Analysis of United States Covid-19 case number trends during late January 2021

DSC 680 Summer 2021

Chris Briggs

Abstract

The winter 2020-21 wave of COVID-19 cases in the United States are modelled using a modified Gaussian curve. These models are used to answer the question: was there an anomalous (from the perspective of such models) drop-off in new daily cases in late January 2021?

Background

Coronavirus disease 2019, abbreviated COVID-19, is an ongoing pandemic, affecting nearly every country in the world (Places Without Reported COVID-19 Cases, 2021). COVID-19 is caused by the pathogen SARS-CoV-2. As of June 2021, the first officially recognized case was in Wuhan, China in December 2019 (Page, Hinshaw & McKay, 2021), but investigations into the origin are ongoing. COVID-19 began sweeping the globe by at least early 2020.

The first COVID-19 case to be reported in the United States occurred in January 2020, causing President Donald Trump to declare a public health emergency shortly thereafter ("COVID-19 pandemic," 2021). Despite subsequently implemented public health control measures, the virus spread through all US states and territories. The behavior of the spread has varied by state. As of June 2021, the total case rate per population in the US varies from 26,000 per million in Hawaii to 145,000 per million in North Dakota. Differences may be due to environmental factors, such as humidity (Allen, Iwasaki, & Marr, 2020), policy, behavior, or other factors. Further supporting the environmental hypothesis is the fact that the virus has spread in distinct waves. Generally in each state, a primary wave during summer 2020 gave way to a decrease in cases through early fall 2020, with cases again rising to an absolute peak in winter 2020-21.

It is a central fact of epidemiology that the ability of a virus (to which natural immunity is possible) to spread depends, among other things, on what fraction of the population is susceptible to infection (Pandemics, 2021). With a novel virus, such as SARS-CoV-2 appears to be, early in the spread practically nobody will have natural immunity. This causes rapid spread of the virus. As more people are infected, the virus spreads more slowly. The reproduction number, or R0, is the number of people on average that an infected person will infect in turn (Biggers, 2020). Daily new cases increase as long as R0 is larger than 1, and they fall as soon as R0 is smaller than 1. The net result is that the cumulative number of infections follows a curve called a *logistic curve* ("Logistic function," 2021). Logistic curves are "S-shaped," levelling off to the right at some height (the total number of people ultimately infected). This height is called the *carrying capacity* of the virus.

The carrying capacity need not be constant. If changing environmental factors change the carrying capacity, then multiple waves of infection may be observed, and this is consistent with observations of COVID-19 (Maragakis, 2020).

So, following the basic theory of epidemiology, the total cases should follow a logistic curve day-over-day, and the daily new cases should follow a bell-shaped curve, also known as a Gaussian curve ("Gaussian function," 2021). However, during late January 2021, cases appeared to be declining precipitously, faster than would be the case if the curve had continued to follow a Gaussian shape. The present study aims to answer the question: was there a faster drop-off in cases during late January 2021 than would otherwise be expected from Gaussian modeling?

Statement of the Problem

During late January 2021, the number of new COVID-19 cases per day dropped precipitously across the United States. Was this drop expected given standard epidemiological modelling approaches, or did the drop in new cases exceed the expectations of such models?

