Process Behavior Charts and Covid-19

The purpose of analysis is insight

Donald J. Wheeler

Since the start of the Covid pandemic I have received many questions about how to analyze the Covid numbers using process behavior charts. Various schemes have been proposed and used. This paper will discuss appropriate ways of analyzing data from epidemics and pandemics.

Now to be clear, data analysis is distinct from modeling. Epidemiological models incorporate subject-matter knowledge to create mathematical models that are useful for understanding and predicting the course of an epidemic. These models allow the experts to evaluate different treatment approaches. While these models are generally refined and updated using the collected data, this is not the same as what I call data analysis. Data analysis can be carried out by non-epidemiologists. This occurs when people try to use the data tell the story of what is happening. This article is about the analysis of the existing data by non-epidemiologists. Nothing in what follows should be construed as a critique of epidemiological models.

PLOT THE DATA

While our brains are not good at processing strings of numbers, they are good at recognizing patterns. This is why we should always start by plotting the data. With time-ordered data the essential plot is the running record or time-series graph. Figure 1 shows the time series for the daily number of deaths from Covid-19 for the U.S.. Here we see the initial growth and the effects of the overloading of the medical services and facilities in and around NYC in mid-April. Then, as the numbers settle down we begin to see a weekly pattern to the reported values. Without making any computations Figure 1 tells us a lot about these data. Figure 2 uses a seven-day moving average to smooth out the weekly pattern and show the underlying trend.

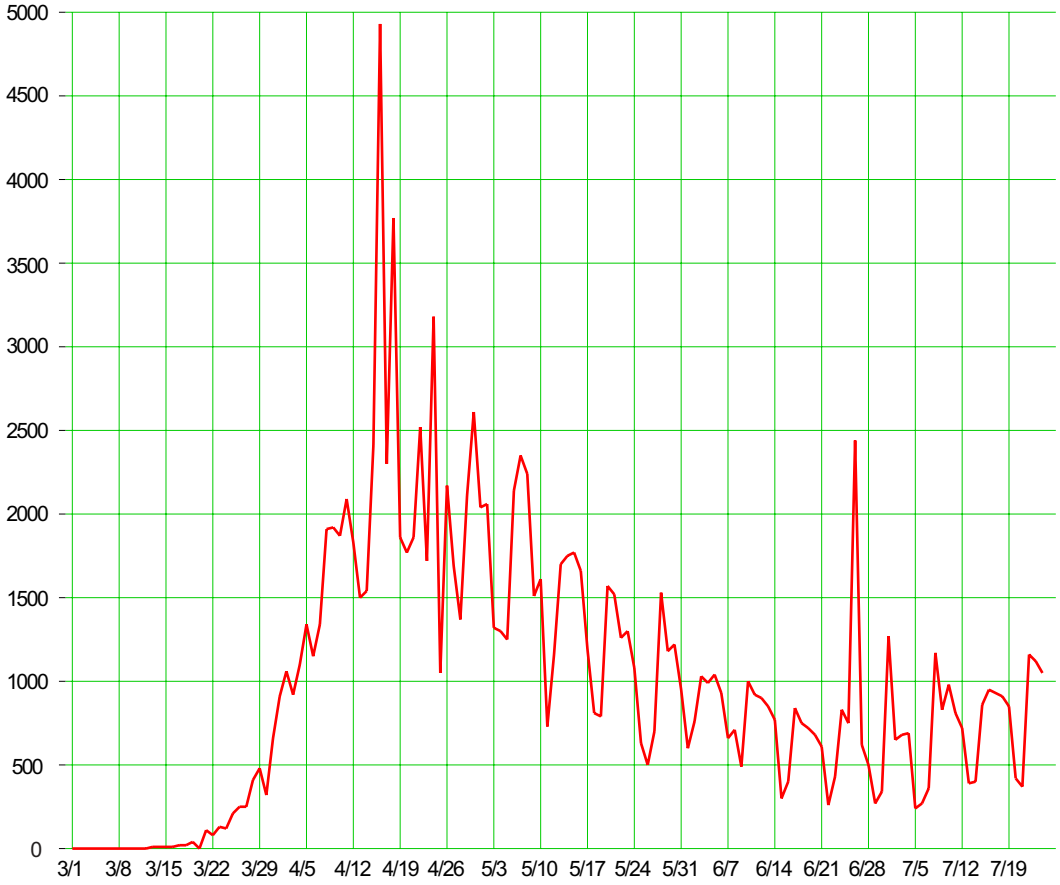

Figure 1: Daily Number of Covid-19 Deaths in the U.S.

