Received: 23 July 2020 | Revised: 28 April 2021 | Accepted: 21 May 2021

DOI: 10.1111/test.12273

ORIGINAL ARTICLE

WILEY

Covid and data science: Understanding R_0 could change your life

Jim Ridgway

School of Education, University of Durham, Durham, UK

Correspondence
Jim Ridgway, School of Education, University of Durham, Durham, DH1 1TA, UK.
Email: jim.ridgway@durham.ac.uk

Abstract

The Covid epidemic has provided an excellent example of the need to call on a wide variety of statistical tools to address a global problem, and can give students insights into some of the dimensions of data science. Here, we describe some of the characteristics of data that students encounter as citizens. We set out some teaching ideas, which focus on a few familiar core ideas—such as exponential growth, estimation, interpreting graphs, measurement, and sampling—set in the authentic context of containing a pandemic. In the final section, we sketch some more ideas on activities to develop student skills essential for civic engagement in a data-rich world.

KEYWORDS

teaching, Covid, data visualization, empowerment, estimation, exponential growth, measurement, teaching statistics

1 | INTRODUCTION

A tenant of epidemiology is that disease "outbreaks are inevitable, but epidemics are optional" [1], where 'optional' here means they can be prevented by appropriate action. A large number of people—3 million people at the time of writing (April 2021)—had died as a result of SARS-CoV-2 or Coronavirus Disease 2019 (hereafter, Covid). Every unavoidable death is a personal tragedy. The epidemic has had an impact on the daily lives and circumstances of billions of people and will continue to do so for the foreseeable future. Some governments imposed Draconian laws on their citizens and some did not. Citizens world-wide, if not impacted directly by illness or death, have been asked at the very least to change their behavior, often at considerable cost to themselves and their families. Within every country, citizens have had to make decisions about their own behavior, within these government frameworks across the world. A very wide variety of approaches has been taken to dealing with the pandemic—life-and-death decisions have been made on the basis of less than perfect information, uncertain models, and rapidly changing knowledge and events. Covid has posed a unique set of challenges; initially, everyone was ignorant about key aspects of the disease: how dangerous was it? how quickly could it spread? what could be done to reduce the damage caused? These are exactly the sorts of questions that provoked the invention of statistics. This is the heartland of data science.

Engel and Ridgway [3] identify some key features of information relevant to social issues that citizens encounter. Data have properties that should be encountered in all introductory statistics courses, including: choices of measures and decisions about the operationalization of concepts (eg, "Covid deaths" or "poverty") are contested; phenomena are multivariate; there are interactions between variables, and

non-linear relationships are typical, not exceptional; data are often aggregated; and indicator systems are common. Rich, new methods of representation and analysis continue to emerge. The variety of data types continues to increase; data are collected in innovative ways; new data visualizations and new analytic tools continue to be invented.

Data science is about using statistical methods, data, and technical tools to solve problems. It can involve creating and implementing more tools when we need them. Data science is characterized by thinking about ways to measure, the consequences of different decisions, making judgements about the quality of available evidence, and making predictions. Here, we set out some teaching ideas, which focus on a few familiar core ideas—such as exponential growth, estimation, interpreting graphs, measurement, and sampling—set in the authentic context of containing a pandemic. In the final section, we sketch some more ideas on activities to develop student skills essential for civic engagement in a data-rich world. Teachers will be sensitive to the needs of their own students. There are clear advantages in illustrating the usefulness of data science to society by using authentic data relevant to a pressing problem. However, for some groups (especially if individuals have been directly affected by Covid), it might be more appropriate to use data relevant to earlier pandemics such as ebola, influenza, or smallpox.

2 | CORE IDEAS

2.1 | Plausible estimation

The essence of plausible estimation is to derive an estimate that is good enough for a decision to be made when necessary, such as under time or information constraints. Interval estimates are much better than point estimates, and should be accompanied by a description of the assumptions that have been made. Covid planning has involved a great deal of plausible estimation. For example:

> How much personal protective equipment (PPE) will be needed by health workers in your country over the course of 1 month?

> If every citizen uses a disposable mask every day, what will be the weight and volume of the waste each week?

> Suppose you have an infinite supply of an effective vaccine. How long would it take to inoculate everyone in your country?