Analysis Approach

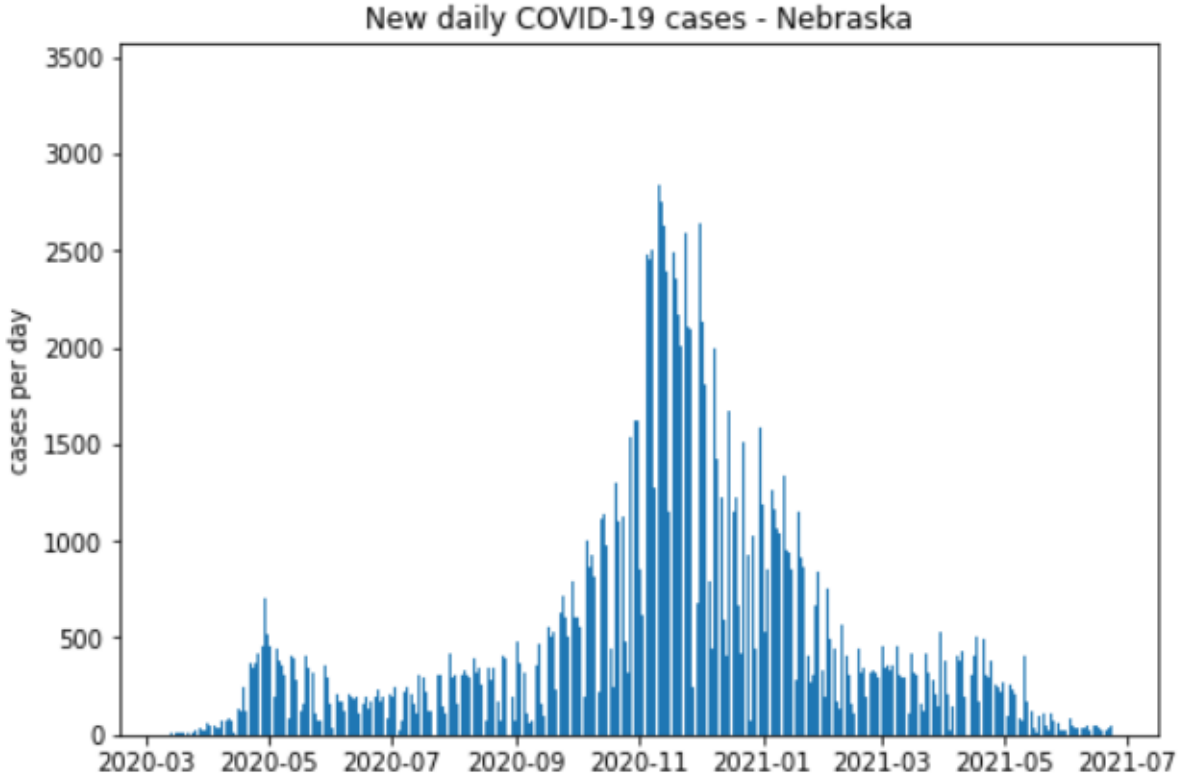
The main idea is to fit Gaussian curves to the data, make predictions in January, and measure the extent to which the actual data in late January deviates below predictions as compared to the deviation in the first half of January. Put another way, we're looking for a "break" in behavior in mid January, in which the curves do a comparatively fair job of predicting cases in early January, then tend to largely overpredict cases in late January.

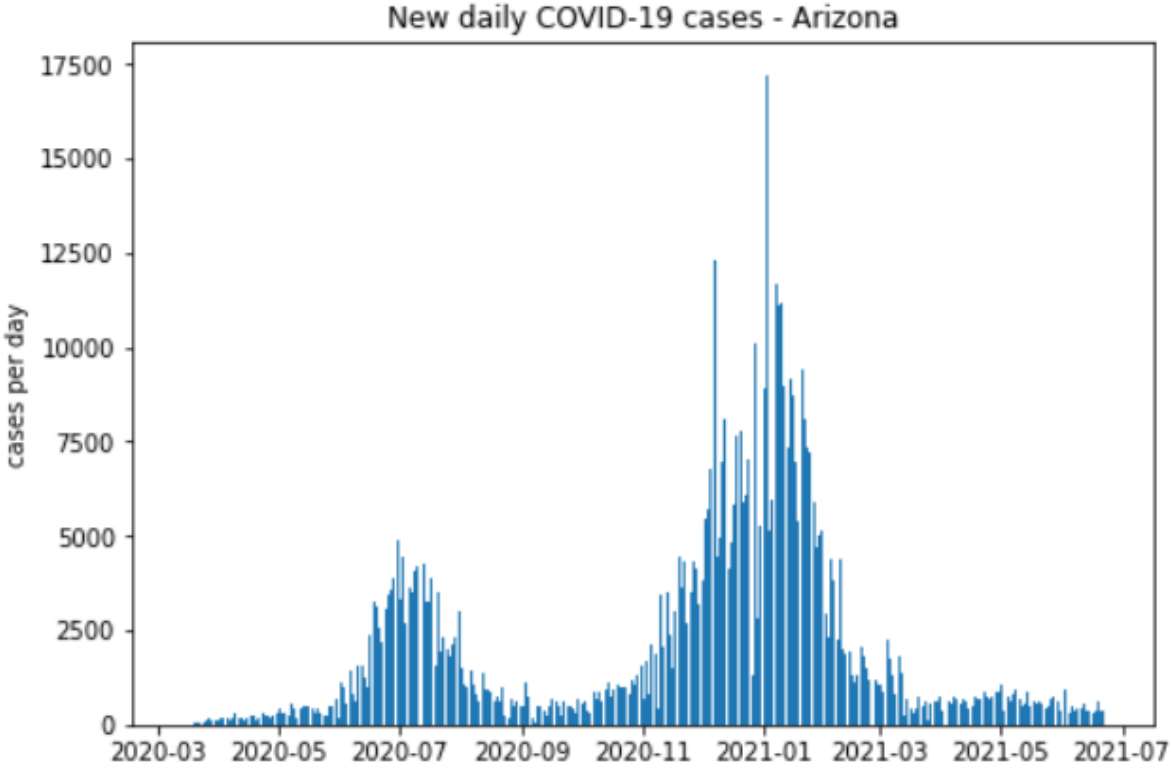
As the pandemic has occurred in multiple waves, a simple Gaussian curve is not fit for the data. There are several ways to proceed from this observation. Perhaps the simplest, and the one opted for here, is to isolate the winter wave and fit to it Gaussian curves. An alternative would be to seek a multi-wave model fit for the data.

