The Hypothesis of Testing: Paradoxes arising out of reported coronavirus case-counts

April 16, 2020

Abstract

Many statisticians, epidemiologists, economists and data scientists have registered their serious reservations regarding the reported coronavirus case-counts. Comparing countries and states using those case-counts seem inappropriate when every nation/state have adopted different testing strategies and protocols. Estimating prevalence of COVID-19 based on these data is a hopeless exercise and several groups have recently argued for estimating the number of truly infected cases by using mortality rates. In this note, we aim to (a) we posit a conceptual mathematical framework to characterize both sampling bias and misclassification/imperfection of the test simultaneously on the estimation of the prevalence rate, (b) review current testing strategies in some of the countries where we have testing data, and (c) provide guidelines for testing strategy/disease surveillance that may help track the pulse of the epidemic, to identify disease free areas and identify disease outbreaks.

1 Introduction

The World Health Organization has declared the coronavirus disease 2019 (COVID-19) a public health emergency. As of March 28th, 2020, a total of 622,450 cases have been confirmed worldwide. As of that afternoon, the New York Times reports at least 113,031 people across the United States have tested positive for the virus, and at least 1,895 patients with the virus have died. Aggressive policies have been put in place across the US with at least 50% of the US population officially urged to stay home via state-wide executive actions.

Despite these necessary steps, the data landscape for understanding COVID-19 remains limited. Public databases maintained by Johns Hopkins University (`https://bit.ly/2UqFSuA`) and the New York Times (`https://bit.ly/2vUHfrK`) provide incoming county-level information of confirmed cases and deaths. Statisticians, epidemiologists, economists, and data scientists the world over have been using this granular data to

The WHO reports, however, only 103,945 total tests performed in the US as of March 19th. Due to limited testing capacity, local and state health departments have focused on testing only those from high-risk populations, resulting in low data quality due to bias in sampling from the overall population. The public health community requires auxiliary

sources of information to improve national and local health policy decisions. Certain survey efforts are underway, but may take time to yield fruit.

The goal of this paper is to express reservations at the use of case-counts as a proxy for prevalence and disease trajectory as well as its use as direct input into estimation of standard epidemiological models for inference and forecasting. The two critical features are selection bias due to varied testing strategies and measurement error due to imperfect tests (i.e., false positives and negatives). We demonstrate via theoretical analysis More testing equals more data but does not imply more information. The key issue is selection bias and measurement-error.

Given population sizes are quite heterogeneous, the current data are of limited use to understand prevalence and even trajectory of the disease.

A critical question is to understand the mismatch between there are alternative data streams that may be leveraged to understand the handling the COVID-19 pandemic in the US.

1.1 Related work

There has been an abundance..

However, most complain about testing *capacity* (cite Nate Silver), there is clearly issues of data quality.

1.2 Outline

- Selection bias + MEM = problems! Scales with population size so cross country comparisons are meaningless.
- What about cross-time same spot? That's equally difficult!
- Present a compartmental model that accounts for these time-varying factors. Discuss identifiability (statistical issue) and how to proceed
- Review testing and building selection models (sensitivity analysis)
- Robustness (what directions are the models most resilient) and the notion of forecasts in bubbles
- Testing, disease surveillance and second-wave testing goals.

2

The naive data analyst may think

As we increase testing capacity, we will learn more about disease prevalence.

Case-count is considered a proxy for disease prevalence and contagion spread.

Let N denote the population size. At a fixed moment in time, let $Y_j \in \{0, 1\}$ for $j = 1, \ldots, N$ denote COVID status. As in survey sampling methodology, we treat these as

fixed, unknown population quantities. For now, we ignore the dynamic nature of the viral outbreak; so either individual j is COVID-19 positive and $Y_j = 1$ or is COVID-19 negative and $Y_j = 0$. A testing indicator, $I_j \in \{0, 1\}$ is an indicator that the individual was selected for testing ($I = 1$) or not ($I = 0$).