In each case, one needs definitions and "facts" as starting points. Who is to be counted as a "health worker," and how many people does this include? How many masks and pairs of gloves does one person need each day? How many people are available who are competent to give injections? Start with strong simplifying assumptions—for example, in the vaccine question, assume that everyone will be able to attend for vaccination; and that vaccination centers will function perfectly, 24 h every day. Then, add more realistic assumptions—bout personal mobility, geography, and the availability of competent staff. In 2001, Swan and Ridgway [14] created and gave detailed lesson plans to support teaching about (and assessment of) plausible estimation.

2.2 | Exponential growth

> Here is a deal for you—On New Year's Eve, I'm going to give you £1 billon! Yes, really! All I want back is £1 on first January, £2 on Jan second, £4 on Jan third, £8 on Jan fourth... until the end of the month. Waddaya say?

Time to reach for a spreadsheet... Students can be asked to create a table and graph to display deterministic exponential growth over days for different exponents. On screen, it is a simple matter to create a display where the user can change the exponent, and see the deterministic effects in both the table and the graphic.

Table 1 shows deterministic exponential growth for different exponents. The column headed £, where the payment doubles every day shows that, after 31 days of daily re-payments, the person offering the deal will be more than £1 billion ahead (the column total—which is £2 147 483 647—minus the original payment). So the response to the offer should be *I've got a better idea—let ME give YOU the billion...*

2.2.1 | On to Covid

The £ column corresponds to a reproduction rate (R_0) of exactly 2; that is, each infected person infects 2 more

TABLE 1 Exponential growth for different parameters

Day	Daily payment £	Daily cases $R_0 = 5$	$R_0 = 1$	$R_0 = 0.9$
1	1	10	10	10
2	2	50	10	9
3	4	250	10	8
4	8	1250	10	7
5	16	6250	10	7
6	32	31 250	10	6
7	64	156 250	10	5
8	128	781 250	10	5
9	256	3 906 250	10	4
10	512	19 531 250	10	4
11	1024	97 656 250	10	3
12	2048	488 281 250	10	3
13	4096	2,441 406 250	10	3
14	8192	12 207 031 250	10	3
15	16 384	61 035 156 250	10	2
16	32 768		10	2
17	65 536		10	2
18	131 072		10	2
19	262 144		10	2
20	524 288		10	1
21	1 048 576		10	1
22	2 097 152		10	1
23	4 194 304		10	1
24	8 388 608		10	1
25	16 777 216		10	1
26	33 554 432		10	1
27	67 108 864		10	1
28	134 217 728		10	1
29	268 435 456		10	1
30	536 870 912		10	0
31	1 073 741 824		10	0

people the next day. The next three columns in Table 1 are introducing ideas of disease spread. Simplistic deterministic assumptions have been made. Starting with 10 people, *if* there is perfect transmission, and every infected person infects exactly 5 others ($R_0 = 5$) by the next day, day 14 would see more than 12 billion newly infected people; that is, the world population (about 8 billion people) would catch the disease within 14 days. There are some big *Ifs* here that will be explored later. *If* each infected person infects exactly one other person ($R_0 = 1$), then the total number of infected people increases linearly by 10 each day; if the exact infection rate is less than 1 ($R_0 = 0.9$), on day 30, the number of new cases falls below 0.5 (0.47) and continues to rapidly fade away.

Apart from the absurdity of considering fractions of people, such a model for the spread of infection might appear too simplistic to be useful, but in the case of Covid it has its uses, with the above results a good approximation on average, with R_0 the average number of people infected by each infected person. A general model for epidemics must take into account many factors including: the virulence of the virus, the nature of the contacts between people in terms of spatial distribution and frequency, the nature of the contacts in terms of the circumstances of meeting (indoors or outdoors, wearing/not wearing masks etc.), and other chance factors including "virus load", individual immunity and the natural variation of infection spread by droplets or aerosol. Models of infection usually need to consider both the number of infected people and the number of those susceptible to infection, and the chance that a contact between an infected person and a non-infected person produces a new infected person. However, in the case of Covid, everyone was susceptible and the virus is easily transmitted, so the simplest model for an epidemic provides some answers. This model assumes an unlimited number of susceptible people, with R_0 the average number infected by each infected person. Although the same R_0 can arise from different assignation of probabilities over the number of possible people infected by each infected person, it is certain that the epidemic will die out if $R_0 < 1$, and grow indefinitely if $R_0 > 1$. How quickly either of these happens depends on both the value of R_0, and the distribution of probabilities. Simple models of this are very easy to explore and simulate by school students, as explained, for example, by Helen MacGillivray [7]. Hence, we see how important it has been to try to estimate R_0, and to reduce it by good hygiene, reducing contact, and finding and applying vaccines.

