ESTIMATING CUMULATIVE COVID-19 INFECTIONS BY A NOVEL "PANDEMIC RATE EQUATION"

David H. Hamilton [1]

Mathematics, University of Maryland, College Park

August 17, 2020

Abstract

A fundamental problem dealing with the Covid-19 pandemic has been to estimate the rate of infection, since so many cases are asymptomatic and contagious just for a few weeks. For example, in the US, estimate the proportion $P(t) = N/330$ where N is the US total who have ever been infected (in millions)at time t (months, $t = 0$ being March 20). This is important for decisions on social restrictions, and allocation of medical resources, etc. However, the demand for extensive testing has not produced good estimates. In the US, the CDC has used the blood supply to sample for anti-bodies. Anti-bodies do not tell the whole picture, according to the Karolinska Instituet [2], many post infection cases show T-cell immunity, but no anti-bodies. We introduce a method based on a difference-differential equation (dde) for $P(t)$. We emphasize that this is just for the present, with no prediction on how the pandemic will evolve. The dde uses only $x = x(s)$, which is the number/million testing positive, and $y = y(s)$, the number/million who have been tested for all time $0 \le s \le t$ (months), with no assumptions on the dynamics of the pandemic. However, we need two parameters. First, ρ, the ratio of asymptomatic to symptomatic infected cases. Second, τ, the period of active infection when the virus can be detected. Both are random variables with distribution which can be estimated. For fixed ρ, we prove uniform bounds

$$(1+\rho)\frac{x(t)}{\rho y(t)+1} \le P(t) \le (1+\rho)\, x(t)\ ,$$

are best possible, with range depending on τ. One advantage of our theory is being able to estimate P for many regions and countries where x and y is the only information available.

[1] dhh@umd.edu or Professor D.H.Hamilton, Dept Math., UMCP, Md 20742

Competing Interest Statement

The author have declared no competing interest.

Funding Statement

No external funding was received.

Author Declarations

I confirm all relevant ethical guidelines have been followed, and any necessary IRB and/or ethics committee approvals have been obtained.

Yes

The details of the IRB/oversight body that provided approval or exemption for the research described are given below:

N/A

All necessary patient/participant consent has been obtained and the appropriate institutional forms have been archived.

N/A

I understand that all clinical trials and any other prospective interventional studies must be registered with an ICMJE-approved registry, such as ClinicalTrials.gov. I confirm that any such study reported in the manuscript has been registered and the trial registration ID is provided (note: if posting a prospective study registered retrospectively, please provide a statement in the trial ID field explaining why the study was not registered in advance).

Yes

I have followed all appropriate research reporting guidelines and uploaded the relevant EQUATOR Network research reporting checklist(s) and other pertinent material as supplementary files, if applicable.

Yes

1 Introduction

There has been some unfortunate modeling of the Covid epidemic[7]. We will stay away from predicting the future, but instead, try to understand the immediate past by using some biology, stats, and analysis, to develop a transparent formula to estimate the proportion $P(t)$ of the US population that have *ever* been infected by the Coronavirus 19 at time t (months, where zero is March 20). This uses the number $y(s)$ of those tested/million, and cases $x(s)$/million testing positive, over the time interval $0 \leq s \leq t$. The aim is to overcome the problem that testing misses those past the $1-3(?)$ week period of infection. The other problem is that asymptomatics are often not being tested, [14], [17], [18], [13]. Asymptomatics represent proportion $r = \rho/(1+\rho)$ of all infected, with the wide estimate $0.5 < r < 0.95$. Ideally, one might try large scale testing for anti-bodies. However, many anti-bodies tests are unreliable, some infected do not develop antibodies. A recent study by King's College [5] showed that a majority of post infection patients loose most of their anti-bodies within a month. In any case, there has been no widespread testing for anti-bodies. On June 24, the CDC [9] announced new estimates using sampling from regional US blood labs, where on average, 8% had anti-bodies. Considering the numbers testing positive in those regions, this implied $\rho \sim 11$. So, it is important to have a good estimate of P.

