On the estimation of the total number of SARS-CoV-2 infections

Carlos Hernandez-Suarez[a,*], Polo Verne[b], Efren Murillo-Zamora[c]

[a] *Facultad de Ciencias, Universidad de Colima, Bernal Diaz del Castillo 340, Colima, Colima, 28040, MEXICO*
[b] *World Bank, 1818 H St, NW Washington, DC 20433, USA*
[c] *Departamento de Epidemiología, Unidad de Medicina Familiar No. 19, IMSS, Av. Javier Mina 301, 28000, Colima, Colima. MEXICO*

Abstract

We introduce a simple methodology to estimate the total number of infected with SARS-CoV-2 based on the number of deaths in households with at least one confirmed case of COVID-19. If we are willing to assume that a single member of a household with n members will infect the remaining members with probability 1, then the number of deaths in a household follows a binomial distribution with parameters $(n-1, p)$ where p is the CFR. Although the method may be affected by classification errors, its simplicity will allow to reduce the error of the estimates by increasing the sample size, since it requires minimal laboratory testing capabilities. We illustrate our methodology with data from Mexico and estimate the CFR in 0.34 %, that is, we estimate that the total number of infections is about 300 times larger than the number of deaths. We specify some dataset limitations. In comparison, using the number of deaths to date and a recently published results from random tests in Iceland, we calculated the ratio estimated infections/deaths in about 200 for that country.

Keywords: COVID-19, SARS-CoV-2, CFR, Asymptomatics, Immunity, Total Infections

*Corresponding author
Email addresses: carlosmh@mac.com (Carlos Hernandez-Suarez), pverme@worldbank.org (Polo Verne), efren.murilloza@imss.gob.mx (Efren Murillo-Zamora)

1. Introduction

It is known that the immune response to SARS-CoV-2 range from fully asymptomatic to exhibit mild or even severe responses that may cause death. Estimates of the probability of presenting a particular response is useful for prevention and attention purposes or even for building appropriate mathematical models that may provide some projections at the population level, specially to analyze the evolution of the immune population with the purpose of economic recovery. These estimates are particularly important to estimate the total number of infections by expanding the fraction of observed in some category, for the instance the number of hospitalized persons or the number of deaths.

Let $\mathbf{p} = [p_1, p_2, \ldots, p_s]$ be the probabilities that an individual will develop reaction i from a possible set reactions, for instance: $S = \{$None, Mild, Severe, Death$\}$ or any other categorization that can be associated to an individual without error and where the categories are mutually exclusive. The idea is that if the number of individuals in some category k is known or can be approximated, say n_k, and its proportion p_k can be estimated, then the total number of individuals in all categories can be estimated with n_k/p_k.

There are current estimates of the probability of showing a specific reaction to infection, for instance, being asymptomatic, presenting mild or severe symptoms [1, 2, 3, 4], but their statistical properties are unknown. A possible design that would allow to estimate $\mathbf{p}$ is random screening for infection or antibodies, and categorizing the response of infected or already immune individuals. The press has announced ongoing studies of this type to estimate the share of immunes which would allow to estimate the spread of the disease, but these studies may face some bias depending on the level of randomness, since in most trials participation is voluntary and individuals that were exposed or believe that were exposed may feel encouraged to participate, contributing to overestimate the spread of the disease.

The ideal random sample would be one extracted from a census database making sure that those dead from the disease are included. Nevertheless, this

is expensive because it requires comprehensive laboratory testing. However, if the fraction of infected or recovered is small, since only those infected at some time provide information, the cost per unit of information may be very high. In addition, a small sample would result in estimates with large confidence intervals.

Here we suggest a simple study design based on the outcomes of households in which there has been at least one infected individual.

Methodology

Let's define an *effective contact* or *contact* for short as any act between an infectious and a susceptible individual that would result in the infection of the susceptible [5]. Let's suppose that we are presented with an individual that had a *contact*, this individual will then provide information on the likelihood of presenting a reaction in the set S. If we are presented with a sample of n individuals that we know had a *contact* with some infectious individual (not necessarily the same infectious individual) then if x_i is the number of individuals that exhibit reaction i, $\hat{p_i} = x_i/n$ is an estimate of p_i, the probability than an individual will develop reaction i to infection. The variance of the estimate is $\hat{p_i}(1 - \hat{p_i})/n$.