This approach assumes veracity of the worldometers data and suitability of the winter wave of COVID-19 cases to modelling by Gaussian curves.

Exploratory Analysis

The winter wave does not start at the same date in each state. So, approximate start dates of the winter wave were manually determined and entered by hand. For a data set large enough to make this manual identification impractical, a good alternative solution would be to smooth the curves over seven or so days, and choose the Fall 2020 minimum point as the start of the winter wave. The following are typical:


New daily COVID-19 cases - Nebraska
cases per day
3500
3000
2500
2000
1500
1000
500
0
2020-03
2020-05
2020-07
2020-09
2020-11
2021-01
2021-03
2021-05
2021-07

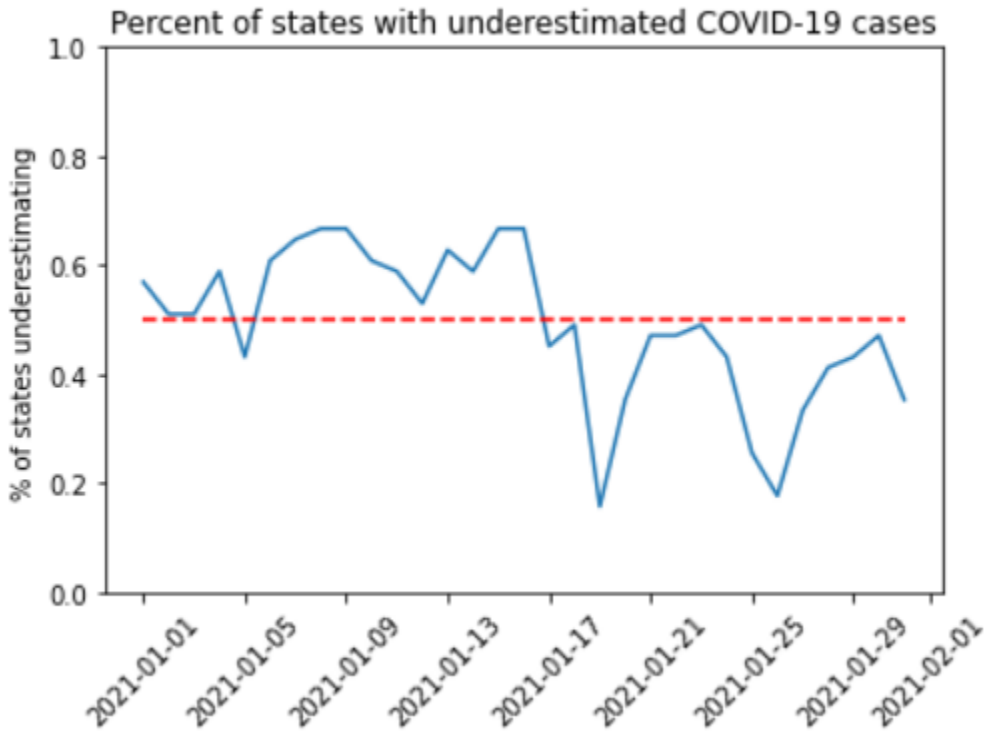

New daily COVID-19 cases - Arizona
cases per day
17500
15000
12500
10000
7500
5000
2500
0
2020-03
2020-05
2020-07
2020-09
2020-11
2021-01
2021-03
2021-05
2021-07

Exploratory analysis reveals that the Gaussian curve is not an ideal fit for the winter wave. Indeed, when matching the height of the wave, a Gaussian curve will drop off too rapidly (the observed tails are fatter than predicted by a Gaussian model). So, I will modify the fit curve by adjusting the exponent in the Gaussian curve. See Appendix A for a comparative discussion of these curves.

