Role of statistics in the era of data science

Rajeeva L. Karandikar

Statistics evolved as a science in an era when the amount of data available was small and efforts were on to extract maximum information from them. Are the techniques developed during those times relevant anymore in the era of data science? We will illustrate using examples that several statistical concepts developed over the last 150 years are as relevant in this era as they were then.

Keywords: Analytics, big data, bias, data-science, regression, statistics.

FOR over a decade now, big data, analytics and data-science have been the buzzwords. As is the trend now, we will just refer to any combination of these three as data science.

A large number of professionals working in the Information Technology (IT) sector have moved to positions in data science and picked up new tools which have mushroomed. Often, these tools are used as a black-box. This is not surprising because most of these people have little background in statistics. We can often hear them make comments such as, 'With a large amount of data available, who needs statistics and statisticians? We can process the data with various available tools and pick the tool that best serves our purpose.'

It can be seen that in a large proportion of the problems that need some decision based on available data, the standard tools in artificial intelligence or machine learning (AIML) and statistics will yield the best or nearly the best answer. Of course, the analyst will still need to use the correct tool.

We hear many stories of wonderful outcomes from what can be termed a pure data-driven approach. This has led to a tendency to just take a large chunk of available data and push it through an AIML engine to derive intelligence without giving a thought to where the data came from, how they were collected and what connection they have with the questions whose answer one is seeking ... . If an analyst were to ask questions about the data like – how and when were they collected, etc. the answer one frequently hears is: 'How does it matter?' Later in the article, we will see that it does matter.

We will also see that there are situations where blind use of the tools with the data may lead to a poor conclusion. While we can construct many such (hypothetical) examples where pure data-driven analysis can lead to incorrect or wrong conclusions, it would be much more convincing to pick real examples from history, and we will later present a few such examples.

The author is in the Chennai Mathematical Institute, Chennai 603 103, India.
e-mail: rlk@cmi.ac.in

As more and more data become available in various contexts, our ability to draw meaningful actionable intelligence will grow enormously. The best way forward is to marry statistical insights to ideas in AIML, and then use the vast computing power available at one's fingertips. For this, statisticians and AIML experts have to work together along with the domain experts.

We will illustrate through examples how ignoring the statistical ideas and thoughts that have evolved over the last 150 years can lead to incorrect conclusions in many situations.

Statisticians are often asked: 'is statistics relevant in the era of big data?'. My take is that statistical ideas are important and will continue to play a big role. For this the way we teach statistics must also change. The focus should be more on concrete applications. Each technique, starting with the estimation or testing of hypotheses, should be taught along with concrete applications from various domains.

Small data is still relevant

First let us note that there is a class of problems where all the statistical theory and methodology developed over the last 150 years continues to have a role – since the data are only in hundreds or at most thousands and never in millions. For example, issues related to quality control, quality measurement, quality assurance, etc. only require a few hundred data points to draw valid conclusions. Finance – where the term VaR (Value at Risk), which is essentially a statistical term – 95th or 99th percentile of the potential loss, has entered the law books of several countries – is another area where use of data has become increasingly common; and here too we work with a relatively small number of data points. There are roughly 250 trading days in a year and there is no point going beyond 3 or 5 years in the past as economic ground realities are constantly changing. Thus we may have only about 1250 data points of daily closing prices to use for say portfolio optimization or option pricing, or for risk management. One can use hourly prices (with 10,000 data points) or even tick-by-tick trading data,

but for portfolio optimization and risk management, the common practice is to utilize daily prices. In election forecasting, psephologists usually work with just a few thousand data points from an opinion poll to predict election outcomes. Finally, policy makers, who keep tabs on various socio-economic parameters in a nation, rely on survey data which, of course, are not in millions.

One of the biggest problems faced by humanity in recent times is the COVID-19 pandemic. From March 2020 till the year end, everyone was waiting for vaccines against COVID-19. Finally in December 2020, the first vaccine was approved and more have followed. Let us recall that the approval of vaccines is based on randomized clinical trials (RCT) which involve a few thousand observations, along with concepts developed in the statistical literature under the theme 'design of experiments'. Indeed, most drugs and vaccines are identified, tested and approved using these techniques.

These examples illustrate that there are several problems where we need to arrive at a decision or reach a conclusion under uncertainty and we do not have millions of data points. We have to do our best with a few hundred or few thousand data points. The problems include identification and approval of medicines and vaccines – which are based on clinical trials. The urgency of getting the medicine or vaccine to the people means we need to work with as little data as possible. So techniques of working with small data (involving a few thousand observations) will always remain relevant.

