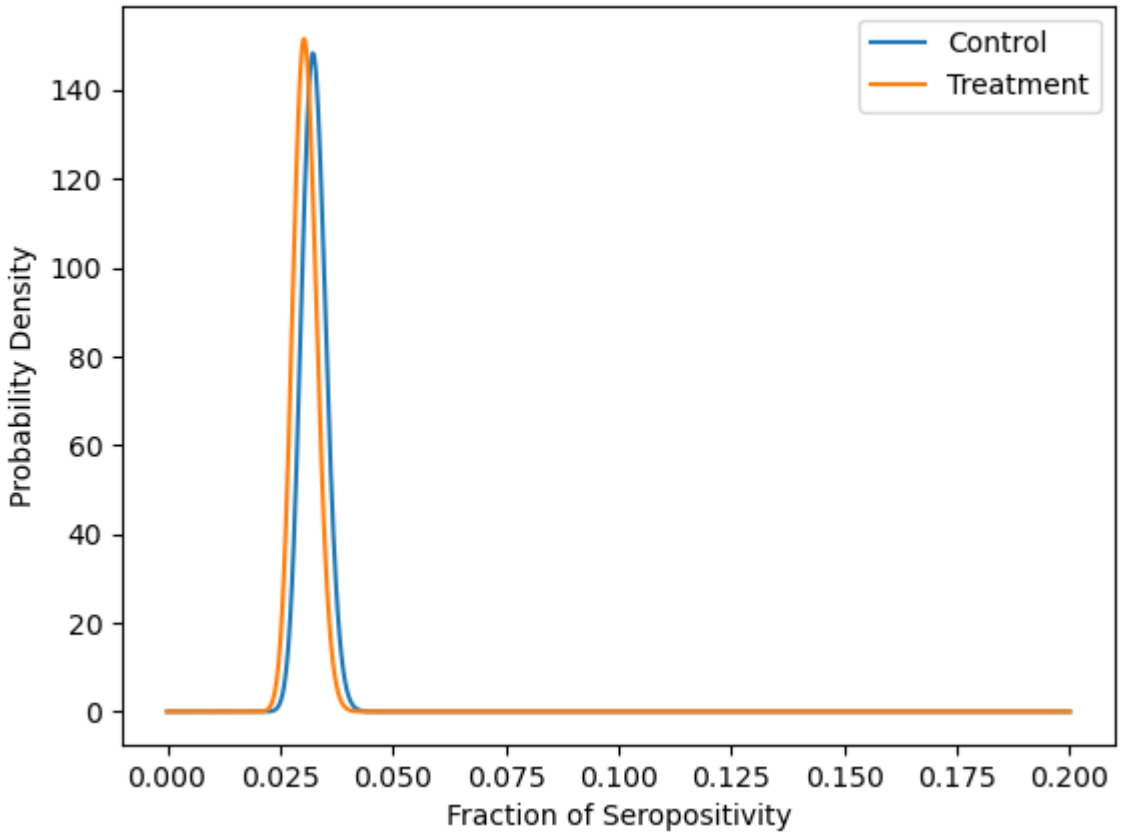
Treatment: Blue Surgical Masks



Treatment: Purple Cloth Masks



Treatment: Red Cloth Masks



One thing this does show is that *if you get symptoms*, you have the same probability as being positive regardless of whether or not your community wears masks. So, one way to interpret this is that wearing masks prevents symptoms which could also be from the flu (which is droplet-spread and a mask could work), but it doesn't prevent symptomatic COVID (except for those defective purple masks...)

False Positives

The other thing that's been nagging at me is the false positive rate on the COVID antibody test. In the paper, they mention the procedure used for the test, but I didn't see a reference to the actual test. Early in the pandemic, I was helping an MD friend of mine try to understand the population-level information gained from seroprevalence test and he sent me a data sheet on a seroprevalance test (attached to this repo as Dimension Vista SARS-CoV-2 Total Antibody IFU.pdf). It's the only seroprevalance test data I have, and I don't know if it's the same test that they're using in this paper, but it gives us a some baseline data on false positive/negatives.

From the paper, the false positive rate was on the order of 3 in 1,800 tests. So if we are testing about 10,000 people, we'd expect false positives on the order of 5-6 people, which is a significant proportion of the difference we're measuring anyway.

Conclusion

I can see why this paper was published. At a naive level it does reach statistical significance and the direction is in the way that the popular opinion goes, so the researchers and the publisher will be happy to push this along to match the narrative.

When you account for things like internal correlation, false positives, and sample size issues all of those effects go away and are lost in the noise.

Those are just the math issues. There are much deeper procedural issues in how the study was performed. Self-selection, consent, and not-fully-blinded researchers wash out the rest of differences.

This study does not point to a benefit to masks, and it clearly doesn't point towards making public policy in favor of masks. However, all of the arguments against this paper are statistical arguments and getting that message across will be very hard. The mistakes made here are only excusable in a first-year grad student, but I don't see how we can convince policy makers of that.

Going Further

If we want to get fancy, I've included two papers in this repo, j.jclinepi.2006.06.010.pdf and j.0006-341x.1999.00137.x.pdf, which talk about more nuanced behavior of the design effect on beta distributions and epidemology, but I think that's just nitpicking when there are much bigger issues (researcher bias) that need to be accounted for first.