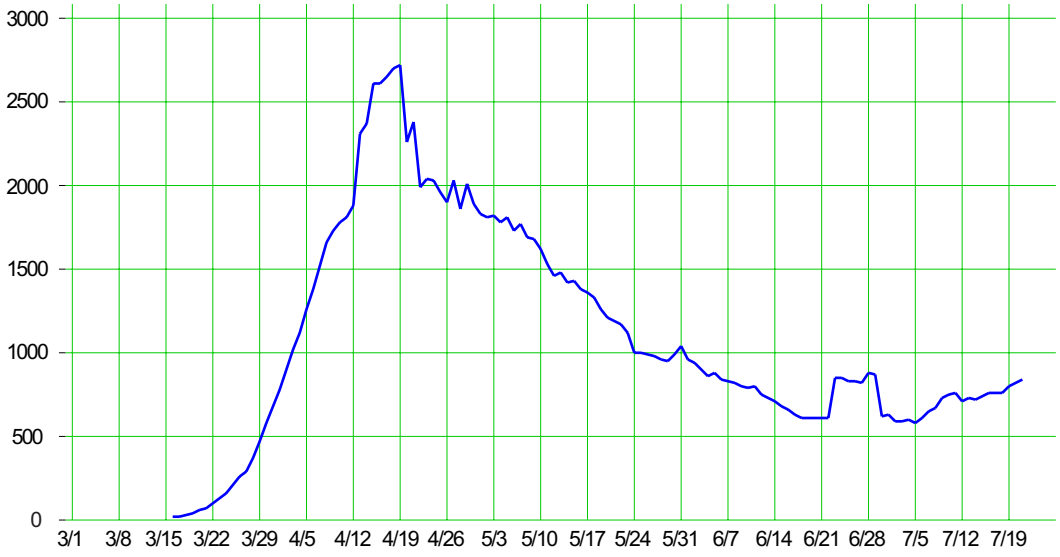

Figure 2: Seven-day Moving Average of the Daily Number of Covid-19 Deaths in the U.S.

The power of running records like Figures 1 and 2 is that they draw our eyes in the direction

our minds wants to go. At this point Figure 2 shows the current average number of deaths per day to be about 850, and this number has been trending up over the past two weeks. (The small plateau between 6/23 and 6/29 is due to the spike of 2437 deaths reported on 6/26.)

In those cases where the running records tell the story we will usually have no need for further analysis. With these two graphs we have pretty well discovered all there is to learn from these data. We can describe the past and project future possibilities.

To describe the past we might want to know the initial growth rate. By computing the ratio of successive seven-day averages for March and April, we can quantify the initial growth rate as averaging about 14% per day. To make predictions we might assume that the average number of deaths per day remains in the neighborhood of 800 per day and extrapolate. When we do this we predict about 80,000 more deaths by Halloween, which would result in a total of 229,000 deaths in the U.S. by that date. (This simple extrapolation is consistent with the projections coming out of the epidemiological models at this time.)

NONSENSE HAPPENS

Nevertheless, people all over the World have tried to use statistical tools to "analyze" the Covid data. After all, we can't possibly consume "raw" data. We have to process it first to make it unintelligible.

One of these statistical tools is the process behavior chart. Process behavior charts are incredibly useful and versatile. They allow us to filter out the noise so we can pay attention to the signals. They allow us to identify those points in time where a change occurs in a system that is supposed to be operated in a steady state. As one colleague of mine quipped, they ought to be called "has-a-change-occurred" charts.

