Covid-19 Data

Jason Schattschneider
UCCS Undergradate
Electrical and Computer Eng.
jschatts@uccs.edu

Abstract- With a lot of Covid-19 data and narratives being brought to light in media and social media. I thought it would be interesting to compare initial covid-19 projections vs where our numbers are today. This data should not be used in any way to predict or be used for medical research. This is a simple graphing and analysis of the initial projection's vs up to date numbers.

I. Introduction

As of writing this paper, it is the beginning of May 2020, which means we have now been living in the COVID-19 era for about 3 months in the United States and 5 months since the disease was observed in Wuhan, China. The goal of this paper is to understand why initial projections are/were off or in some cases close to the actual number.

In this report, I will explore some data from the IHME (Institute for Health Metrics and Evaluation) COVID-19 health service utilization team. This report was used quite widely for projections, one of the reasons being because it was funded by the Bill and Melinda gates foundation and that the IHME is an independent global health research center at the University of Washington.

First, I would like to go over the initial report as detailed as possible and how IHME developed its data set for the paper.

Secondly, I would like to present a merge of past and present IHME data. On their website they display their current data, but you cannot go back and compare projections to current. This is probably for good reason because people will want to point out obvious error with initial projections. But in this report, we will make non-conspiracy claims and justifications to why any data maybe out of place.

Disclaimer, I am not a statistician or medical professional. There is likely going to be error in my work because this is paper not peer reviewed and I am not out to disprove the work of IHME, or approve of their work. I also encourage readers to see the original report which is posted below.

II. THE Initial report

The initial report can be found here https://www.medrxiv.org/content/10.1101/2020.03.27.20043752v1.full.pdf .

This report was set out to answer the question "Assuming social distancing measures are maintained, what are the forecasted gaps in available health service resources and number of deaths from the COVID-19 pandemic for each state in the United States?"

Note: IHME paper was published March 30th, 2020.

Overall, I think the report does a good job using reliable measures. Although, one draw back is the one of the main data sets the report was essentially forced to use was data from Wuhan, China. Because this was the biggest data set that was available at the time.

There are reports that the CIA received intelligence that China was woefully underreporting the spread and death rate of the coronavirus in China (Barnes, 2020).

Now they do not state this in the report that China's numbers may have been altered, but this could have influenced their projections. They also could have altered their own simulations without reporting this although, this is strong speculation.

Other data sets they used in their projections were from local and national governments, and WHO websites. Other countries listed in there data set were Italy, South Korea, and at the time limited data from the US. They also considered when those governments made emergency declarations when to implement social distancing.

Finally, they combined this data set from all the countries and made an indirect standardized model. (Read full report for more detail in respect of standardized model).

III. Formula

In their projections they used a curve fitting tool of a parametrized Gaussian error function. Also known as the error function and sometimes notated classically as the erf(x) function.

Equation 1 Gaussian Error Function from IHME Report

$$D(t;\alpha,\beta,p) = \frac{p}{2}\left(\Psi(\alpha(t-\beta)\right) = \frac{p}{2}\left(1+\frac{2}{\sqrt{\pi}}\int_0^{\alpha(t-\beta)} \exp\left(-\tau^2\right) d\tau\right)$$

Now lets dive into the variables and then we will also rewrite the equation in a friendly notation.

Ψ = The Gaussian Error Function
P = Maximum death rate at each location
t = Time since death rate exceeded .31 per million
ß = Time at which death rate is maximum (per location)
α = Location specific growth parameter

Note: IHME used the .31 per million as a minimum for the disease hence time since death rate exceed .31 per million.

Now to show the formula in the more well-known notation just for reference.

$$D = \frac{p}{2}(1 + \text{erf}(\propto (t - ß)))$$

If you are unfamiliar with this equation here is a quick rundown of the equation from the Imperial college of London https://www.youtube.com/watch?v=26QbWYBCw7Y.

The basic premises is looking at a bell curve weather standard or modified.

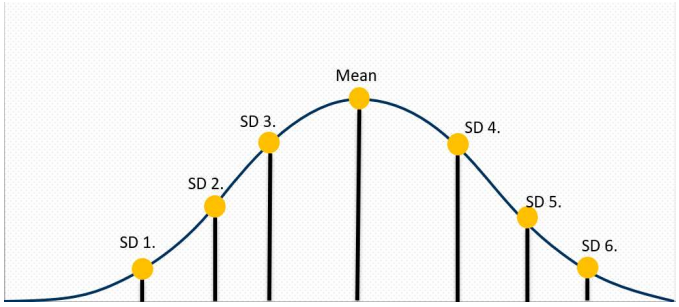


Figure 1: credit Jason Schattschneider

Above is an example of a normal bell curve and surprisingly enough the area underneath the curve is $\sqrt{\pi}$. This comes from the equation.