To reiterate, we introduce a new method to estimate $P(t)$, the proportion of the population who would test positive for the virus at any time before time t, i.e. all those who have ever been infected. This is a more robust measure as the pcr test, etc., for the virus is fairly reliable. Our method also requires parameter τ, the time period during which individuals test positive. We show τ is quite significant, indeed, we show how it affects the CDC estimate of ρ.

As " for every infection only $\rho/(1+\rho)$ is symptomatic", a simple estimate is

$$P \sim (1+\rho)x = P_1 \qquad\qquad (PRE\ 1)\ ,$$

with PRE standing for "pandemic rate equation". However, this assumes only symptomatic cases are counted, and ignores large scale testing which uncovers and counts asymptomatic cases.

Trying to count the asymptomatic cases, leads to the quadratic

$$x = (1-r)P + rP(y-(1-r)P)$$

which has solution $P_2(t) =$

$$\frac{ry + (1-r) - \sqrt{(ry+(1-r))^2 - 4xr(1-r)}}{2r(1-r)} \geq \frac{(1+\rho)x}{\rho y + 1} \quad (PRE\ 2)$$

This second estimate assumes all testing done once at time t, whereas testing is spread over time. Furthermore, those who have been infected test positive only for a relatively short time, and testing past this period will not count them. Nevertheless, at the beginning of the pandemic, the bounds P_1 and P_2 were close. Now (at $t = 4.6$), we find for the US, assuming $r \sim .93$ with $x(4.6) = 1.5\%$ and $y(4.6) = 19.5\%$, that $P_1 \sim 21.4\%$ while $P_2 \sim 6\%$.

Our main result models P as the solution of the "pandemic rate equation":

$$P'(t) = \frac{x' - ry'(P(t) - P(t-\tau))}{(1-r)(1-r(P(t)-P(t-\tau))}\ , \qquad (PRE)$$

with initial conditions $x(t) = y(t) = P(t) = 0,\ t \leq 0$. Our assumption is that symptomatic cases are immediately treated, and thus tested (and counted). In actual fact, symptomatic cases take about a week before going for treatment/testing[2]. This delay could be entered as another random variable, but it is simpler to understand that our estimate will always be about a week out of date. Also, we assume that asymptomatic cases are discovered by essentially random testing from the entire population.

On June 24, the CDC[10] estimated $r = .9125$, i.e. the ratio of asymptomatic cases to symptomatic is about $\rho = 11$. The results of the Karolinska Inst.[2] shows $\rho \sim 14$ is more plausible, i.e. $r = .93$. Of course, τ and r are random variables whose distributions can be estimated. Simulations with lognormal distributed $\tau = 0.5 \pm 0.25, r = 0.93 \pm 0.03$ (68% CI), our best estimate is

$$P = 18\% \pm 4\% \qquad 68\%\ \text{CI (August 8, 2020)}$$

with the main error coming from uncertainty in ρ.

[2]Our first version[1] used another model, where symptomatic cases are not immediately tested. We found that even with a delay of two weeks, there was no significant difference from the present immediate response model- whose mathematics is much easier to handle.

2 The Theory

Now for the mathematics, first the complete statements, then the proof.

2.1 The Formulae

As before, $P = P(t)$ is the proportion of those in the population who would ever test positive for the virus in time interval $[0, t]$ months, with initial time $t = 0$, March 20. As before, the parameters r and τ denote the proportion of the infected who are asymptomatic, and the length of infection, respectively. To begin with, we now assume the parameters $r$and τ are fixed for the population. The functions x, y, are smoothed with derivatives x', y'.

We prove that P satisfies a nonlinear difference-differential equation:

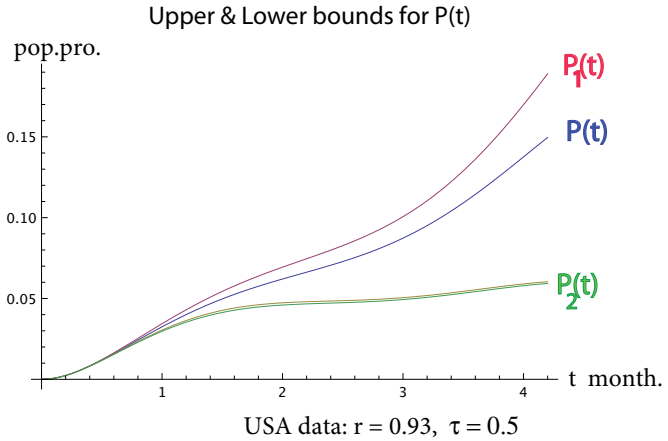
$$P'(t) = \frac{x' - ry'(P(t) - P(t-\tau))}{(1-r)(1-r(P(t) - P(t-\tau))} \qquad (3)$$

with initial conditions $x(t) = y(t) = P(t) = 0,\ t \leq 0$. The two simple minded estimates are essential bounds for the general case.

THEOREM *Suppose that x, y, P, are increasing functions. Then, for fixed ρ, independently of τ the solution of (3) satisfies the bounds*

$$\frac{(1+\rho)x(t)}{\rho y(t) + 1} \leq P(t) \leq (1+\rho)x(t)\ ,$$

and, furthermore, these bounds are best possible.



2.2 Derivation of Pandemic Rate Equation

We prove PRE. There is a large population of total size n million. Also, we use $X = X(s)$ as the number who have tested positive, and $Y = Y(t)$ as the number who have been tested up till time $s \leq t$ (months). As before, the parameters r, τ, denote the proportion of the infected who are asymptomatic, and the length of infection, respectively. We now assume the parameters r, τ, are fixed for the population of size n. People are sick for a time interval $[t - \tau, t]$ when they test positive. So, it is only during this time interval that the virus can be detected. Consider a very short time interval $[t - \Delta t, t]$, during which the quantities x, y change by $\Delta x = x(t) - x(t - \Delta t)$, $\Delta y = y(t) - y(t - \Delta t)$. The data for testing y and cases x is discrete and discontinuous, but we assume to be differentiable. Thus $\Delta x \sim x'(t)\Delta t$, $\Delta y \sim y'(t)\Delta t$.

The total number of cases comes from those who are symptomatic, take themselves for treatment and hence are tested; and those, who we assume are randomly tested, and a certain proportion turn out to be infected (but asymptomatic). We assume symptomatic and asymptomatic are only infected for a time period of length τ during which they test positive. There is the question of how quickly symptomatic cases go for treatment. We assume they immediately go for testing. [3]

We also use $Z(t)$ the total number of symptomatic cases, and $W(t)$ the number of asymptomatic cases. Evidently, $X = Z + W$. At time t, the number of past infections is $nP(t)$, so in time period $[t - \Delta, t]$, the number of new symptomatics is $(1 - r)nP'(t)\Delta t \geq 0$, assuming P increasing. Thus,

$$Z'(t)\Delta t = (1 - r)nP'(t)\Delta t$$

There is also $\Delta Y \sim Y'(t)\Delta t$ new tests, which counts those who sought help, and those asymptomatic cases caught up in essentially random sampling. Thus, not counting the symptomatics gives

$$Y'(t)\Delta t - (1 - r)nP'(t)\Delta t \geq 0 \ ,$$

assuming $Y' \geq Z'$. The proportion of these testing positive is

$$r(P(t) - (P(t - \tau))$$

[3] In [1] we considered the plausible idea that there was delay $\sim \tau$.

Thus, the new asymptomatics is

$$\Delta W = r(P(t) - (P(t-\tau))(Y'(t) - (1-r)nP'(t))\Delta t$$

Now, as $\Delta X = \Delta Z + \Delta W$, the new cases during $[t-\Delta, t]$ is $\Delta X \sim X'(t)\Delta t =$

$$\{(1-r)nP'(t) + r(P(t) - (P(t-\tau))(Y'(t) - (1-r)nP'(t))\}\,\Delta t$$

With $\Delta t \to 0$, and $x = X/n, y = Y/n$ gives the delay-differential equation

$$x' = (1-r)P' + r(P(t) - P(t-\tau))(y' - (1-r)P')$$

which simplifies to (3).

The nonlinear dde cannot be solved explicity in general, but is well suited for numerical solutions such as NDSolve in Mathematica, which has routines for delay-differential equations. The formula is for fixed ρ and τ, but in reality there is a distribution of values. We handled these by stochastic simulations.