From here, the importance of finding individuals that we know had a *contact*. But these individuals are easy to find: several studies have shown that household transmission as well as familial transmission is high [6, 7, 8, 9, 10, 11, 12] or even in offices for relative short interactions [13]. Therefore, if we are willing to concede that all the members of a household with a diagnosed individual had a *contact* with the initial infected in the household, the fraction of the remaining members of the household that exhibited reaction i is an estimate of p_i and we can pool data from several households to obtain a better estimate. In what follows, we formalize this estimate.

Define a household as an *infected household* if there has been *at least* one confirmed COVID-19. Suppose we have several infected households, with n_j

inhabitants in house j. Call the initial infected in the household *infected zero*. Assume that:

(i) The *infected zero* will infect the remaining $n_j - 1$ susceptible in the household with probability 1.

(ii) Once infected, the responses of individuals in an *infected household* are independent, that is, the responses of the remaining susceptible members in a household follows a multinomial distribution with parameters $n_j - 1$ and $\mathbf{p}$.

Observe that (i) implies that when two or more individuals are infected in the household, the probability that any one of the remaining susceptible will be infected is not increased. Also, it implies that all infected individuals are equally infectious, regardless of their symptomatic response to infection.

In our approach, it is required to know the total number of individuals in a specific category of responses. The simplest approach is to consider the number of deaths, as this is likely the most reliable observed indicator to proxy the corresponding statistics in the population. Hereafter, we will refer exclusively to this response to infection and thus our set consists of two responses S={Recovered, Dead}. This is preferable to use than the total number of individuals attending hospitals or receiving intensive care, for instance, which depends on the availability of health facilities and case definitions, which may vary between countries. Thus, the individual responses within a household follow a Bernoulli distribution with parameter p, where p is the Case Fatality Ratio (CFR).

Estimation of total number of infections

A list of confirmed cases can be used to obtain a sample of *infected houses*. Suppose that sample is of size m. Let n_j be the size of household j and $n = \sum_j^m n_j$ be the sum of all members in all households in the sample. Let x_j be the number of deaths (excluding *infected zero*) in household j and let $x = \sum_j^m x_j$. The estimate of p, the CFR measured at the household level is $x_j/(n_j - 1)$. Using all households data in the sample, the estimate of p is:

$$\hat{p} = \frac{x}{n-m} \tag{1}$$

with variance $\hat{p}(1-\hat{p})/(n-m)$.

With one further assumption, one can estimate the number of infections for the total population from these same data. If we assume that the number of COVID-19 deaths recorded includes all deaths from COVID-19, we can simply estimate the number of infected people in the population by expanding the fraction of infected people estimated from the sample of observed households. This should provide a simple but statistically sound estimate of the total number of infected people in the population.

The estimate of the total number of infections per death is about $\theta = 1/\hat{p}$. The approximate variance of $\hat{\theta}$ is:

$$\mathrm{Var}(\hat{\theta}) = \frac{1-\hat{p}}{(n-m)\hat{p}^3}$$

Let M be the total number of deaths from COVID-19 in the population, the estimate of the total number of infected individuals in the population, N is:

$$\hat{N} = M\frac{(1-\hat{p})}{\hat{p}} + M = M/\hat{p} \tag{2}$$

with approximate variance:

$$\hat{\sigma}^2_N = \frac{M^2}{n-m}\frac{1-\hat{p}}{\hat{p}^3} \tag{3}$$

The effect of external infection in the household

The probability that, among the remaining susceptible in the household, one or more will become infected by a different individual than *infected zero* is negligible, mainly because of the comparative pressure of *infected zero* on all members of the household. Nevertheless, assume that this happens and one of the susceptible in the household is infected by someone outside the household. At a glance, it seems that the correct estimate at the household level is now $x_i/(n_j-2)$ because there are only $n_j - 2$ remaining susceptibles, but this is

incorrect. The simplest explanation is the following: we know that in an *infected household* with n_j members, there are $n_j - 1$ individuals that have been subject to a *contact*. If one of the members of the household has a *contact* with an individual outside its household, its response still counts regardless of where the infection was acquired. Recall that we are estimating the probability of having a specific reaction to infection, not the probability of infection. This is the rationale we use to select a member at random from the duplicates in the list of deaths. It is not relevant who is the *infected zero*, we only need to ensure there was enough pressure of infection to guarantee a *contact*.