Equation 2

$$\sqrt{\pi} = \int e^{-t^2}\, dt$$

Where exp(-t^2) is the equation of a bell curve. This is important to how this study gets their data because then you can do definite integrals to approximate how many people maybe in that area (IE fit into death category or otherwise).

The final portion to make the erf equation (gaussian error function) is the add a $\frac{2}{\sqrt{\pi}}$ and this is to cancel the other side of the area.

When you add the 1+ to the erf() function, this says what is the area from negative infinity, till a point x or in this case ($\propto (t - ß)$.

Again, just to solidify the basics if we take the standard bell curve from Figure 2 and we have this equation.

Equation 3

$$y = (1/2)[1 + \frac{2}{\sqrt{\pi}}\int_{-\infty}^{mean} e^{-t^2} dt]$$

The area underneath the curve would now $\frac{\sqrt{\pi}}{2}$. Using this concept is how Statisticians can estimate how many people may fall into a category.

In this case and what we will be focusing on is how many daily deaths per day.

The good news is you can still understand the experiments I do in this paper without being an expert at this process. Because I simply model data, they derived from these initial projections and plot them against what we now know as the real numbers.

IV. Looking at Cases

Before we start looking at some of the data I complied I would like to go over the how and why.

IHME released their data points on their website and they update it daily. With what data came out from the real events.

So, if one day they project daily deaths to be 10 but then after that day passed and the actual death rate was 8 per day they go back and backfill the data. Also, looking at these numbers more closely this is a rate and not how many people died that day.

Additionally, the two data sets I grabbed from their website were from March 25th 2020, which will be used as our projected data set.

Then the data set I used for Actual is from 29 April 2020. My graphs I included do included dates past the **29 April point. On the Actual line (in the graphs) dates past 29 April will still be projections.**

To make these graphs I used matplotlib which is a common python library used for graphing data. It is a simple open sourced graphing software package like MATLAB but not tended to by a professional company like MATLAB.

I also used NUMPY another python library for doing some math operations. The data included in these sets are the same categorially. But the 3 I was mainly concerned about are the daily deaths per day on the actual set. Additionally, to see the lower and upper bound projections of daily deaths from the March data set to see if the projections at least landed within the bounds in some cases.

In this report I want to cover a worst-case projection and best-case projection. Where the projection was really inaccurate, and one were the projection was spot on.

Additionally, all the code and original graphs will be posted on GitHub here : https://github.com/schattz/Statsproject

V. Worst Case

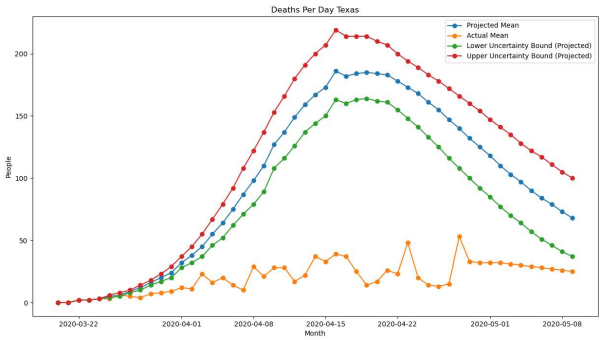


Figure 2: Texas State Plot

Please refer to appendix or GitHub page for larger photo.
Note: The "actual" numbers are noisy because these are exact numbers recorded by IHME. It is going to be "noisy" and not follow a true curve.

Texas's initial projections were not very accurate, not only did they not fall within the margin of error, but the daily death count was vastly overstated.

Looking at the raw CSV files located on GitHub or at this link: http://www.healthdata.org/covid/data-downloads

12 April 2020, being the most off number in this projection. Where the projected mean number for that day was 182 (rounded) daily deaths and a lower bound of 160 deaths finally, an upper bound of 214 deaths per day. The actual death rate for the day was 17.

Using a percent error calculation for these numbers.
Equation 4

$$\text{Percent Error} = \frac{Projected\ value - actual}{actual}$$

Projected Value	Actual	Error
Mean: 182	17	970.5882%
Upper bound: 214	17	1,158.8235%
Lower bound: 160	17	841.1765%

So why was this projection so far off?

Well some skeptics may say well the virus just isn't as bad as we thought.

But how would that explain New York's situation where they are devasted by the virus. Now a lot of this could be due to the proximity that New Yorker live next to each other.