A primary question is the prevalence o We are interested in the population average $\bar{Y} = \frac{1}{N}\sum_{j=1}^{N} Y_j$. Suppose we observe a sample of n tests. A natural candidate for prevalence is the proportion of positive tests $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

Prevalence is important in the long-run because the

Based on Meng (2019), we can express the error between $\bar{y}$ and the true proportion as

$$\bar{y}_n - \bar{Y} = \rho_{I,Y} \times \sqrt{\frac{1-f}{f}} \times \sigma_Y$$

Under random sampling $E[\rho_{I,Y}] = 0$ and there is no bias. Under selective testing,

Case-count Figure **??** presents testing per capita across reporting countries. We see th at

Two countries with the same testing strategy (i.e., $\rho_{I,Y}$ equal) can yield wildly different estimates due to pop size.

2.1 Imperfect tests and

Meng's analysis is critical when selection bias is present. For the current crisis, the analysis should be extended in two directions. First, tests are imperfect. Measurement-error.

Let P_j be an indicator that equals 1 when we observe reverse of the true outcome and equals 0 when we observe the true outcome. We suppose this is a stochastic variable that satisfies $\text{pr}(P_j = 1 \mid Y_j = 1) =: FN$ is the false-negative rate and $\text{pr}(P_j = 1 \mid Y_j = 0) =: FP$ is the false-positive rate. If individual j is selected (i.e., $I_j = 1$) then the $Y_j^\star = Y_j(1 - P_j) + (1 - Y_j)P_j$.

$$\bar{y}_n - \bar{Y} = \sqrt{\frac{1-f}{f}}\left[\rho_{I,Y} \times \mu_Y(1-\mu_Y) + \rho_{I,PZ} \times \sigma_{PZ} + \sqrt{\frac{f}{1-f}}\left(FP - (FP+FN)\bar{Y}\right)\right].$$

where $\sigma_{PZ} = \sqrt{2FP}\sqrt{\bar{Y}\left[1 - \frac{FP+FN}{FP}\bar{Y}\right]}$.

This is quite complex but he main point is that the sign of the new terms can be either positive or negative. Moreover, there is an interaction between... This means that raw application is, of course, biased but the expected bias directionality is not guaranteed. There's a clear interaction between selection bias, measurement-error, prevalence, and data quantity!

To see this

The attentive data analyst will recognize the estimator $\bar{y}_n$ will be biased even for simple random samples and suggest the alternative $\tilde{y}_n = \bar{y}_n + (1 - \bar{y}_n)FP + FN\bar{y}_n$. We again wish to express the error $\tilde{y}_n - \bar{Y}$ in statistical terms. Let $Z_j = 1 - 2Y_j$; this converts the binary 0/1 variable into a sign indicator, equal to 1 when the outcome is. In the appendix, we show

that the error now can be expressed as

$$\tilde{y}_n - \bar{Y} = (1 + FP + FN)\rho_{I,Y} \times \sqrt{\frac{1-f}{f}} \times \sigma_Y + \rho_{I,PZ} \times \sqrt{\frac{1-f}{f}} \times \sigma_{PZ}.$$

The first term is the same as before but increased by $(1 + FP + FN)$ to account for the additional uncertainty due to measurement-error. The second term is the interaction between selection bias and measurement error. Interestingly we can show that the sign is reversed, i.e., $sgn(\rho_{I,PZ}) = 1 - sgn(\rho_{I,Y})$, leading to an absolute *reduction* in the error.

For this adjusted estimate, we can see that the directionality is maintained. Therefore, the benefit is if the expected correlation is positive, we can feel like the estimate is likely an over-estimate and vice versa. You pay a penalty in terms of variance and therefore the MSE may not be much smaller (CHECK!)

2.2 Testing strategies and population size: a tango

Data analysts have argued whether

2.3 Regrettable rates

The data analyst, now wary of estimating prevalence and total counts, pauses and thinks. They retur n shortly thereafter, a bit agitated but still *Ok, perhaps total counts is a lost cause. Certainly, however, we can estimate the rate of growth. All I want to know is when we hit the point at which the curve flattens and number of deaths decrease. That can't be too hard, surely!*

The parameter R_0 in the ubiquitous SIR (susceptible-)

Ratio estimator $\bar{y}_2/\bar{y}_1$.