2.3 Proof of THEOREM

This was motivated by considering two extreme cases. First case:
$\tau = 0 \Rightarrow P(t) - P(t-\tau) = 0$, and the dde becomes

$$P'(t) = \frac{x'}{1-r} \Rightarrow P(t) = \frac{x}{1-r}$$

Secondly: once infected always infected, i.e. $\tau = \infty \Rightarrow P(t-\tau) = 0$, dde is

$$P'(t) = \frac{x' - ry'P(t)}{(1-r)(1-rP(t))} \qquad (4)$$

We are assuming x, y, are both are increasing with $x' \leq y'$. Writing $y = x + u$ we have $u' > 0$ and $\int_0^t u'ds = y(t) - x(t)$ viewed as constraints on u. Hence,

$$P(t) = \int_0^t P'(s)ds = \int_0^t \frac{x'(s)}{1-r}ds - \int_0^t \frac{ru'(s)}{1-r}\frac{P(s)}{1-rP(s)}ds$$

Now $0 < P(s) < 1$ is increasing so

$$\frac{P(s)}{1-rP(s)}\ ,\ 0 \leq s \leq t\ ,$$

is maximized at $s = t$. Therefore, the second integral is maximised if u' is a Dirac measure concentrated at $s = t$. It follows that P has lower bound

$$\frac{x(t)}{1-r} - \frac{r(y(t)-x(t))}{1-r}\frac{P(t)}{1-rP(t)} = \frac{x(t)-ry(t)P(t)}{(1-r)(1-rP(t))}$$

Solving for $P(t)$ gives

$$P(t) \geq \frac{ry+(1-r)-\sqrt{(ry+(1-r))^2-4xr(1-r)}}{2r(1-r)} \geq \frac{x}{ry+(1-r)}$$

which proves the lower bound. Observe that for smooth x, y, this lower bound is not achieved, even in the limit as $\tau \to \infty$. Indeed, numerical solution of (4) for given x and y provides a somewhat better lower bound than P_2.

Finally, we prove $P(t)$ is sandwiched between P_1 and P_2 by showing the solution is monotone decreasing in τ. Consider

$$H(t,\tau) = \frac{x'(t)-ry'(t)(P(t)-P(t-\tau))}{(1-r)(1-r(P(t)-P(t-\tau)))}$$

Now, as $0 < r < 1$ and $x' < y'$ we see that $H \geq 0$ as $0 < P(t) - P(t-\tau) < 1$. As τ varies from 0 to ∞, we observe $P(t) - P(t-\tau)$ increases from 0 to $P(t)$. Define function

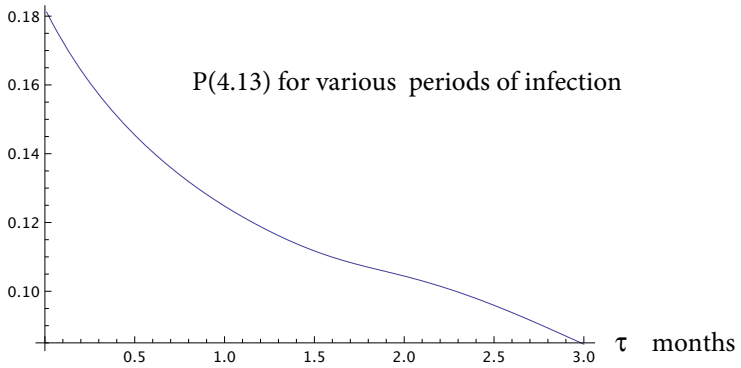
$$h(u) = \frac{1-au}{1-u}\ , a \geq 1\ ,$$

which is decreasing as

$$h'(u) = \frac{-a}{(1-u)^2} \leq 0\ ,$$

Thus $P' = H$ is monotone decreasing in τ. Hence, P decreases as τ increases.