Analysis Results

For each state, the daily new cases were modeled, and the number of daily new cases was predicted for early January and late January as described above. The error was computed as observed minus predicted, as is standard. The fifty states could be viewed individually, but aggregate effects would be difficult to tease out of such a large collection of graphs. Aggregate figures can be found in Appendix B. To reduce the information and more clearly answer the question, I present two simplified views.

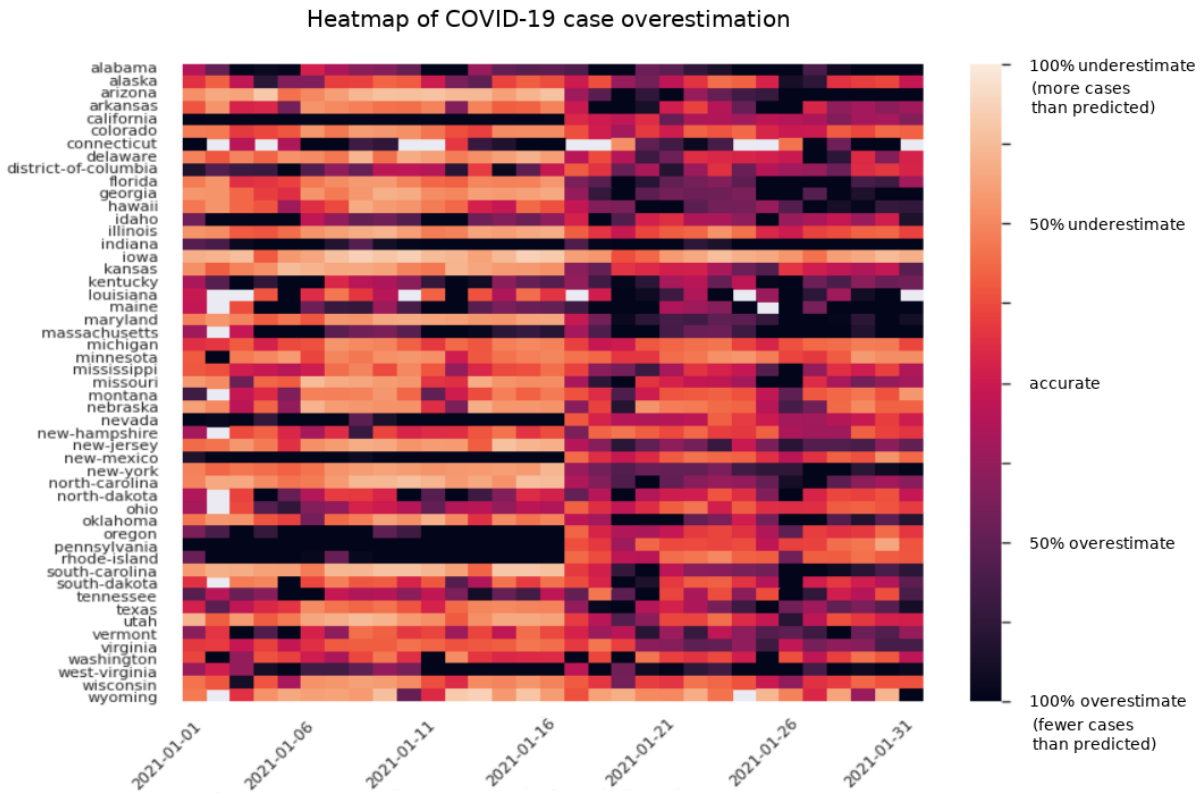
First, for each day in January, I computed the fraction of states in which the model produced an underestimate of the cases.



When the blue line is at the middle, that is along the red dashed line, half of the states produced overestimates and half produced underestimates. Due to random noise in daily cases, even a perfectly fit model would have some variation, so we would expect some variation around the middle line. What we see is that for the first half of January, the blue line is significantly above the red line, indicating higher than expected cases. Then, strikingly, the blue line falls

dramatically below the red line in mid January. This indicates that most states are producing overestimates for the daily case numbers during this period.