Example

In this example we build an approximation to (1) using a database from Mexico's IMSS (Instituto Mexicano de Seguro Social), the Mexican Institute for Social Insurance. The main problem with the database is estimating how many households there are (m) and the total population living in those m households, n. This is due to the fact that state, county, city and street are known, but in most cases there is no street number, so, in this approximation we considered two cases in the same street as belonging to the same household, which underestimates the number of households. Observe that the denominator in (1) can be written as $m(\mu - 1)$, where μ is the average household size, thus this approach tends to overestimate $\hat{p}$ in (1).

The database has 1180 confirmed COVID-19 cases from March 2 to April 16, 2020. Outcome of cases (death, recovered) was missing in several cases which were excluded. In an attempt to consider only households with final outcomes we excluded cases with symptoms onset in the last 21 days, that is, we considered only cases from March 2 to April 11, 2020. We also removed cases with lost addresses, leaving a final sample of 502 cases. The mean age of this final set was 47.3 years with a standard deviation of 16.1 years with median 47 years. From these, there were 61 % males and 39 % females. In this set, 43 % were at least 50 years old.

The total number of households was $m = 488$ and there were a total of $x = 3$ deaths. Since the total number of individuals in all households in the sample (n) i s not known, we vary the average household size in the sample (μ) to calculate $n = m\mu$ and estimate $\hat{p}$ using (1). The results are summarized in Figure 1.

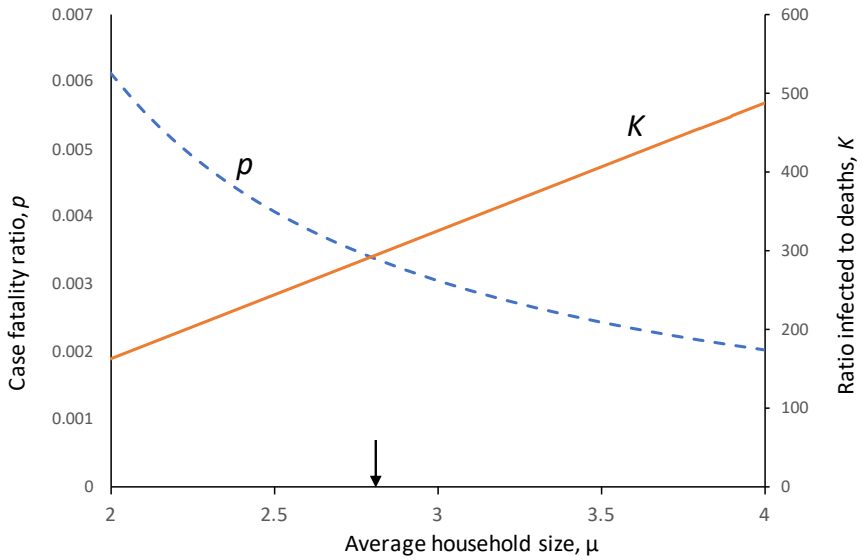

Figure 1: Plot of μ, the average household size vs $\hat{p}$ and $K = 1/\hat{p}$, the ratio of the total infected to deaths. The arrow points the average family size for México, 2.8 At this average, $\hat{p} = 0.0034$ and $K = 293$.

Discussion

First we must mention that our goal here is not to provide precise estimates of p for Mexico but to illustrate a simple methodology to estimate the true number of infections in a population using available information on confirmed individuals. As mentioned before, the database we used does not allow for a direct calculation of the number of households which is underestimated and thus, the CFR is overestimated.

Our estimate from the IMSS data at $\mu = 2.8$ is $p = 0.0034$, which is 3.5 times smaller than the CFR for the *Diamond Princess* with CFR= 0.012 and mean

age of 58 years [14] and three 3.4 larger than the reported so far for the *USS Theodore Roosevelt*, with CFR= 0.001 with an evident lower mean age [15]. In conclusion, we estimate one death per 300 infected individuals.

A recent study in Iceland [16] reports that from 2,283 persons selected at random there where 13 positive to SARS-CoV-2, for an prevalence of 0.0057 The total Iceland population is 364,000 thus an estimate of the total number of infected is about 2,073 On the other hand, the reported number of deaths in Iceland as of April 23 is 10, so we estimate that the number of infections to be $10/\hat{p} = 10/0.0034 = 2,941$ Recall the number of deaths depends on the availability and quality of health services in every country, thus, the estimate of the CFR and the number of deaths should be calculated ideally from the same region. Special care must be taken since both, the number of observed deaths in Iceland and the study in Mexico are small.