Taylor series expansion:

$$\begin{aligned}
\frac{\bar{y}_2}{\bar{y}_1} - \frac{\bar{Y}_2}{\bar{Y}_1} &= \frac{\bar{y}_2}{\bar{Y}_1\left(1 + \rho_{I,Y_1}\sqrt{\frac{1-f}{f}}CV(Y_1)\right)} - \frac{\bar{Y}_2}{\bar{Y}_1} \\
&\approx \frac{\bar{y}_2}{\bar{Y}_1}\left(1 - \rho_{I,Y_1}\sqrt{\frac{1-f}{f}}CV(Y_1)\right) - \frac{\bar{Y}_2}{\bar{Y}_1} \\
&\approx \frac{1}{\bar{Y}_1}\left[\rho_{I,Y_2}\sqrt{\frac{1-f}{f}}\sigma_{Y_2} - \rho_{I,Y_1}\sqrt{\frac{1-f}{f}}\sigma_{Y_1}\left(\frac{E_{J_2}[I_{J_2}Y_{2,J_2}]}{E_{J_2}[I_{J_2}]} + \frac{E_{J_2}[I_{J_2}]E_{J_2}[Y_{2,J_2}]}{E_{J_2}[I_{J_2}]} - \frac{E_{J_2}[I_{J_2}]E_{J_2}[Y_{2,J_2}]}{E_{J_2}[I_{J_2}]}\right)\right] \\
&\approx \frac{1}{\bar{Y}_1}\left[\rho_{I,Y_2}\sqrt{\frac{1-f}{f}}\sigma_{Y_2} - \rho_{I,Y_1}\sqrt{\frac{1-f}{f}}\sigma_{Y_1}\left(\rho_{I,Y_2}\sqrt{\frac{1-f}{f}}\sigma_{Y_2} + \mu_{Y_2}\right)\right]
\end{aligned}$$

We talk about wanting to know if we are at the peak. Let's define that time as when $\bar{Y}_2/\bar{Y}_1 = 0$. Then $\mu_{Y_1} = \mu_{Y_2} = \mu$ and $\sigma_{Y_1} = \sigma_{Y_2} = \sigma$. If selection at each time is independent

and the selection process is stationary, we have bias of

$$\rho\sqrt{\frac{1-f}{f}}\sigma\sqrt{\frac{1-\mu}{\mu}}\left[\sqrt{\frac{1-\mu}{\mu}}-\rho\sqrt{\frac{1-f}{f}}\right]$$

If we collect $f = 0.01$ and the peak occurs at $\mu = 0.03$ with selection bias of $\rho = 0.05$. $1/\rho^2 \times (\mu/(1-\mu))^4$

Then we can be even further off than the prevalence by a factor of the odds ratio!

Suppose we collected $f = 1/2$ and peak occurs at $\mu = 0.10$ and $\rho = 0.005$, then $9 - \rho 3 = 9 - 0.005 \times 3 = 8.985$ and

2.4 Testing increase changes data quality

Many (cites) hve

3 Compartmental model

(A) Heterogeneity in R_0 (B) Selection bias (C)

Above we took a general design perspective. Here we take a model-assisted perspective.

Susceptible individuals can become exposed to the disease.

Selection bias occurs at 2 phases. Phase 1 is self-selection to seek

Phase 2 is the decision to receive testing. This is performed by the

The result is we have multiple types of people: (A) Those disease-free who seek treatment due to high risk.

We must then layer on measurement-error. Tests can

The SIR model with parameters.

Increasing testing leads to lower correlation in

3.1 Heterogeneiety SIR with measurement-error and

Here we explain a general state-space process model as compared to the measurement-process. While related, the distinction drives our key results.

Observer controls testing, FP and FN. Patients

Their behavior may depend on forecasts and therefore the notion of a counterfactual is not well defined. For example, suppose an highly prominent model....