One article states how Texas may have gotten a little "lucky" (Prohov, 2020) and they acted quickly to implement social distancing. The lucky portion was most students were on spring break from school when cases were starting to increase nationwide and those students never came back. By then Texas enacted social distancing measures.

It is also important to note Texas does have some of the worst testing rates in the country. While cumulatively they test a lot but per their population, they are one of the worst at 48th rank among 50 states.

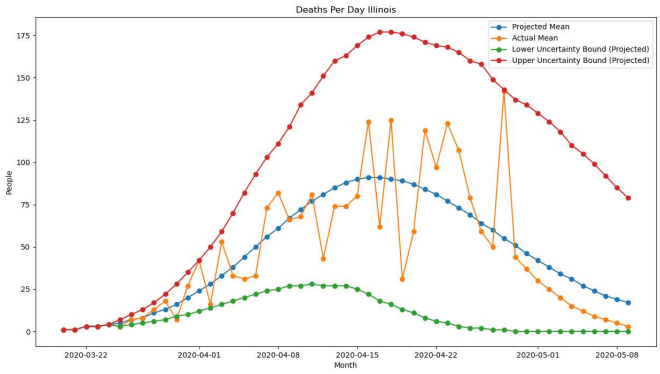
They have tested 777 per 100,00 people (As of April 24th). This can also account for lack of confirmed cases and confirmed deaths from that region. (Mekelburg, 2020)

Hopefully, there numbers continue to stay low and there is another reason they are performing so well in comparison to other states.

VI. Best Case Projection

For the best initial projection, It was probably between Illinois and New York. New York's numbers were worse than the projected, but its actual mean bell curve is pretty impressive in that it follows a true bell curve well.

But Illinois followed the initial projection closely.
Figure 3: Credit: Jason Schattschneider



Looking at this graph let us look at its worst point.
April 28th, 2020.

Projected Value	Actual	Error
Mean: 55	142	61.267605%
Upper bound: 1	142	99.29578%
Lower bound: 143	142	0.70422535%

Even at its worst point this projection was pretty good in comparison to the worst case. This projection kept the actual mean in between its error bounds as well as staying fairly accurate to the projection.

So why is the projection so much better?

As of April 24th, ,Illinois testing is 1,281 per 100,000 people. Better than Texas but not nearly close to Louisiana (best at testing) where they tested 3,000 per 100,000. (Mekelburg, 2020)

Which Louisiana's projections was actually very accurate as well but since it broke its lower bound a few times I ruled it as "best case".

I think the answers to these questions will require some extensive studying because correlation maybe hard to find or not exit.

VII. Disscusion

Is there any correlation to these results?

I took a look at a ranking chart of test's per 1 million (Molla, 2020) and at first glance (not exact) testing doesn't seem to have a direct impact to more accurate projections.

But this when these projections were being made testing was still very low in the United States. It does seem states that have been affected more seriously had more accurate projections. Michigan and New York's projections seemed to follow a curve well although they had some bound breaking.

But looking at these results I think the one true result is these projections are hard to project accurately.

Coming from a Computer Engineering background if I had a resistor that was working with a 66% error, I would not be happy.

But in Electromagnetism we understand the mechanics completely solved. In the medical industry there are thousands of more factors. How close are the people together, domestic, and international travel, climate, medical infrastructure, testing, economics, hygiene of area, underlying medical conditions (hundreds of these), age of population etc.

There is one thing everyone can agree on is more medical funding and preparation for Epidemic events. Just like when we have an economic crash, we need to have more measures in place to prevent a complete crash. I think as a global population this is our Hurricane Katrina.

We were not prepared, and our best statisticians and medical professionals could not more accurately predict this pandemic and doesn't seem to be to their own fault. There was a minimal data set to project this Pandemic and hopefully we come up with something equivalent to a Nyquist frequency. Where we must have 2 times the sampling rate to have a good sample of a signal. But in this case to make an accurate projection of the virus. (Now this maybe already established in a medical term, but I am not sure.)

VIII. Conclusion

I implore you too look at all the data sets and double check the data (I am not an expert). But reiterating what I think is the main takeaway is statistics is hard and making projections with small data sets is near impossible.

I think that is whereas a society we can improve. One being becoming a more mathematically literate society and be able to be given information and interpret the data honestly and fairly.

Provide more information to our public stating how accurate results maybe and say after looking at this data we think this and as a population be ok that it is not going to be 100% accurate.