Table 1 assumes that the number of new infected people each day increased by a factor of exactly R_0. If, instead, we assume that R_0 is the mean number of new infected people per day from each infected person, and consider $R_0 = 0.9$, but that the number of new daily infected people from an infected person has a normal distribution with mean 0.9 and SD 0.3, we can examine the distribution of infected people after a certain period.

Figure 1 shows the results of 100 iterations of this process (starting with 100 cases) applied for 20 successive days. (Note that the values simulated from the normal have been rounded to whole numbers).

Of course, public health measures are designed to reduce R_0; so estimating R_0 is an important challenge for

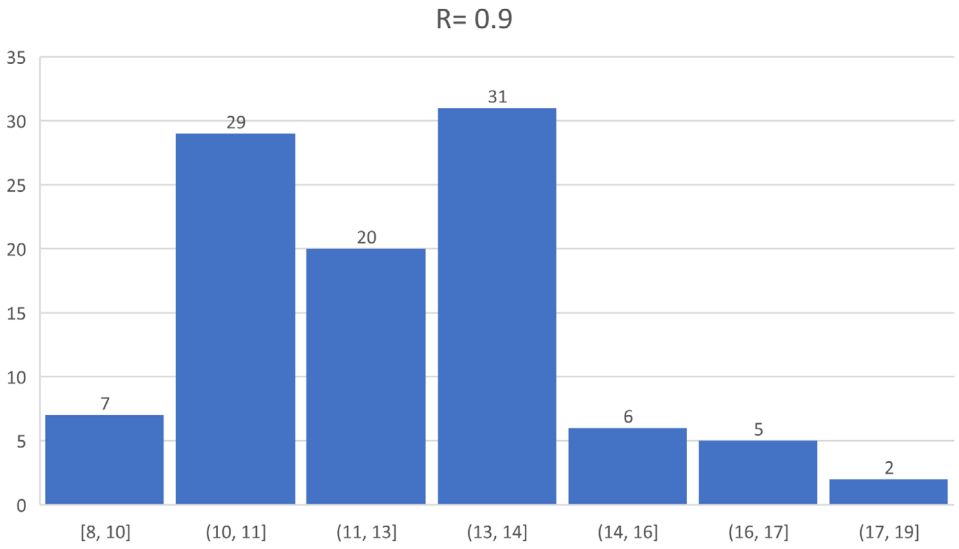

FIGURE 1 One simulation of the number of new daily Covid cases after 20 days when we assume the number of new daily cases arising from each infected person is N(0.9, 0.3) [Colour figure can be viewed at wileyonlinelibrary.com]

