Evidence vs. truth[1]

Andrew Gelman[2]

17 Jun 2020

In early April 2020, a team of researchers based at Stanford University conducted an opt-in survey in the surrounding county, testing for coronavirus antibodies (Bendavid et al., 2020). The result was that 50 out of 3330 people in the survey—1.5%—tested positive. Extrapolating this to the population of the county as a whole yields an estimate of 29,000 exposed, which was much larger than the number of confirmed positive cases in the county (under 1000 at the time). Coronavirus tests were hard to come by at that time, and everyone knew that the number of confirmed cases was much less than the total number of people exposed, but it was not clear how much lower.

The Stanford study was posted on the preprint server medRxiv on 11 April, and its authors were writing op-eds and explaining the implications of their findings on national television. The key result from their preprint was a range of estimates of 2.5% to 4.2% for the prevalence rate, implying "between 48,000 and 81,000 people infected in Santa Clara County by early April, 50-85-fold more than the number of confirmed cases."

Three statistical questions arose:
1. Can we trust the results, given that the survey was not a random sample?
2. How did the raw rate of 1.5% in the data become an estimate of 2.5% to 4.2% in the preprint?
3. Where did the range of uncertainty come from, and is it appropriate given sampling variability in the data?

Questions 1 and 2 go together: the increase from 1.5% to 2.5% or more comes from a statistical adjustment done by the authors to correct for the sample not matching the population (as summarized by census totals for the county) by sex, ethnicity, and zip code. Unfortunately there are a few reasons we do not feel comfortable with these adjustments: first, they don't adjust for age; second, the adjustment for zip code is potentially very noisy (there are 58 zip codes in the county, which makes adjustment difficult, given that the sample contains only 50 positive tests); third, there is concern that, even after demographic and geographic adjustment, people who were more at risk were more likely to get tested; and, fourth, there are many "researcher degrees of freedom" in the adjustment process, leading us to be skeptical of any particular published result.

Question 3 is more challenging than it might seem at first, given that any estimate of prevalence must account for testing errors. But these error rates are not precisely known;

---

[1] To appear in *Chance*. Much of first part of this article is scheduled to appear in the International Society for Bayesian Analysis Bulletin (Gelman and Carpenter, 2020b).

[2] Department of Statistics and Department of Political Science, Columbia University, New York.

they are estimated based on results from testing known positive and negative blood samples, and they can vary according to testing conditions.

During the week after the Stanford study appeared, there was increasing concern on social media regarding its data collection and statistical analysis, and it became clear that the calculations of confidence intervals in the preprint were wrong, even setting aside concerns about the demographic and geographic adjustments (Fithian, 2020, Gelman, 2020). In retrospect, it was not so easy to use classical statistical methods to account for all these uncertainties and adjustments at once. In addition, concerns were raised about data collection and conflicts of interest (Lee, 2020a,b).

On 27 April, the Stanford team issued a revised version of the report, but this too reported uncertainty intervals that were much too narrow given the information in the data. Following our own analysis of the available data, we summarized (Gelman and Carpenter, 2020a): "For now, we do not think the data support the claim that the number of infections in Santa Clara County was between 50 and 85 times the count of cases reported at the time, or the given interval for the coronavirus infection fatality rate. These numbers are consistent with the data, but the data are also consistent with a near-zero infection rate in the county. The data of Bendavid et al. (2020a,b) do not provide strong evidence about the number of people infected or the infection fatality ratio; the number of positive tests in the data is just too small, given uncertainty in the specificity of the test."

This example foregrounds the distinction between *evidence* and *truth*.

These data provide weak evidence; not much can be learned from them, beyond that the rate of infections in the sample was very low in early April. With assumptions about the sampling, some inference can be made about the general population of the county, but the resulting uncertainty interval for the prevalence goes almost all the way down to zero.

That's evidence; what do we know about the truth? We don't know much, but authors of the Stanford study have been on record as saying that they think the infection fatality rate of coronavirus was very low and that number of untested infected people was 50 or more times the number of people who had tested positive by early April. These beliefs are possible—they might be true—and they are consistent with the available data.

As statisticians, we focus on what can be learned from the data at hand plus whatever assumptions (about randomization, representativeness, bias, variance, and so forth) we are willing to assume. We call it a statistical error if you make a quantitative claim that is not supported by your data and assumptions.

But scientists and policymakers are often less interested in evidence and more interested in truth. If the scientific or policy claims are true (and, in this case, they might be), it is considered forgivable to overstate the evidence.