The transition from susceptible to infected is a complex function. It is based on personal choices; the second choice B

Non-identifiable and rely on assumptions. As per Cox's statement.

This makes models *fragile* and rely on strong data-generating assumptions. For a disease

4 Decision-making: data versus information.

Above we point to flaws in using observed data to reason about the

4.1 Interventions on network effects

While we think of the intervention as reducing risk COVID-19; in reality the interventio nis

Classical contact pretends the graph is fixed. Here we assume the intervention is to reduce degree and remove edges. Contact tracing is indeed the way we can

How does R_0 relate to connectivity? This accounts for the heterogeneity. Epi knows this, but I think this point is salient.

5 That which does not break us, makes us stronger (but potentially not smarter)

A slightly altered version of Nietzsche's famous quote has become a mantra for post-pandemic thinking: *That which which does not break us, makes us stronger*. While potentially true via viral resistance, it is not clear that governments have yet to learn lessons on pandemic response.

During the current wave, understanding prevalence is key. It helps us determine long-term risk for a community and make targeted interventions. This will allow When prevalence is below a certain threshold, we can return to daily life.

Once we "return to normal"[1], it is clear that testing strategies should focus on early detection, heading off future outbreaks. *Anyone who wishes to go back to work*. While laudable, without complete compliance, we may be riddles with selection bias and measurement-error. This ignores even the practical and ethical quandaries of how to .

What do we do when we are faced with? We design experiments with our objectives in mind.

(A) Prevalence: Targeted shutdowns (B) Risk Detection: Contact tracing and altering

Use prior results to show that ratio estimator under SRS is a good predictor of potential outbreak.

"Disease free"

6 Conclusion

Interventions in this area are designed experiments. Unlike Fischer's null, here the objectives are settings, statisticians

A Setup and calculation

I focused this note on the very simply setting of understanding the error when both selection bias and measurement-error are present. I worked within Meng's derivation as it provides a guide for how to think this error in our current context. The conclusion so far is that the mathematics will be quite cumbersome, but my hope is that we can potentially draw out some interesting conclusions. In particular, my hope is to provide theoretical confirmation

[1] or at least the new normal

that the interaction of selection bias and measurement-error does not lead to "attenuation" of the effect as in a simple measurement-error context. This is a very very early draft but I wanted to share current "state-of-affairs" to see whether you thought further investigation along this line of thinking was merited.

- Finite-population of size N
- Binary outcome $Y_j \in \{0, 1\}$ for $j = 1, \ldots, N$. As in survey methodology, we treat these as fixed, unknown population quantities. For now, we ignore the dynamic nature of the viral outbreak. So you are either COVID positive ($Y = 1$) or negative ($Y = 0$)
- Selection indicator $I_j \in \{0, 1\}$. This is an indicator that the individual was selected for testing ($I = 1$) or not ($I = 0$).
- Measurement-error: Let P_j be an indicator that equals 1 when we observe reverse of the true outcome and equals 0 when we observe the true outcome. We suppose this is a stochastic variable that satisfies $\text{pr}(P_j = 1 \mid Y_j = 1) =: FN$ is the probability of a false-negative and $\text{pr}(P_j = 1 \mid Y_j = 0) =: FP$ is the probability of a false-positive.
- If we select individual j (i.e., $I_j = 1$) then we observe $Y_j^\star = Y_j(1 - P_j) + (1 - Y_j)P_j$
- We are interested in the population average $\bar{Y} = \frac{1}{N}\sum_{j=1}^{N} Y_j$
- Consider the mean estimator

$$\bar{y}_n = \frac{1}{n}\sum_{j=1}^{n} Y_j^\star = \frac{\sum_{i=1}^{N} I_j Y_j^\star}{\sum_{j=1}^{N} I_j} = \frac{\sum_{i=1}^{N} I_j \left[Y_j(1 - P_j) + (1 - Y_j)P_j\right]}{\sum_{j=1}^{N} I_j}$$