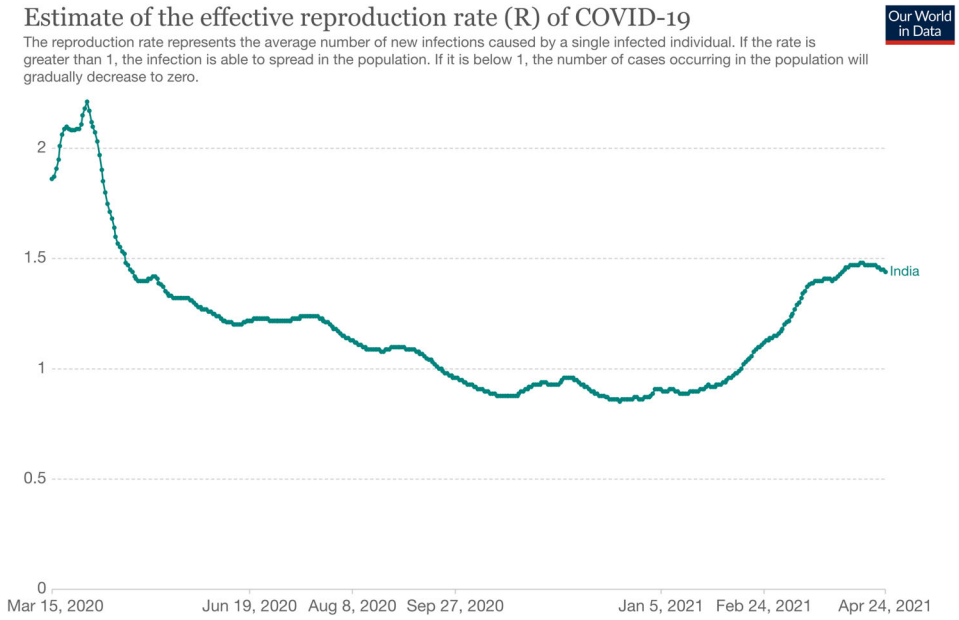

FIGURE 2 Estimates of R_0 for India over a 1-year period [Colour figure can be viewed at wileyonlinelibrary.com]

planning during a pandemic. Figure 2 shows estimates of R_0 for India, over the course of a year (sampling problems associated with all aspects of Covid are discussed later). It is clear that successes in curbing the virus in the autumn of 2020 were followed by a second wave of infection in 2021 (with disastrous human consequences).

2.3 | Interpreting graphs

Graphs are being used increasingly to convey complex information in the media. A picture may be worth 1000 words, but sometimes a graph needs 1000 words of explanation. Figures 3 and 4 present graphs downloaded from the *Our World in Data* website [9] about the spread of the disease in the United States and the United Kingdom. Each graph uses data from the same data set.

The *Our World in Data* grapher offers a number of choices about how data are to be displayed. One can choose

- Deaths or cases
- Daily deaths/cases or cumulative deaths/cases
- Raw numbers or numbers per million of population
- A linear or log scale

Students can be asked to describe to each other what they see in each graph—and if they believe, the graphs are based on the same data set. Then, tell students the graphs *do* show data from the same data set, and pose these challenges.

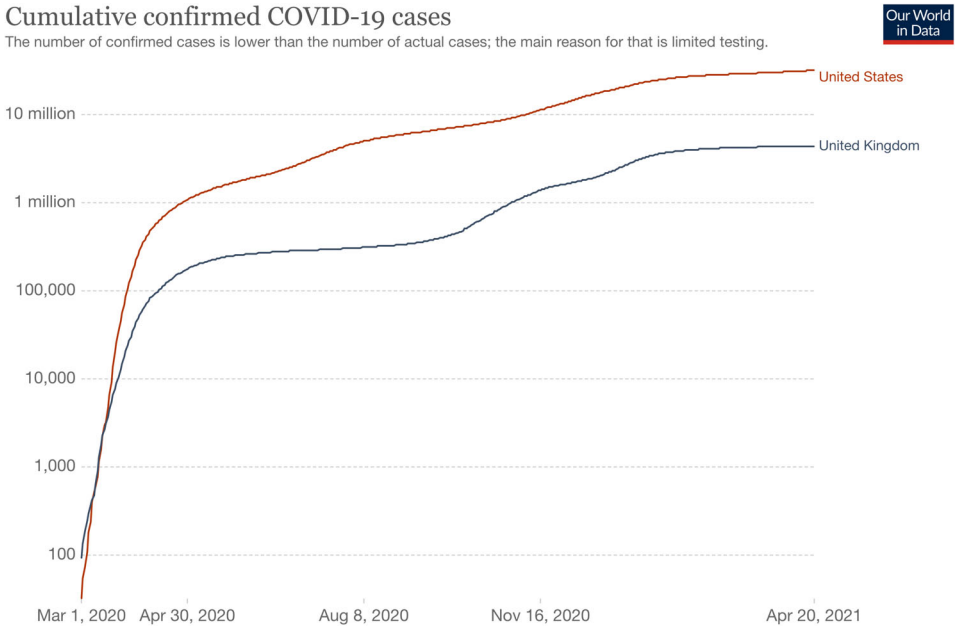

FIGURE 3 Cumulative confirmed Covid-19 cases in the United Kingdom and the United States [Colour figure can be viewed at wileyonlinelibrary.com]