But, as we have seen, an epidemic is anything but a steady-state system. It grows, changes, and evolves. We do not need to ask the question "Has a change occurred?" because we know that by its very nature every epidemic is constantly changing. This is why any attempt to use a process behavior chart to analyze the daily Covid values is a misapplication of the technique. It is conceptually equivalent to someone computing the average for a list of telephone numbers.

With the Covid-19 pandemic we are more interested in the longer-term trends as seen in Figure 2 rather than the daily values shown in Figure 1. The volatility of the daily values, along with the weekly cycles, make it virtually impossible to detect changes in the epidemic by directly analyzing the daily values.

However, this does not deter those who get so completely carried away with using their analysis techniques that they lose sight of what the data represent. Example after example can be found where analysts have used process behavior charts with the daily Covid numbers, or with values computed from those daily numbers. As the pandemic evolves, one or more points will inevitably begin to fall outside the limits, and the analysts can claim to have proven what anyone can see in Figures 1 and 2: the pandemic is changing. A particularly egregious example of this follows.

SEVEN STEPS TO NOWHERE

Some analysts have put out a You-tube video proposing a seven-step approach to analyzing Covid-19 data. These seven steps are as follow:

(1) Begin by *assuming* the daily counts are modeled by a Poisson distribution, then use a special type of process behavior chart known as a "c-chart" to identify when the epidemic starts to "grow."

(2) When the epidemic has grown to the extent that a point falls outside the limits of the c-chart, take all of the data to date and *transform* them by computing their natural logarithms.

(3) Obtain a *regression equation* using the logarithms of the daily counts as the dependent variables (the Y values) and the number of days since the start of the data set as the independent variables (the X values) in order to estimate the growth rate for the epidemic.

(4) Create a *sloping XmR* chart by placing limits around the regression equation:

(4a) To do this use the two point moving ranges obtained from the logarithms found in Step 2 to compute an average moving range.

(4b) Multiply this average moving range by 2.66 to get the three-sigma distance.

(4c) Add and subtract the three-sigma distance to and from the intercept term of the regression equation to get equations for the sloping limits of the *XmR* chart.

(5) *Exponentiate* the sloping limits using the Napierian constant $e = 2.7183$ and plot the resulting values as two exponentially increasing curves on a graph showing the original counts.

(6) Continue to *plot* new points on the graph found in step 5 until you get a point outside the exponentially increasing limits (on the right-hand side).

(7) On the date corresponding to this point outside the limits declare the epidemic to have "*peaked*."

To recap, before we are supposed to be able to interpret graphs like Figure 1, we need to perform seven steps: Assume, Transform, Regress, Create sloping *XmR*, Exponentiate the limits, Plot the points, and Find the peak.

At this point, in spite of all these computations, *no new insight has been created.* Everything "found" by this seven-step analysis (the initial growth, the growth rate, and the beginning of the flattening of the curve) is already made visible in the simple running records of Figures 1 and 2 or can be found using some simple computations. No value has been added by the seven step analysis. It simply uses a lot of computing power in order to decorate Figure 1 with what Edward Tufte calls "non-data ink", or more concisely, "chartjunk." And of course they will be glad to sell you some software to create all this chartjunk.

SUMMARY

Many schemes, ranging from simple to complex, using process behavior charts with Covid data have been tried. But regardless of their complexity, they all come up against the fact that epidemiological data do not represent a steady-state system where we need to discover if assignable causes are present. Process behavior charts simply ask the wrong questions here. When dealing with data from a dynamic system where the causes are well understood, the data will create a running record that can be interpreted at face value. The long-term changes will be sufficiently clear so that further data analysis becomes moot.

So, while specialists may use epidemiological models, when it comes to data analysis by non-specialists we do not need more analysis, but less. We need to draw the graphs that let the data speak for themselves, and then get out of the way. As always, the best analysis is the simplest analysis that provides the needed insight.