- We wish to express the error $\bar{y}_n - \bar{Y}$ in statistical terms
- For any set of numbers $\{A_1, \ldots, A_N\}$ we can view it as the support of a random variable A_J induced by the random index J defined on $\{1, \ldots, N\}$. When J is uniformly distributed $E_J(A_J) = \sum_{j=1}^{N} A_j/N \equiv \bar{A}_N$.
- Then

$$\begin{aligned}\bar{y}_n - \bar{Y}_N &= \frac{E_J\left[I_J\left[Y_J(1 - P_J) + (1 - Y_J)P_J\right]\right]}{E_J[I_J]} - E_J[Y_J] \\ &= \frac{E_J\left[I_J P_J(1 - 2Y_J)\right]}{E_J[I_J]} + \left(\frac{E_J[I_J Y_J]}{E_J[I_J]} - \frac{E_J[Y_J]E_J[I_J]}{E_J[I_J]}\right)\end{aligned}$$

- The term in parentheses can be re-written as

$$\begin{aligned}\frac{E_J[I_J Y_J] - E_J[Y_J]E_J[I_J]}{E_J[I_J]} &= \frac{E_J[I_J Y_J] - E_J[Y_J]E_J[I_J]}{\sqrt{V_J(I_J)V_J(Y_J)}} \frac{\sqrt{V_J(I_J)}}{E_J[I_J]} \times \sqrt{V_J(Y_J)} \\ &= \rho_{I,Y} \times \sqrt{\frac{(1 - f)}{f}} \times \sigma_Y\end{aligned}$$

  which agrees with the decomposition by Meng (2019).

- For the other term, first define $Z_j = 1 - 2Y_j$. Then $Z_j = 1$ if $Y_j = 0$ and $Z_j = -1$ if $Y_j = 1$. Then the other term can be re-written as

$$\frac{E_J\left[I_J P_J (1 - 2Y_J)\right]}{E_J[I_J]} = \left(\frac{E_J\left[I_J P_J Z_J\right]}{E_J[I_J]} - \frac{E_J\left[P_J Z_J\right] E_J[I_J]}{E_J[I_J]}\right) + \frac{E_J\left[P_J Z_J\right] E_J[I_J]}{E_J[I_J]}$$

- The term in parentheses can be re-expressed using Meng's techniques as:

$$\rho_{I,PZ} \times \sqrt{\frac{1-f}{f}} \times \sigma_{PZ}$$

where now the "data defect" and "problem difficulty" are with respect to PZ rather than Y.

- The final term is equal to

$$\begin{aligned} E_J[P_J Z_J] &= E_J[E_J[P_J Z_J \mid Y_J]] \\ &= \text{pr}(P = 1 \mid Y = 0)(1 - \bar{Y}) - \text{pr}(P = 1 \mid Y = 1)\bar{Y} \\ &= FP - (FP + FN) \cdot \bar{Y} \end{aligned}$$

- Combining these yields:

$$\bar{y}_n - \bar{Y} = \sqrt{\frac{1-f}{f}} \left[\rho_{I,Y}\sigma_Y + \rho_{I,PZ}\sigma_{PZ} + \sqrt{\frac{f}{1-f}}\left(FP - (FP + FN)\mu_Y\right)\right]$$

- We know for binary outcome Y that $\sigma_Y = \sqrt{\mu_Y(1 - \mu_Y)}$.

- We can try and understand σ_{PZ}:

$$\begin{aligned} V_J(P_J Z_J) &= E_J[(P_J Z_J)^2] - E[P]E[Z] \\ &= E[P] - E[P](1 - 2\mu_Y) = 2\mu_Y E_J[P_J] \\ &= 2\mu_Y\left(FP(1 - \mu_Y) + FN\mu_Y\right) \\ \Rightarrow \sigma_{PZ} &= \sqrt{2\mu_Y\left(FP(1 - \mu_Y) + FN\mu_Y\right)} \end{aligned}$$