The method presented here is simple enough to be applied in countries with relatively few tracking capabilities. All it is needed is a list of households (a sample may suffice) with the total number of members in the household and the number of deaths for COVID-19 in each household. The precision of estimate (1) depends on the sample size m, and the precision of estimate (2) depends in addition on how good is our estimate of the actual number of deaths from COVID-19 to date. Overall, the precision will depend on our ability to diagnose COVID-19 related deaths.

Assumption (i) is central for this proposal, but there is a way to avoid it although clearly at a larger economic cost: this consists in testing all the members of the household of a confirmed case. The estimate (1) can still be applied using only data of confirmed cases, but now x is the number of deaths among all confirmed cases in all households (excluding the *infected zero*) and n the total number of confirmed cases in all households (including the *infected zero*).

In a following step, we can obtain the same probabilities for the whole population of positive cases by matching the household sample of tested households with households in the census. In other words, we only need to make sure

that the sample of households retained from the interviews is representative of the national sample of households. This can be done, *ex-ante* with a sample of available infected households or, if this information is not available, *ex-post* by matching the interviewed sample of households with the national census of households. Something that can be done with matching or machine learning methods. This provides the distribution of cases between any categorization of symptoms for the population of infected people in a population. A direct approach from stratified sampling may use some demographic knowledge of the population which would allow us to weight for differential response to the infection. Suppose that we classify a population in K categories (e.g., age) at relative frequencies f_i. Let $x^{(i)}$ and $n^{(i)}$ be respectively the total number of deaths and total number of individuals in category i in all households in the sample of size m, then a better estimate of p would be:

$$\hat{p} = \sum_{i=1}^{K} f_i \theta_i, \qquad \text{with } \theta_i = \frac{x^{(i)}}{n^{(i)} - m} \tag{4}$$

with variance

$$\hat{p} = \sum_{i=1}^{K} f_i^2 \, \frac{\theta_i}{n^{(i)} - m} \tag{5}$$

This $\hat{p}$ must be plugged in (2), with variance (3). We can divide then population in Mexico in two categories: age ≤ 50 years and age > 50 years, at respective proportions $f_1 = 0.9$ and $f_2 = 0.1$ [17]. The CFR in the first category was 0.002 and in the second 0.0052. From (4) we have $\hat{p} = 0.0023$ for the whole population, the weighted estimate suggests the number of total infected is about 400 times larger than the number of deaths.

One of the most important sources of bias in this method, is that some observations may be censored. Perhaps death has not occurred yet in a given household and thus the probability of death is underestimated. We tried to control this by using only data where the onset of symptoms was at least 21 days old so that the outcome is very likely observed, but in principle, we should

use households were there is enough evidence to believe that we can observe final outcomes.

Conflict of interest

Authors declare no conflict of interest.

Funding

This work is part of the program "Building the Evidence on Protracted Forced Displacement: A Multi-Stakeholder Partnership". The program is funded by UK aid from the United Kingdom's Department for International Development (DFID), it is managed by the World Bank Group (WBG) and was established in partnership with the United Nations High Commissioner for Refugees (UNHCR). The scope of the program is to expand the global knowledge on forced displacement by funding quality research and disseminating results for the use of practitioners and policy makers. This work does not necessarily reflect the views of DFID, the WBG or UNHCR. This study had approval R-2020-601-07 by the Health Research Ethics Committee (601) of the IMSS.

References

[1] Y. Liu, L.-M. Yan, L. Wan, T.-X. Xiang, A. Le, J.-M. Liu, M. Peiris, L. L. Poon, W. Zhang, Viral dynamics in mild and severe cases of covid-19, The Lancet Infectious Diseases.

[2] K. Mizumoto, K. Kagaya, A. Zarebski, G. Chowell, Estimating the asymptomatic proportion of coronavirus disease 2019 (covid-19) cases on board the diamond princess cruise ship, yokohama, japan, 2020, Eurosurveillance 25 (10) (2020) 2000180.

[3] H. Nishiura, T. Kobayashi, T. Miyama, A. Suzuki, S. Jung, K. Hayashi, R. Kinoshita, Y. Yang, B. Yuan, A. R. Akhmetzhanov, et al., Estima-

tion of the asymptomatic ratio of novel coronavirus infections (covid-19), medRxiv.