The second view shows a heat map with percentage errors in the prediction. The percentage error was computed as the (actual - predicted) / actual cases.



In this view, darker regions correspond to overestimated cases - actual cases being lower than predicted by the models. Again, we see significantly more darkness in the right half (late January) than the left half. Also conspicuous in this view are several states, such as Pennsylvania, Rhode Island, and New York, which were departing downward from the modeling curves already by early January, as indicated by the dark bars on the left half of the heatmap. Conversely, Iowa was still dealing with significantly more than the expected number of cases in the second half of January.

Conclusion

In conclusion, the evidence gathered supports an affirmative answer to the research question. Specifically, the number of COVID-19 cases dropped faster in late January than is otherwise explained by previous trends in the progression of daily cases.

Challenges Encountered

The first major challenge was to determine how to handle the fact that the winter waves began at different times in different states. As mentioned before, the solution was to record by hand these approximate start dates.

The second major challenge was that the Gaussian curve did not fit as expected from theory. By modifying the form of the curve, I was able to get a better-fit curve to the data.

When training and predicting, the first approach taken was to train on data up to a given day to predict each next day. There was a major drawback here: if there was indeed an unexpected drop-off in cases in the second half of January, then this phenomenon would begin to be incorporated into the fitted curve, obscuring the fact that these two weeks collectively violated previous trends. It was here that I decided to predict the first half of January from a single curve, and the second half from another curve.

Limitations

The results indicate that there was a faster-than-expected drop in cases during late January. The study cannot point to a cause for the identified phenomenon. Some commonly hypothesized explanations are: increasing COVID-19 vaccination rates, improved compliance with public health measures, changing seasons (although this does little to explain a mid-winter shift), or less testing (Why are coronavirus cases, 2021).

Moving Beyond (how it can be "taken to the next level")

Comparison with the control of the summer drop-off. This would significantly complicate the analysis, as the summer dropoff happened at different times in different states, so much care would be required to get an aggregate first-wave drop-off pattern to compare to the second wave drop-off.

Rather than modelling the winter 2020-21 wave as a single Gaussian curve, a more sophisticated curve could be fitted to the entire multi-wave data set. This curve would need to account for the varying carrying capacity of the virus (Meyer & Ausubel, 1999). The carrying capacity is of course not known, and will vary by region (according to e.g. local environment and policy). The carrying capacity may be deduced by the shape of the curves per region.

References

1. Allen, J., Iwasaki, A., and Marr, L. (2020, Nov 18). This winter, fight covid-19 with humidity. Retrieved from https://www.washingtonpost.com/opinions/2020/11/18/winter-covid-19-humidity/

2. Biggers, A. (2020, Apr 20). What Is R0?. Retrieved from https://www.healthline.com/health/r-nought-reproduction-number

3. COVID-19 pandemic (n.d.). Retrieved June 26, 2021 from https://en.wikipedia.org/wiki/COVID-19_pandemic

4. Gaussian function (n.d.). Retrieved June 26, 2021 from https://en.wikipedia.org/wiki/Gaussian_function

5. Logistic function (n.d.). Retrieved June 26, 2021 from https://en.wikipedia.org/wiki/Logistic_function

6. Maragakis, L. (2020, Nov 17). Coronavirus Second Wave? Why Cases Increase. Retrieved from https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/first-and-second-waves-of-coronavirus

7. Meyer, P and Ausubel, J. (1999, Feb 05). Carrying Capacity: A Model with Logistically Varying Limits. Technological Forecasting and Social Change 61(3):209--214, 1999.

8. Page J, Hinshaw D, McKay B (26 February 2021). "In Hunt for Covid-19 Origin, Patient Zero Points to Second Wuhan Market – The man with the first confirmed infection of the new coronavirus told the WHO team that his parents had shopped there". The Wall Street Journal. Retrieved 27 February 2021.