- Then the formula for the error is given by:

$$\sqrt{\frac{1-f}{f}} \left[\rho_{I,Y}\sqrt{\mu_Y(1 - \mu_Y)} + \rho_{I,PZ}\sqrt{2\mu_Y\left(FP(1 - \mu_Y) + FN\mu_Y\right)} + \sqrt{\frac{f}{1-f}}\left(FP - (FP + FN)\mu_Y\right)\right.$$

- Let's consider a simple case first and assume no false positive results, i.e., set $FP = 0$. Then $\sigma_{PZ} = \mu_Y\sqrt{2FN}$. Then the math greatly simplifies:

$$\bar{y}_n - \bar{Y} = \mu_Y \sqrt{\frac{1-f}{f}} \left[\rho_{I,Y}(1 - \mu_Y) + \sqrt{FN}\left(\sqrt{2}\rho_{I,PZ} - \sqrt{\frac{f}{1-f}}\sqrt{FN}\right)\right]$$

- We see an interesting trade-off in the second term. The term in parentheses is positive iff

$$\rho_{I,PZ}^2 \geq \frac{f}{1-f}\frac{FN}{2}$$

- Thus the sign of the additional error depends on selection bias ($\rho_{I,PZ}$), false negative rates (FN), and sample fraction f. Moreover, the scale of the additional term depends on the false negative rate and the infection rate $\mu_Y = \bar{Y}$.

- Let's consider the other simple case and assume no false negative results, i.e., set $FN = 0$. Then $\sigma_{PZ} = \sqrt{2FP\mu_Y(1-\mu_Y)}$. Then the math greatly simplifies:

$$\bar{y}_n - \bar{Y} = \sqrt{1-\mu_Y}\sqrt{\frac{1-f}{f}}\left[\rho_{I,Y}\mu_Y\sqrt{1-\mu_Y} + \sqrt{FP}\left(\sqrt{2\mu_Y}\rho_{I,PZ} - \sqrt{\frac{f}{1-f}}\sqrt{FP}\sqrt{1-\mu_Y}\right)\right]$$

- We see an interesting trade-off in the second term. The term in parentheses is positive iff

$$\rho_{I,PZ}^2 \geq \frac{f}{1-f}\frac{FP}{2} \times \underbrace{\frac{1-\mu_Y}{\mu_Y}}_{\text{Inverse Odds Ratio}}$$

- Thus the sign of the additional error depends on selection bias ($\rho_{I,PZ}$), false positive rates (FP), sample fraction f, AND the odds ratio. Moreover, the scale of the additional term depends on the false positive rate and one minus the infection rate $1-\mu_Y = 1-\bar{Y}$.

- Let's compute the MSE

- If we adjust $\bar{y}_n$ by taking $\bar{y}_n + (FP + FN)\bar{y}_n - FP$; then

$$\sqrt{\frac{1-f}{f}}\left[\rho_{I,Y}\sigma_Y(1+FP+FN) + \rho_{I,PZ}\sqrt{2\mu_y\left(FP(1-\mu_y) + FN\mu_Y\right)}\right]$$

- In FN = 0; we have

$$\sqrt{\frac{1-f}{f}}\sqrt{\mu_Y(1-\mu_Y)}\left[\rho_{I,Y}(1+FP) + \rho_{I,PZ}\sqrt{2FP}\right]$$

- So the false positives actually offset some of the positive correlation.

$$\rho_{I,PZ} = \frac{C(I,PZ)}{\sqrt{V()}}$$

B Quotes

I prefer to think of a statistical sensibility rather than statistical thinking. It's "less than an agenda but more than an attitude." It allows for methodological preference while avoiding dogma. Paired with data analytic humility and I think you have proper "data science"

"Routine statistical questions are less common than questionable statistical routines..." McCullagh (2005).

If an issue can be addressed nonparametrically then it will often be better to tackle it parametrically; however, if it cannot be resolved nonparametrically then it is usually dangerous to resolve it parametrically." (p.96)

A test of meaningfulness of a possible model for a data-generating process is whether it can be used directly to simulate data." (p.104). In our current setting, this most certainly related to simulation while accounting for measruement error.