We also carried out simulations for varying τ which gives consistent results:



2.4 Estimating ρ and τ

The estimate of ρ (or r) is from a variety of sources, collected in [12]. In China, for children $r \sim .95$; at a NYC hospital where all pregnant women were tested $r \sim .9$; while $r \sim .5$ for passengers on cruise ships. WHO currently has $r \sim .8$ which is consistent with results from Iceland, where it is estimated $r > 0.8$ (where many "symptomatics" really had the flu which reduced the ratio).

We expect r to be different for different populations, indeed, be age dependent. Nevertheless, we want an average value for the entire population. By our formula, for early in the epidemic, we have $P \sim (1 + \rho)x$. Now, on March 29, for NYC, the CDC[9], [8] used anti-bodies to estimate $P \sim .07$, while $x \sim .004$, giving $r \sim .943$. Later, on May 29, the CDC[10] estimate was $P \sim .23$, while $x \sim .024$, giving $r \sim .9$. Similar results were found in different regions of the US giving the CDC average estimate $r = 0.9$. However, the Karolinska Inst. [2] estimates at least 25% more infections from their study of T-cell immunity. So, we take $r = 0.93 \pm 0.03$.

We make the important observation that $P \sim (1+\rho)x$ is only valid for small x, y, t, so using it to estimate ρ on present US data underestimates ρ by $\sim 20\%$. This underestimate will grow as x and y become larger.

3 Results

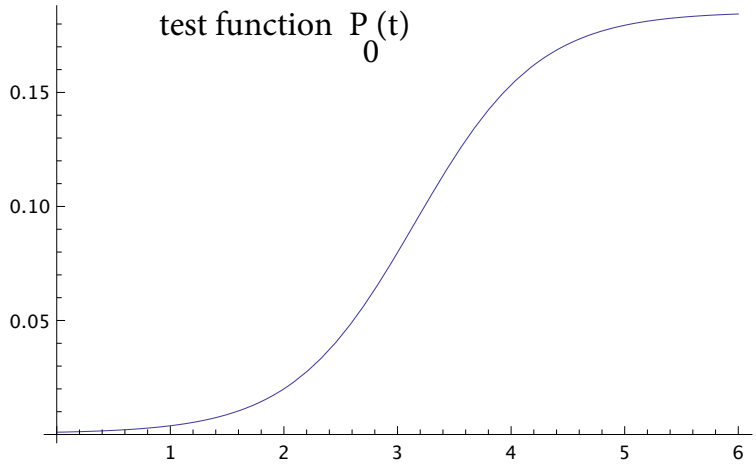
Our method was first tested on a hypothetical case. Then we use USA data.

3.1 Test cases

As nobody knows the true level of infection, we test by using a hypothetical infection which has proportion P_0 in the population governed by:

$$P_0' = R(P_0(t) - P_0(t - \tau_0))(1 - 2P_0(t))$$

This assumes about 50% of the population can be infected (exactly once) and the period of contagion is τ_0. R is the rate of infection. Also some small initial value $P_0(t) = \phi(t),\ t < 0$ is required. For $R = 5$ and $\tau = 0.25$, this was solved numerically:

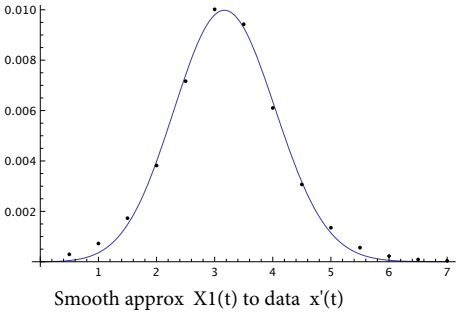


We need some testing function $y'(t)$, choose $y' = 0.1$ constant. Then, the function x for the number of cases/mill. satisfies the dde

$$x' = (1 - r_0)P_0' + r_0(P_0(t) - P_0(t - \tau))(y' - (1 - r_0)P_0')$$

where the proportion $r_0 = 0.9$ or $\rho_0 = 9$ is given.