[4] J. T. Wu, K. Leung, M. Bushman, N. Kishore, R. Niehus, P. M. de Salazar, B. J. Cowling, M. Lipsitch, G. M. Leung, Estimating clinical severity of covid-19 from the transmission dynamics in wuhan, china, Nature Medicine 26 (4) (2020) 506–510. doi:10.1038/s41591-020-0822-7.
URL https://doi.org/10.1038/s41591-020-0822-7

[5] F. Brauer, C. Castillo-Chavez, Z. Feng, Mathematical Models in Epidemiology, Springer, 2019.

[6] Y. Bai, L. Yao, T. Wei, F. Tian, D.-Y. Jin, L. Chen, M. Wang, Presumed asymptomatic carrier transmission of covid-19, Jama.

[7] J. F.-W. Chan, S. Yuan, K.-H. Kok, K. K.-W. To, H. Chu, J. Yang, F. Xing, J. Liu, C. C.-Y. Yip, R. W.-S. Poon, et al., A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster, The Lancet 395 (10223) (2020) 514–523.

[8] Z. Hu, C. Song, C. Xu, G. Jin, Y. Chen, X. Xu, H. Ma, W. Chen, Y. Lin, Y. Zheng, et al., Clinical characteristics of 24 asymptomatic infections with covid-19 screened among close contacts in nanjing, china, Science China Life Sciences (2020) 1–6.

[9] P. J. Lillie, A. Samson, A. Li, K. Adams, R. Capstick, G. D. Barlow, N. Easom, E. Hamilton, P. J. Moss, A. Evans, et al., Novel coronavirus disease (covid-19): the first two patients in the uk with person to person transmission, Journal of Infection.

[10] X. Pan, D. Chen, Y. Xia, X. Wu, T. Li, X. Ou, L. Zhou, J. Liu, Asymptomatic cases in a family cluster with sars-cov-2 infection, The Lancet Infectious Diseases 20 (4) (2020) 410–411.

[11] G. Qian, N. Yang, A. H. Y. Ma, L. Wang, G. Li, X. Chen, X. Chen, A covid-19 transmission within a family cluster by presymptomatic infectors in china, Clinical Infectious Diseases.

[12] P. Yu, J. Zhu, Z. Zhang, Y. Han, A familial cluster of infection associated with the 2019 novel coronavirus indicating possible person-to-person transmission during the incubation period, The Journal of infectious diseases.

[13] C. Rothe, M. Schunk, P. Sothmann, G. Bretzel, G. Froeschl, C. Wallrauch, T. Zimmer, V. Thiel, C. Janke, W. Guggemos, et al., Transmission of 2019-ncov infection from an asymptomatic contact in germany, New England Journal of Medicine 382 (10) (2020) 970–971.

[14] T. W. Russell, J. Hellewell, C. I. Jarvis, K. Van-Zandvoort, S. Abbott, R. Ratnayake, S. Flasche, R. M. Eggo, A. J. Kucharski, C. nCov working group, et al., Estimating the infection and case fatality ratio for covid-19 using age-adjusted data from the outbreak on the diamond princess cruise ship, medRxiv.

[15] New York Times. Sailor on Roosevelt, whose captain pleaded for help, dies from coronavirus [online] (April 13, 2020) [cited April 22,2020].

[16] D. F. Gudbjartsson, A. Helgason, H. Jonsson, O. T. Magnusson, P. Melsted, G. L. Norddahl, J. Saemundsdottir, A. Sigurdsson, P. Sulem, A. B. Agustsdottir, B. Eiriksdottir, R. Fridriksdottir, E. E. Gardarsdottir, G. Georgsson, O. S. Gretarsdottir, K. R. Gudmundsson, T. R. Gunnarsdottir, A. Gylfason, H. Holm, B. O. Jensson, A. Jonasdottir, F. Jonsson, K. S. Josefsdottir, T. Kristjansson, D. N. Magnusdottir, L. le Roux, G. Sigmundsdottir, G. Sveinbjornsson, K. E. Sveinsdottir, M. Sveinsdottir, E. A. Thorarensen, B. Thorbjornsson, A. Löve, G. Masson, I. Jonsdottir, A. D. Möller, T. Gudnason, K. G. Kristinsson, U. Thorsteinsdottir, K. Stefansson, Spread of sars-cov-2 in the icelandic population, New England Journal of Medicine 0 (0) (0) null. arXiv:https://doi.org/10.1056/NEJMoa2006100, doi:

10.1056/NEJMoa2006100.

URL https://doi.org/10.1056/NEJMoa2006100

[17] Encuesta Intercensal, INEGI, Recovered from: http://www. beta. inegi. org. mx/proyectos/enchogares/especiales/intercensal.