9. Pandemics: How Are Viruses Spread? (n.d.). Retrieved June 27, 2021 from https://www.nctm.org/Classroom-Resources/Illuminations/Interactives/Pandemics-How-Are-Viruses-Spread/

10. Places Without Reported COVID-19 Cases (2021, June 21). Retrieved from https://www.usnews.com/news/best-countries/slideshows/countries-without-reported-covid-19-cases

11. Why are coronavirus cases suddenly dropping? Here are 4 key factors. (2021, Feb 17). Retrieved from https://www.advisory.com/en/daily-briefing/2021/02/17/coronavirus-cases

Appendix

A. The modelling curve

A standard Gaussian curve is a function of the form

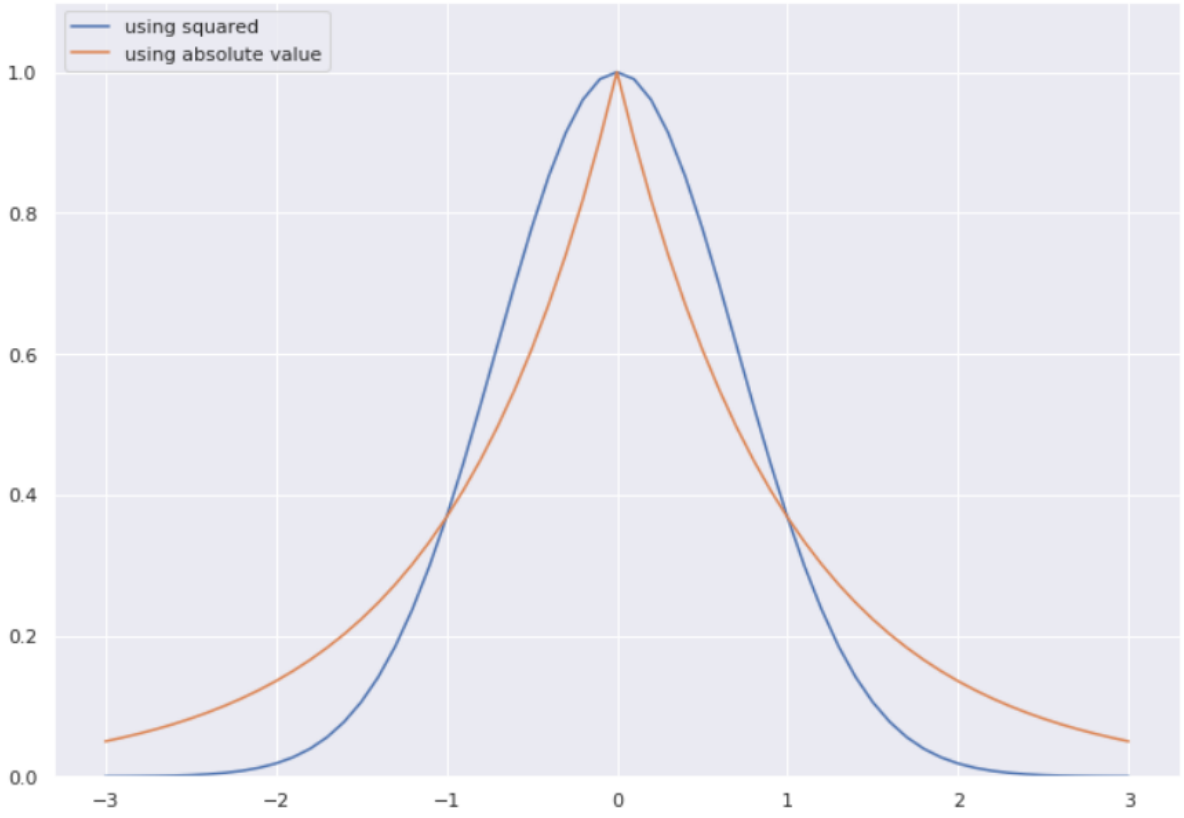
$$f(x) = a \cdot \exp\left(-\frac{1}{2}\left(\frac{x-b}{c}\right)^2\right).$$

The center of the curve is at $x = b$. As x is removed from b in either direction (to the right or left), the negative exponent on e increases rapidly. This rapid increase is largely due to the square in $\left(\frac{x-b}{c}\right)^2$. Another important feature of the square is that it gives the curve symmetry over $x = b$.