IX. References

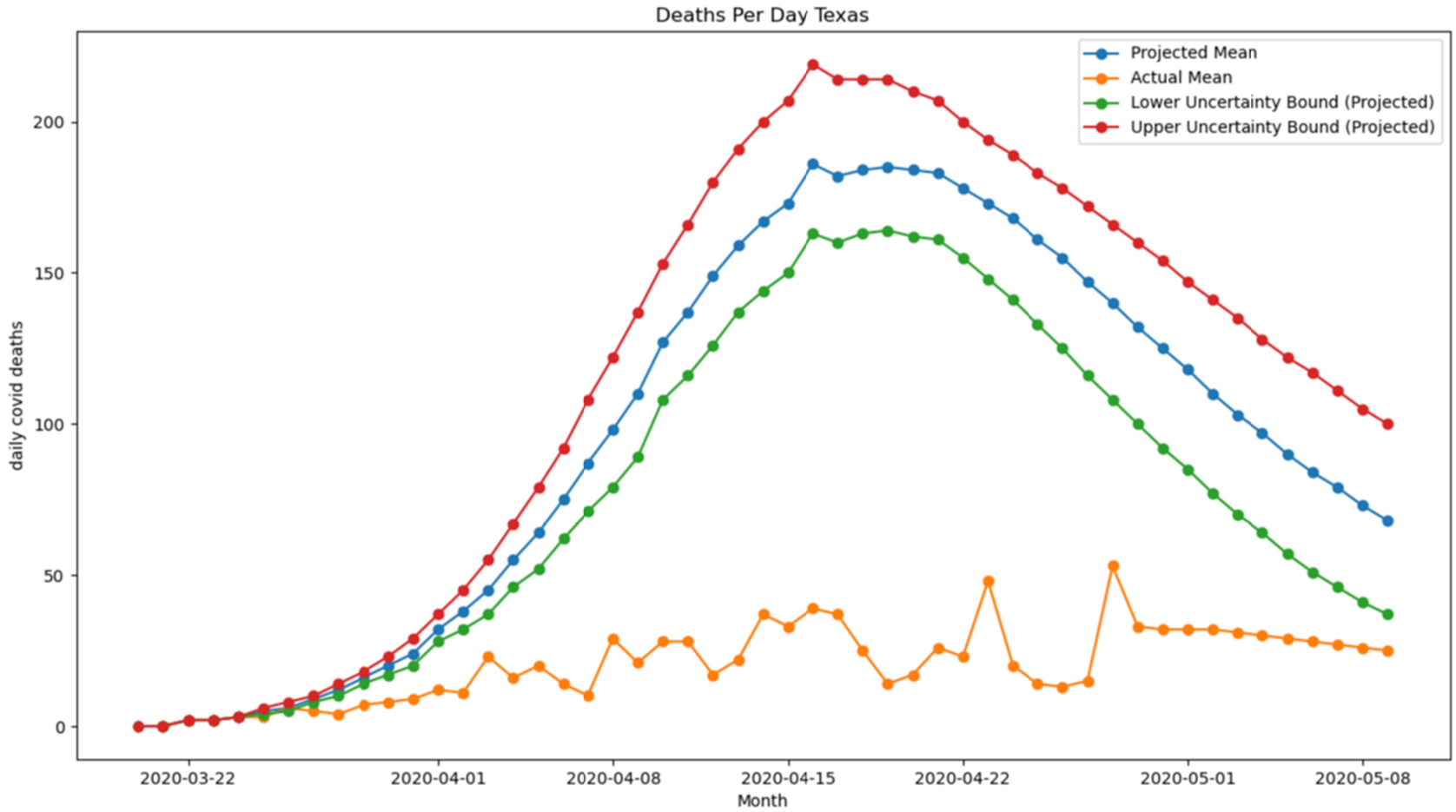
Barnes, J. E. (2020, 4 2). *C.I.A. Hunts for Authentic Virus Totals in China, Dismissing Government Tallies*. Retrieved from New York Times: https://www.nytimes.com/2020/04/02/us/politics/cia-coronavirus-china.html