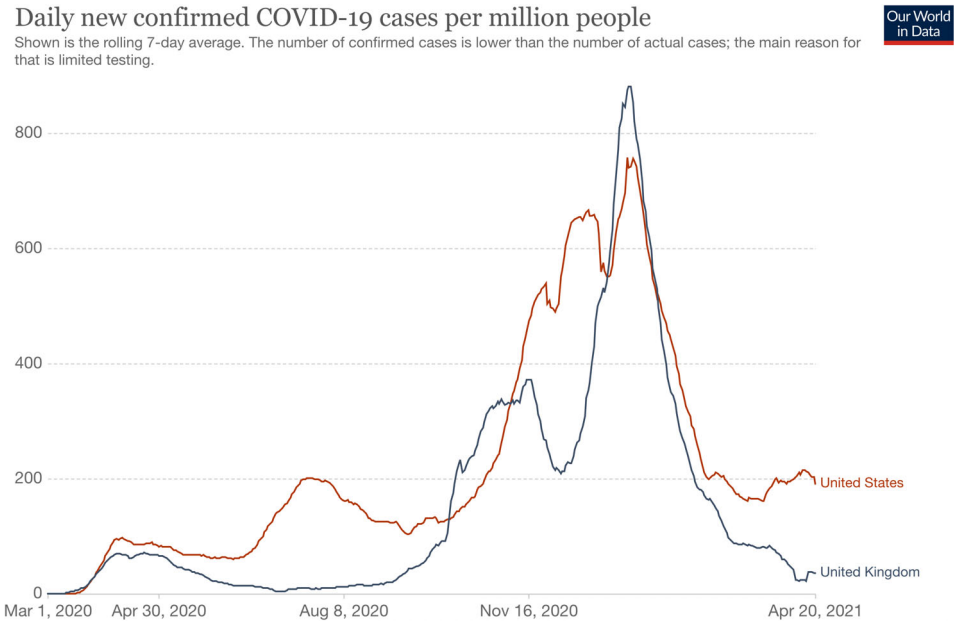

FIGURE 4 Daily new confirmed cases per million people in the United Kingdom and the United States [Colour figure can be viewed at wileyonlinelibrary.com]

The graphs look very different—explain why.
What type of scale is being used in each graph?
What conclusions can you draw from each graph?
In the context of Covid
When should you use a *linear scale*, and when should you use a *log scale*?
When should you use *raw numbers*, and when should you use *events per million of population*?
When should you use *deaths*, and when should you use *cases*?
When should you use *new cases*, and when should you use *cumulative cases*?

Linking Table 1 and Figure 3 highlights the most obvious advantage of log scales—the slope of the curve at any point gives a direct indication of the speed of spread

of the disease; changes in the slope show changes in disease acceleration (positive or negative). For instance, epidemics at an early stage can be compared with those at a later stage, even though there are big differences in actual deaths. However, log scales can be hard to understand, and can be visually misleading—for example, a change of one unit on the y-axis corresponds to both a rise from 100 to 1000 cases, and for a rise from 1000 to 100 000 cases. A linear scale gives a clearer impression of the size of the epidemic.

Raw numbers are essential for planning; scaling numbers by the size of the population is not useful at the start of an epidemic, but later gives some indication of the success of eradication programs in different countries, and the load on a country's resources. Estimates based on small samples are usually less stable than those based on larger samples; this is true of countries, too—those with small populations often have very high "cases per million" (eg, Andorra and Montenegro), and also have very low "cases per million" (eg, Mauritius, Fiji).

2.4 | Measurement issues

The discussion of cases vs deaths leads directly to questions about measurement. Clearly, identification of cases depends on the extent of testing; if testing is sparse, the count of *cases* will be too low. Even with comprehensive testing, a test may fail to detect Covid in the early stages of the disease. Further, some people get the virus and have mild symptoms or no symptoms at all. For example, in 2020, Pollan et al [11] conducted a survey of 61 000 randomly selected people in Spain, completed in May 2020, to determine the prevalence of Covid. About 5% of the sample tested positive; of these, about 1 in 3 were asymptomatic. So, the true number of cases is probably much higher than the reported number.