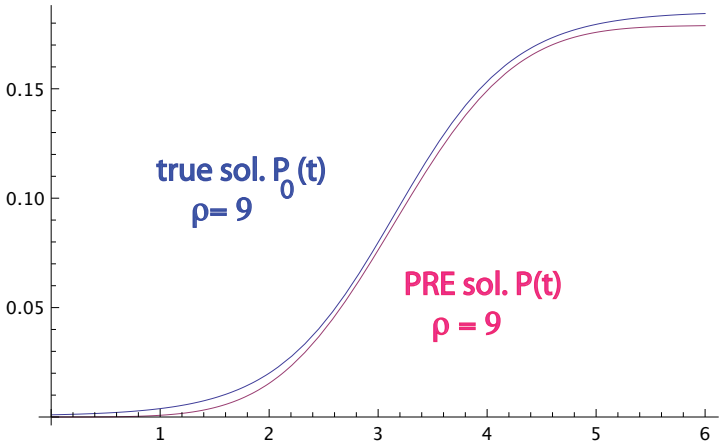
Next, we apply our process: first approximate x', y' by $X1, Y1$. A not very good approximation is shown:



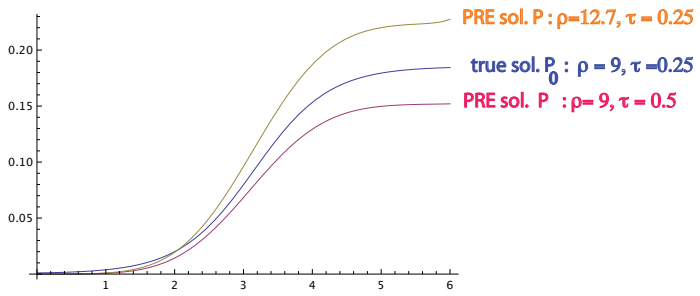
Then, obtain the predicted proportion $P(t)$ using PRE:

$$P'(t) = \frac{X1 - rY1(P(t) - P(t-\tau))}{(1-r)(1-r(P(t)-P(t-\tau)))}$$

Now, of course, we have no apriori way of choosing the right ρ and τ. So these "would be obtained by field data". However, if we choose the right ones, in this case $\tau = 0.25, \rho = 9$



We find if r, τ, are close to the true r_0, τ_0, then the predicted proportion P is close to the true proportion P_0. However, if say r is twice r_0, then P is approximately twice the value of P_0. On the other hand, the solution P is relatively less sensitive to variations in the period of infection τ.



What this means is that the PRE is well posed, indeed, it is linear in x', y' and roughly linear in ρ.

3.2 USA data

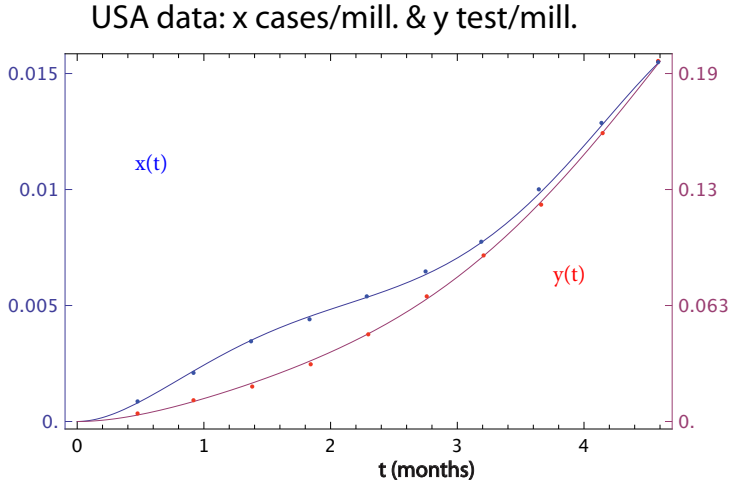
For the US over time, $0 \leq t \leq 4.6$ months, the data is approximated by