This example also raises a set of ethical questions, which we will discuss in the order that they arose:

1. Is it an ethics violation to make serious statistical errors in high-profile, massively publicized research? No, this is not alone an ethics violation. Statistics is hard, in the usual practice of research, we make statistical errors all the time. The obligation is not to avoid mistakes but (a) to acknowledge and correct them as soon as possible, and (b) to facilitate conditions so that others can easily find out where and how we went wrong.
2. Is it an ethics violation to not engage with relevant experts before releasing and publicizing a report? This is a tougher call. Having no statistics experts involved in what is essentially a statistics project (estimating population prevalence from a sample) represents a sort of meta-ignorance, not just about statistics but about the very fact that statistics is a field with specialized knowledge.
3. Is it an ethics violation to release and publicize a report without making data and code available? Again, this is a borderline call. It can take effort to release data that are sufficiently de-identified to satisfy institutional and legal requirements, and it is unfortunately not yet standard in many areas of science to share computer code. But not releasing data and code has negative consequences, as it makes it more difficult for outsiders to figure out what went wrong. Beyond this, the very act of releasing data and code might well motivate researchers to perform their analyses more carefully.
4. Is it an ethics violation to not acknowledge errors in one's work that have been noted by outsiders? Yes.

What about evidence and truth? Given that lives and livelihoods are at stake, are statisticians being picky by harping on evidential gaps? We don't think so. It should be possible for researchers and even the news media to distinguish between truth and evidence.

For example, the Stanford team could say: "Our report had errors. As was pointed out by several statisticians, the data we reported are consistent with near-zero prevalence of coronavirus in Santa Clara County in early April. However, under certain assumptions the data are also consistent with higher rates in the 4% range, and for reasons separate from this study we believe these higher rates are correct."

It should be possible to say that your data are consistent with your prior beliefs while acknowledging that these data also admit other interpretations, and it should be possible to move beyond the norm in which researchers do not reassess their conclusions when they learn of errors in their work.

References

Bendavid, E., Mulaney, B., Sood, N., Shah, S., Ling, E., Bromley-Dulfano, R., Lai, C., Weissberg, Z., Saavedra-Walker, R., Tedrow, J., Tversky, D., Bogan, A., Kupiec, T., Eichner, D., Gupta, R., Ioannidis, J., and Bhattacharya, J. (2020a). COVID-19 antibody

seroprevalence in Santa Clara County, California, version 1. https://www.medrxiv.org/content/10.1101/2020.04.14.20062463v1

Bendavid, E., Mulaney, B., Sood, N., Shah, S., Ling, E., Bromley-Dulfano, R., Lai, C., Weissberg, Z., Saavedra-Walker, R., Tedrow, J., Tversky, D., Bogan, A., Kupiec, T., Eichner, D., Gupta, R., Ioannidis, J., and Bhattacharya, J. (2020b). COVID-19 antibody seroprevalence in Santa Clara County, California, version 2. https://www.medrxiv.org/content/10.1101/2020.04.14.20062463v2

Fithian, W. (2020). Statistical comment on the revision of Bendavid et al. https://www.stat. berkeley.edu/~wfithian/overdispersionSimple.html

Gelman, A. (2020). Concerns with that Stanford study of coronavirus prevalence. Statistical Modeling, Causal Inference, and Social Science blog, 19 Apr. https://statmodeling.stat.columbia.edu/2020/04/19/fatal-flaws-in-stanford-study-of-coronavirus-prevalence/

Gelman, A., and Carpenter, B. (2020a). Bayesian analysis of tests with unknown specificity and sensitivity. https://www.medrxiv.org/content/10.1101/2020.05.22.20108944v2

Gelman, A., and Carpenter, B. (2020b). Using Bayesian analysis to account for uncertainty and adjust for bias in coronavirus sampling. International Society for Bayesian Analysis Bulletin.

Hemenway, D. (1997). The myth of millions of annual self-defense gun uses: A case study of survey overestimates of rare events. Chance 10 (3), 6–10.

Lee, S. M. (2020a). A Stanford professor's wife recruited people for his coronavirus study by claiming it would reveal if they could "return to work without fear." BuzzFeed News, 24 Apr. https://www.buzzfeednews.com/article/stephaniemlee/stanford-coronavirus-study-bhattacharya-email

Lee, S. M. (2020b). JetBlue's founder helped fund a Stanford study that said the coronavirus wasn't that deadly. BuzzFeed News, 15 May. https://www.buzzfeednews.com/article/stephaniemlee/stanford-coronavirus-neeleman-ioannidis-whistleblower