Measuring *deaths* is not without its problems. Is everyone who dies tested for Covid? In hospital settings, the death count is likely to be reliable (because it is important to know which patients had Covid); in community settings, such as care homes for the elderly, or in people's homes, data may be unreliable. Measures themselves can change; for example, moving from simply counting deaths in hospital to including deaths in prison, care homes, and the community produces major changes in the numbers reported. There are problems comparing numbers from different countries—different measures are used in different countries, and countries differ in the extent to which official statistics are independent of political dictat. The recoding system itself may be unreliable (as in some poorer countries). All of this can be used to draw students' attention to the critical importance in data science of understanding what is being measured and how, and to the importance of accessing and understanding metadata. It illustrates the reason that agencies concerned with cross-country comparisons (such as Organisation for Economic Co-operation and Development, Eurostat, and the United Nations) place such emphasis on the development of measures, reaching international agreement on methods of measurement, along with their insistence on linking metadata descriptions to datasets. Let us consider some of these measurement problems in more detail.

> For everyone who dies, suppose we know exactly who did, and did not, have Covid. Is this a good measure of deaths attributable to Covid?

Someone could have Covid and die from heart disease. This is problematic in care homes for the elderly, where the mortality rate is high, so Covid deaths may be exaggerated (conversely, one could argue that Covid made the heart attack far more likely). A bigger problem was discussed in a paper by Loke and Heneghan in 2020 [6] from the Centre for Evidence-Based Medicine in the United Kingdom. In England, there is a register of everyone who has ever been diagnosed with Covid-19. In July 2020, when someone died, if they were registered as having had Covid-19, they were recorded as a Covid death. So someone who recovered fully, but subsequently died in a traffic accident, was recorded as a Covid death (this recoding method has now changed). Covid deaths were recoded differently in Scotland, making it difficult to draw comparisons between the two countries.

> Use the (historical) English model for recording Covid deaths. Assume a constant rate of infection, and a constant true death rate. Sketch a graph of UK recorded Covid deaths over time.

Are there other ways to estimate Covid deaths?

Figure 5 shows weekly *total deaths* in the United Kingdom, together with weekly *total deaths* averaged over the previous 5 years, and *deaths* attributed to a number of causes. There are obvious peaks in April both in *total deaths* and *deaths attributed to Covid-19*.

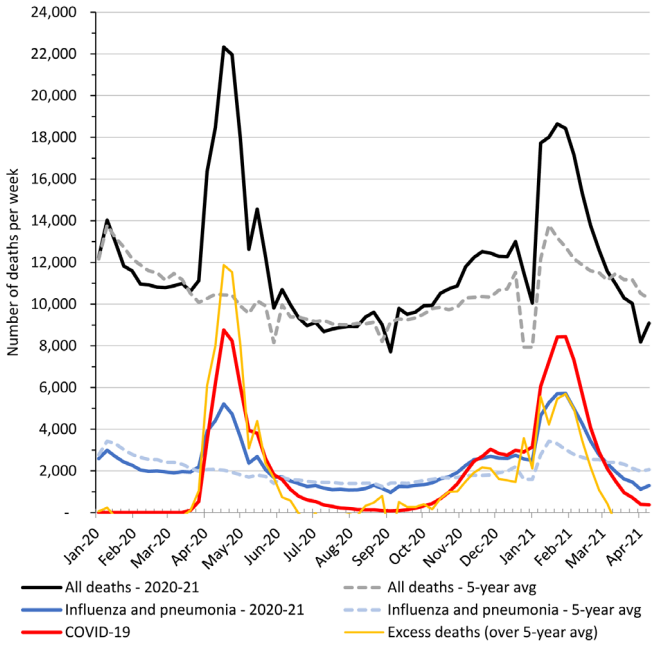

FIGURE 5 Weekly deaths: data from the Office for National Statistics CC BY-SA 4.0 [8] [Colour figure can be viewed at wileyonlinelibrary.com]

> We have data on total deaths each week over several years. So, it is easy to calculate *excess deaths*—the number of deaths that are higher than expected. Is this a good measure of deaths attributable to Covid?

Excess mortality data (available for different countries on the *Financial Times* website [4]) needs accurate historical data on deaths; these data are rare in middle-income and poor countries. Excess deaths might be underestimated if, for example, influenza deaths are lower in a particular year or if there are fewer deaths from other causes, such as road traffic accidents, or deaths attributable to air pollution, because people work from home.