$$y(t) = (0.01539t^2 - 0.00422t^3 + 0.000661t^4)U(t) \ ,$$

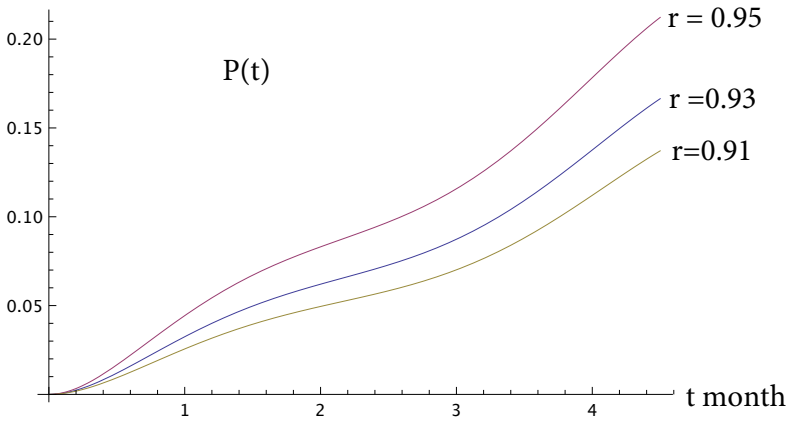
and by the function

$$x(t) = (0.00473t^2 - 0.0030094t^3 + 0.0007488t^4 - 0.00006139t^5)U(t) \ ,$$

using the Heaviside Step Function $U(t)$ to emphasize $x, y = 0$ for $t \leq 0$. This is simply a smooth interpolation and not meant to be any prediction, although the 3-4 term polynomials are a good fit for data at 11 points.



We then solve PRE for P with various values of r:



USA data , τ= 0.5

4 Conclusions

Our theory shows that it is possible to estimate the proportion of all who have been infected by methods easy enough to perform on any computer with *Mathematica*. The estimate is as accurate as sampling, and stable. Thus, it is not necessary to sample the whole population.

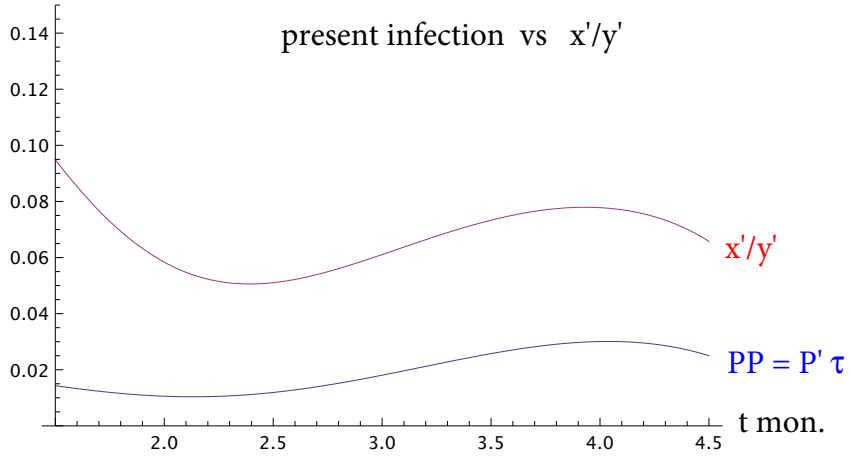
Using this method, one can make predictions on the saturation point, i.e. the proportion σ of susceptibles in the population. Of course, this depends also on the population profile, in particular the amount of older, sick people. Applying our formula to data from the states NY and NJ one sees that $P > 30\%$, which suggests that σ lies in the range $35-45\%$, which is consistent with their x curves.

The formula shows proportion presently infected is

$$PP(t) = P(t) - P(t-\tau) \sim P'(t)\tau = \frac{(x' - ry'(P(t) - P(t-\tau)))\tau}{(1-r)(1-r(P(t) - P(t-\tau)))}$$

Various authorities have tried to estimate PP: they use the proportion of infected among those daily tested , analytically this is the function $x'(t)/y'(t)$.

Comparing $x'(t)/y'(t)$ with $P'(t)\tau$ for USA data:



USA data : r =0.93, $\tau = 0.5$

At the present time, the $x'(t)/y'(t)$ estimate is out by about a factor two, but nonetheless, in the right ballpark (and erring on the safe side).

The CDC has used x/P to estimate ρ. This is accurate enough for small t, but for current time, our theory shows, with $\tau \sim 0.5$, that the CDC method underestimates ρ by at least 20%, with the discrepancy growing with time.