If we wish to fatten the tails of the distribution, we might take the absolute value instead of squaring. Our function then has the form

$$g(x) = a \cdot \exp\left(-\frac{1}{2}\left|\frac{x-b}{c}\right|\right).$$

For comparison, here are plots of f and g using the parameters $a = c = 1, b = 0$:



B. Aggregate predictions, observations, and errors for January

	predicted	actual	error
2021-01-01	278458	243566	-34892
2021-01-02	282631	221741	-60890
2021-01-03	287129	230621	-56508
2021-01-04	291695	208214	-83481
2021-01-05	296556	179535	-117021
2021-01-06	301852	236596	-65256
2021-01-07	307598	257252	-50346
2021-01-08	313820	279178	-34642
2021-01-09	320530	294659	-25871
2021-01-10	327755	257623	-70132
2021-01-11	335074	219580	-115494
2021-01-12	342552	207889	-134663
2021-01-13	350598	228948	-121650
2021-01-14	359242	230754	-128488
2021-01-15	368513	233786	-134727
2021-01-16	371900	239546	-132354
2021-01-17	222700	203937	-18763
2021-01-18	223231	179829	-43402
2021-01-19	223851	146818	-77033
2021-01-20	224557	175235	-49322
2021-01-21	225348	181710	-43638
2021-01-22	225913	193083	-32830
2021-01-23	226523	188172	-38351
2021-01-24	227162	171233	-55929
2021-01-25	227849	140955	-86894
2021-01-26	228609	143956	-84653
2021-01-27	229110	150957	-78153
2021-01-28	229400	153835	-75565
2021-01-29	229655	164933	-64722
2021-01-30	229995	164571	-65424
2021-01-31	230418	144399	-86019

Ten Anticipated Questions

1. **How can the model be made more accurate?**
   The approach here modelled only the winter 2020-21 wave. The model can be made more accurate by using a multi-wave approach, accounting for the varying virus carrying constant.

2. **Is there a way to control for environmental factors, such as humidity?**
   Not really - there are not two states identical in every way except for humidity. The pandemic did hit different regions at different times and rates, suggesting weather conditions were at play, but this is not proof.

3. **Is there evidence of environmental factors in the data?**
   The best example in the data may be Hawaii. Hawaii's weather, of course, varies little from season to season compared to the continental US. If weather affected the virus carrying constant, we would expect to see a flatter curve in Hawaii than elsewhere (lower peaks and higher tails). This is indeed what we observe. Hawaii's cases did still vary at roughly the same intervals as the mainland, and this might be explained by seeding of local outbreaks from the mainland.

4. **What factors may weaken this analysis?**
   This viral outbreak was extraordinary in the measures taken to counteract it. Economies around the globe were locked down to help suppress transmission. Periods of heavier and lighter restrictions are discontinuous and will cause sharp breaks in the data. Ditto for holidays, during which no cases are reported and more transmissions may occur, etc.

5. **Assuming the conclusion is valid, how might you explain the occurrence?**
   The four main commonly floated explanations are increasing vaccination rates, better adherence to public health directives, changing weather, and changes in testing behavior. The latter may include changes to policy (e.g. the number of tests being administered), or changes in the ways the tests are conducted, affecting sensitivity and specificity of the tests.

6. **Does your study support any of those explanations?**
   No. The study only identifies that there was a rapid drop-off, but does not explain why the drop-off may have occurred.

7. **What would need to happen in order to support one of those explanations?**
   That would require the testing of individual factors as best as can be done. The required information is certainly not in the data gathered here. Expanding the perspective to the whole globe may provide some extra insight - Israel made headlines for a very quick vaccination rate, and may be otherwise comparable to some other country. But the only way to really be sure is with controlled studies, which are not practical in this case for both ethical and practical reasons.

8. **Can the study be improved at all using only the information available?**
   Yes, to be more precise the study can be extended to an A/B test in order to give exact probabilities of a deviation from some prior trend.

9. **How do you justify skipping the late December data? Would this not skew results?**
   Yes, it certainly does. However, we have to do the best we can with the data we have. We can't eliminate the holiday effect from the data - it's there. The choices were (1) model using the holiday numbers, (2) impute data in its place, or (3) train without it. I chose the third. The second or third would be fair options, while the first would probably skew the model most.

10. **Does this study support any policy recommendations?**
    There's something going on with the numbers which is not entirely straightforward. It should be a high priority to study what factors may have contributed to the drop. This policy can have a huge impact on the health of businesses and individuals if it can anticipate earlier opportunities for reopening. For example, Texas Governor Greg Abbott was widely panned for lifting COVID-19 restrictions in early March. In hindsight, had other states followed suit, much economic damage may have been mitigated in the meantime.