Excess deaths might be overestimated if a pandemic results in increased deaths from other causes, such as resources being directed away from treating diseases such as cancers or HIV/AIDS, or if people die because they were unwilling to go to hospitals (eg, for emergency care) out of fear of contracting the disease.

So *confirmed deaths* associated with Covid (assuming these are not the result of a statistical anomaly!) and *excess deaths* are reflecting similar but not identical things. Covid deaths do reflect the cause of death, but probably underestimate the death toll (unless Public Health England were counting). Excess deaths are giving an overall impression of the effect of the pandemic, but can give an estimate of deaths directly attributable to the disease that might be overestimated (because of [say] more heart deaths associated with unwillingness to seek treatment) or underestimated (because of [say] fewer deaths associated with traffic accidents). Detailed lesson plans to support teaching about (and assessment of) inventing measures have been created by Swan and Ridgway [13].

2.5 | Sampling

Sampling is one of the Big Ideas in statistics. Covid demonstrates clearly that this Big Idea has not been grasped (or acted upon) by very many decision makers worldwide. For planning and action, we need to be able to estimate a number of parameters—how many people in the population are *susceptible*? *Infected*? How long will infected people (as a function of age, obesity, severity of infection, and other co-morbidities) stay in hospital? What is the *case fatality rate*? What proportion of people who have recovered are *immune*? (and for how long?). To determine these parameters, one needs to take sampling very seriously. Too much of the early work on estimation was based on opportunistic sampling. There are very big local variations in all these parameters, and parameters change over time, so careful testing needs to be an ongoing process.

3 | FURTHER ACTIVITIES

In this paper, we have confined the discussion to some fundamental data science ideas that students need to acquire. However, Engel and Ridgway [3] offer a longer list of skills necessary for informed citizenship. We conclude by recommending ways in which the Covid epidemic can be a focus for inculcating some more of these skills (see also [7] and [12]).

- Use a wide range of *data sources*—from hospitals, official statistics agencies, newspapers, university departments, web searches and tracking apps;
- Explore a variety of *data collection methods*—analyzing clinical reports, mass testing, surveys such as those described by Pew [10] on attitudes to Covid;
- Use a *wide range of analytic techniques*—from curve fitting to modelling—for example by exploring the effects of changing parameters in the SIR (Susceptibles, Infectives, Removed) approach to modelling epidemics using the model in a New York Times article by Kristof and Thompson [5];

- Use data sets that are *multivariate*—(incidence, deaths, recovery rates and geographies, etc) for example from worldometer [16] or WHO [15];
- Discuss the advantages and disadvantages of using *aggregated* and *disaggregated* data—use graphics from *Our World in Data* [9], or the *Financial Times* [4], which track cases and deaths in each US state. Does it make sense to add deaths in New York to deaths in Alaska?;
- Discuss and design *indicator systems*—for example using "expected quality of life years" to make decisions about which patients to treat if resources are limited;
- Explore *interactions* between variables—the likelihood of contracting Covid appears to be a function of (at least) poverty, age, obesity, and perhaps ethnicity;
- Point to the fact that *data may be time critical*—failing to act on early evidence about the spread of Covid lead to the preventable deaths of tens of thousands of people; and that *parameters change over time*, as a function of human behavior (eg, the chance of contracting Covid reduced when social distancing was observed);
- Teach beyond "graph reading" to "deconstructing and interpreting novel visual displays" by working with *innovative visualizations*;
- Teach "criticality"—that is, the ability to map out the logical structure of claims being made, and to evaluate both the logic of the argument, and the strength of the evidence on which it is based; statistical arguments are often *embedded in rich text*—so ask students to deconstruct and reconstruct stories about data in a variety of media;
- Discuss *causality*—this is a difficult arena that we should explore with students—the study of disease is an excellent context. For example, in 2015, DeBold and Friedman [2] writing in the *Wall Street Journal* present displays of the number of infected people in every US state over a 70 year time period for different diseases, along with a line corresponding to the time when vaccination was introduced. In some cases, the evidence to support a causal claim that vaccination had a dramatic effect seems overwhelming; for some other diseases, less so;
- Teach about *risk*—a central issue—we need to go beyond probability and embrace the *consequences of different outcomes*—for example, the UK Prime Minister boasted on March 3, 2020 about shaking hands "with everybody" at a hospital with confirmed Covid patients (the same day that the government advisory group warned specifically against this). Four weeks later, he was a Covid patient in an intensive care ward.