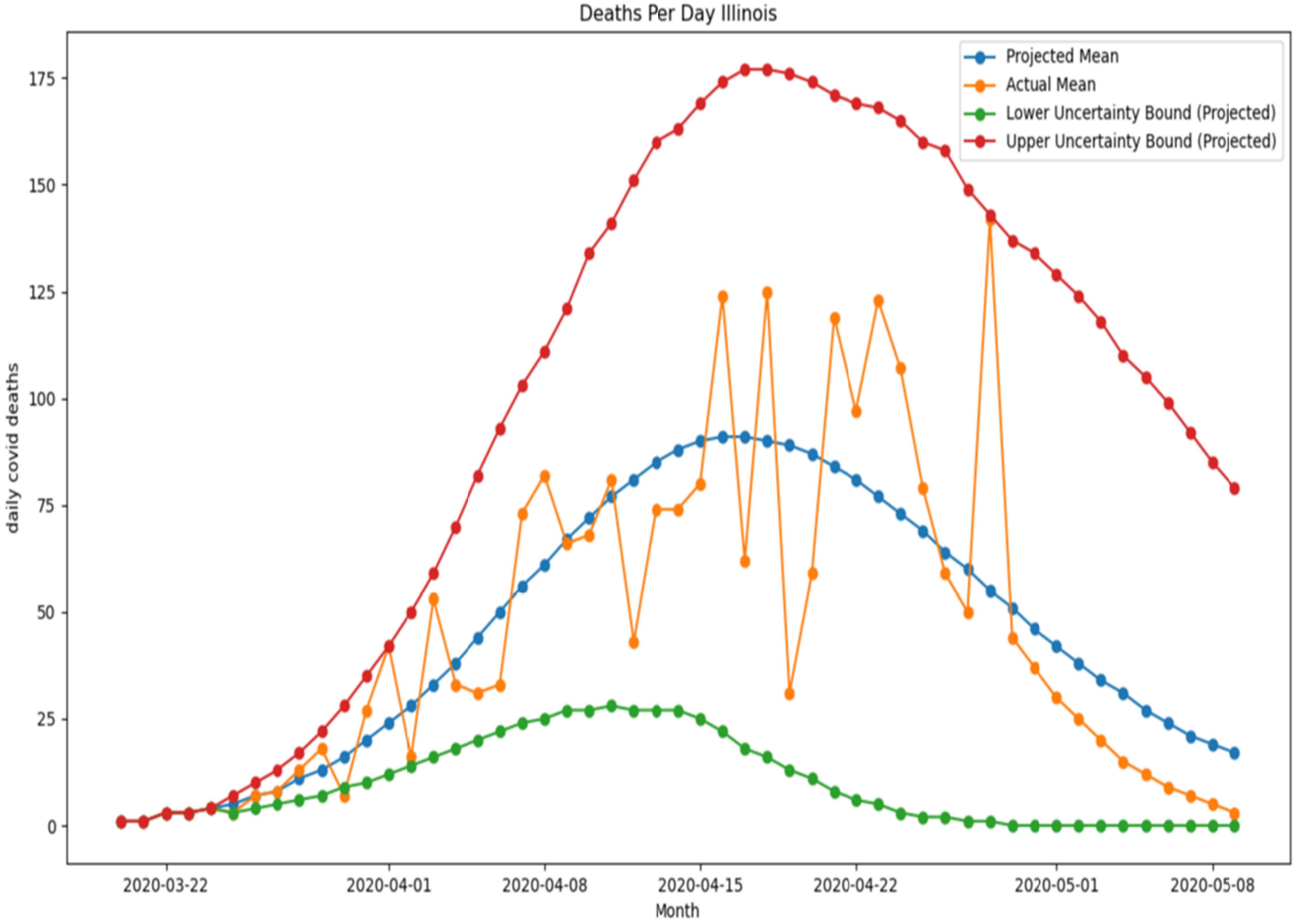
IHME COVID-19 health service utilization forecasting team. (2020). *Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilatordays and deaths by US state in the next 4 months.* Washington : IHME.

Mekelburg, M. (2020, 4 24). *Fact-check: Where does Texas rank on coronavirus testing?* Retrieved from Statesman: https://www.statesman.com/news/20200424/fact-check-where-does-texas-rank-on-coronavirus-testing

Molla, D. S. (2020, 05 4). *Charting the coronavirus pandemic state by state*. Retrieved from VOX: https://www.vox.com/2020/3/26/21193848/coronavirus-us-cases-deaths-tests-by-state

Prohov, J. (2020, 04 22). *When will COVID-19 deaths peak in Texas?* Retrieved from WFAA: https://www.wfaa.com/article/news/health/coronavirus/coronavirus-deaths-texas-university-peak-model-prediction-reopening-social-distance/287-705f8cbd-4d2f-4eff-8cd5-01ca67bc0f11


Deaths Per Day Texas
Projected Mean
Actual Mean
Lower Uncertainty Bound (Projected)
Upper Uncertainty Bound (Projected)
daily covid deaths
0
50
100
150
200
2020-03-22
2020-04-01
2020-04-08
2020-04-15
2020-04-22
2020-05-01
2020-05-08
Month


Deaths Per Day Illinois
Projected Mean
Actual Mean
Lower Uncertainty Bound (Projected)
Upper Uncertainty Bound (Projected)
daily covid deaths
0
25
50
75
100
125
150
175
2020-03-22
2020-04-01
2020-04-08
2020-04-15
2020-04-22
2020-05-01
2020-05-08
Month