One could also make quick estimates of P for various nations. Our main example is the USA, which at $P \sim 18\%$, has not too far to go before reaching saturation. On the other hand, my native Australia, which had a severe lockdown, seems to have $P \sim 1\%$, i.e. it has a long, long way to go.

5 Data and computations

There is considerable data available from CDC, Oxford University [4], and Los Alamos[6]. All computations done on an iMac using *Mathematica*.

References

[1] Hamilton, D. H. *Estimating Cumulative Covis-19 Infections*, technical report, Department of Mathematics, UMD, July 2020.

[2] Karolinska Instituet COVID-19 Study Group: *Robust T cell immunity in convalescent individuals with asymptomatic or mild COVID-19* bioRxiv preprint: https://doi.org/10.1101/2020.06.29.174888. (posted June 29, 2020).

[3] coronavirus data/ CDC.gov

[4] *Our world in data* coronavirus/ourworldindata.org

[5] Jeffrey Seow,Carl Graham,Blair Merrick et al *Longitudinal evaluation and decline of antibody responses in SARS-CoV-2 infection* bioRxiv preprint: https://doi.org/10.1101/2020.07.09.20148429

[6] Los Alamos National laboratory. *COVID-19 Confirmed and Forecasted Case Data*. 2020: https://covid-19.bsvgateway.org.

[7] John P.A. Ioannidis et al, *Forecasting for COVID-19 has failed* Stanford Prevention Research Center, Department of Medicine, and Departments of Epidemiology and Population Health, of Biomedical Data Science, and of Statistics, Stanford University, and Meta-Research Innovation Center at Stanford (METRICS), Stanford, California, USA

[8] John P.A. Ioannidis *The infection fatality rate of COVID-19 inferred from seroprevalence data* medRxiv preprint doi: https://doi.org/10.1101/2020.05.13.20101253. posted July 14, 2020.

[9] *Large-scale Geographic Seroprevalence Surveys*. https://www.cdc.gov /coronavirus/2019- ncov/cases - updates/geographic - seroprevalence - surveys.html. Accessed June 8, 2

[10] *Commercial Laboratory Seroprevalence Survey data .* https://www.cdc.gov /coronavirus/2019- ncov/cases-updates/commercial-lab-surveys.html. Accessed July 21

[11] *Seroprevalence of Antibodies to SARS-CoV-2 in 10 Sites in the United States, March 23-May 12, 2020* Fiona P.Havers,CarrieReed, TravisLim, JAMA Intern Med.Published online July 21, 2020. doi:10.1001/jamainternmed.2020.4130

[12] https://www.cebm.net/covid-19/covid-19-what-proportion-are-asymptomatic

[13] Claire J. Steves et al *Estimates of the rate of infection and asymptomatic COVID-19 disease in a population sample from SE England* Department of Twin Research, King's College London, St Thomas? Hospital, London SE1 7EH, UK, Claire.J.Steves@kcl.ac.uk.

[14] Yang et al, *COVID-19 Asymptomatic Infection Estimation* National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China medRxiv preprint doi: https://doi.org/10.1101/2020.04.19.20068072.this version posted April 23, 2020.

[15] Masahiro Sonoo, et al *Estimation of the true infection rate and infection fatality rate of coronavirus disease 2019 in each country.* Department of Neurology, Teikyo University School of Medicine, Tokyo, Japan Corresponding Author: Masahiro Sonoo (sonoom@med.teikyo-u.ac.jp)

[16] Akiva B Melka, Yoram Louzoun, *Evaluation of the number of COVID-19 undiagnosed infected using source of infection measurements* medRxiv preprint doi: https://doi.org/10.1101/2020.06.09.20126318.this version posted June 15, 2020

[17] Nishiura H, Kobayashi T, Suzuki A,et al.*Estimation of the asymptomatic ratio of novel coronavirus infections (COVID-19).*Int J Infect Diseases,Published Online First: 13 March 2020.doi:10.1016/j.ijid.2020.03.020

[18] Karl M. Aspelund et al *Identification and Estimation of Undetected COVID-19 Cases Using Testing Data from Iceland* medRxiv preprint doi: https://doi.org/10.1101/2020.04.06.20055582.this version posted June 23, 2020

.