Covid offers a context for presenting data science as an overarching structure for empowering students (and citizens, and decision makers). Data science and statistics are inextricably bound together—fundamental statistical ideas can be exemplified vividly by devoting time to real-world problems, and grand principles can be illuminated by working with authentic data presented in exciting and engaging ways. Pandemics are optional; if we choose not to act, pandemics pose an existential threat to individuals and societies; students can see that, indeed, understanding R_0 is a matter of life and death.

ORCID

Jim Ridgway https://orcid.org/0000-0002-0826-4815

REFERENCES

1. L. Brilliant, Outbreaks are inevitable, but pandemics are optional, 2020, available at https://www.youtube.com/watch?v=nVWoHmURDTQ.
2. T. DeBold and D. Friedman, *Battling Infectious Diseases in the 20th Century: The Impact of Vaccines*, Wall Street J. (2015). http://graphics.wsj.com/infectious-diseases-and-vaccines/.
3. J. Engel and J. Ridgway, *Back to the future-rethinking the purpose and nature of statistics education*, in *Statistics for Empowerment and Social Engagement: teaching civic statistics to develop informed citizens*, J. Ridgway, Ed., Springer, Berlin, 2021 (in press).
4. Financial Times, *Coronavirus tracked: the latest figures as countries start to reopen*, 2020, available at https://www.ft.com/content/a26fbf7e-48f8-11ea-aeb3-955839e06441.
5. N. Kristof and S. Thompson, *Trump wants to "Reopen America." Here's What Happens if We Do*, New York Times, 2020, available at https://www.nytimes.com/interactive/2020/03/25/opinion/coronavirus-trump-reopen-america.html.
6. Y. Loke and C. Heneghan, *Why no-one can ever recover from COVID-19 in England-a statistical anomaly*, 2020, available at https://www.cebm.net/covid-19/why-no-one-can-ever-recover-from-covid-19-in-england-a-statistical-anomaly/.
7. H. MacGillivray, *Charting statistical courses*, Teach. Stat. **42**(2) (2018), 33–35.
8. Office for National Statistics, 2020, available at https://commons.wikimedia.org/wiki/File:ONS_weekly_COVID-19_deaths_E%26W.svg#/media/File:ONS_weekly_COVID-19_deaths_E&W.svg.
9. *Our World in Data Coronavirus pandemic (COVID_19)*, available at https://ourworldindata.org/coronavirus.
10. Pew Research Center, *Coronavirus Disease (COVID-19)*, 2020, available at https://www.pewresearch.org/topics/coronavirus-disease-2019-covid-19/
11. M. Pollan et al., *Prevalence of SARS-CoV-2 in Spain (ENE-COVID): a nationwide, population-based seroepidemiological study*, Lancet **396** (2020), 535–544.
12. J. Ridgway and R. Ridgway, *Civic statistics in the time of Covid*, in *Statistics for empowerment and social engagement: teaching civic statistics to develop informed citizens*, J. Ridgway, Ed., Springer, Berlin, 2021.
13. M. Swan and J. Ridgway, *Classroom assessment techniques: "creating measures" tasks*, 2001, available at http://archive.wceruw.org/cl1/flag/cat/math/measures/measures1.htm.

14. M. Swan and J. Ridgway, *Classroom Assessment Techniques: "plausible estimation" tasks*, 2001, available at http://archive.wceruw.org/cl1/flag/cat/math/measures/measures1.htm.
15. WHO Coronavirus Disease, *(COVID-19) Dashboard*. https://covid19.who.int/info/.
16. Worldometer, *Coronavirus (COVID-19) Mortality Rate*, 2020, available at https://www.worldometers.info/coronavirus/coronavirus-death-rate/#who-03-03-20.

How to cite this article: J. Ridgway, *Covid and data science: Understanding* R_0 *could change your life*, Teaching Statistics **43** (2021), S84–S92. https://doi.org/10.1111/test.12273