In this article we will illustrate that even in problems where we may have large amounts of data available, several statistical ideas have a role.

Perils of purely data-driven inference

The first example we discuss goes back nearly 150 years. Francis Galton was a cousin of Charles Darwin, and as a follow-up to Darwin's ideas of evolution, Galton was studying inheritance of genetic traits from one generation to the next. He had his focus on how intelligence is passed from one generation to the next. To understand the context and his views on the theme, see Galton[1].

Studying inheritance, Galton wrote 'It is some years since I made an extensive series of experiments on the produce of seeds of different size but of the same species. They yielded results that seemed very noteworthy, and I used them as the basis of a lecture before the Royal Institution on 9 February 1877. It appeared from these experiments that 'the offspring did not tend to resemble their parent seeds in size, but always to be more mediocre than they – to be smaller than the parents, if the parents were large; to be larger than the parents, if the parents were very small.'[2] (also see Galton[3]). Galton firmly believed that this phenomenon will be true for humans as well and for all traits that are passed on genetically, including intelligence.

To illustrate his point, Galton obtained data on heights of parents and their (grown-up) offspring. He chose height as it was easy to obtain data on it. Analysis (given in Galton[4]) of the data confirmed his hypothesis, quoted above in italics. He further argued that this phenomenon would continue over generations, and its effect would indicate that heights of future offspring will continue to move towards the average height. He argued that the same will happen to intelligence and thus everyone will only have average intelligence. This is why he chose the title of the paper as 'Regression towards mediocrity in hereditary stature'[3].

The conclusion drawn by Galton is fallacious as can be seen by analysing the same data by interchanging roles of height of offspring and mid-height of parents leading to an exactly opposite conclusion – namely that if the offspring is taller than average then the average height of parents will be less than that of the offspring, while if the offspring is shorter than average, then the average height of parents will be more than the child. This appears to contradict the conclusion drawn by Galton.

The variation in heights (variance of heights) in the two generations was comparable, whereas if there was regression towards the mean, variance would have decreased. Thus, Galton's conclusion about regression to mediocrity over generations is not correct. However, the methodology that he developed for the analysis of inheritance of heights has become a standard tool in statistics and continues to be called regression (this explains why a term with an apparently negative connotation is being used for one of the most popular statistical techniques).

Galton, who made many contributions to the development of theory and applications of statistics in his time apart from regression, was so convinced of his theory that he just looked at the data from one angle and got confirmation of his belief. This phenomenon is called confirmation bias – a term coined by English psychologist Peter Wason in the 1960s.

This belief, of regression to mediocrity, was so entrenched in Galton's thinking, that it became the central theme in a subject he named eugenics, and also became a big movement. Karl Pearson was an avid follower of Galton, and was the first eugenics professor at University College London. Pearson introduced the χ^2-test for testing goodness of fit, concept of P-value, the method of moments for estimation of parameters and principal component analysis (PCA) for analysis of multivariate data – a technique extremely popular today. He also laid the foundation on which R. A. Fisher later proposed the theory of testing of hypothesis. Yet Pearson went along with Galton's broad conclusions, though he did not stress upon it. Galton and Pearson's writings on eugenics are today considered out and out racist.

Some people may think that if only Galton had a very large dataset, he may not have drawn a wrong conclusion. Suppose that a pair of random variables (X, Y) has a

bivariate normal distribution, with mean and standard deviation of each of the components being 173 and 8 cm respectively, and correlation between the two being 0.8. Simulating a million data points (x_j, y_j) from this distribution, we can observe the same effect (taking X as the mean height of the father and Y as the height of the son): if we predict the height of the son (Y), based on height of the father (X), we will see the phenomenon alluded to by Galton, while if we use the same simulated data to predict height of the father (X), based on the height of the son (Y), we will see the other effect – which seems to contradict the first. This phenomenon of two regression lines is well known and is generally covered in the first course on regression.

Are data representative of the population?

Given data, even if they are, one must first ask how they were collected. Only after knowing this can one begin to determine if the data are representative of the population one wishes to draw conclusions about.

In India, if one website of a news channel (A) runs a poll on its website on a political question, say a policy announced by the Central Government recently and at the same time, another news channel (B) also runs a poll at its website on the same question. Even if both sites attract a large number of responses, it is very likely that the conclusions will be diametrically opposite – the poll on (A) could show that the vast majority supports the government policy while the poll on (B) may be diametrically opposite. It is easy to explain as neither website represents the Indian population. In view of their known positions on political questions, more pro-Government citizens may visit and respond on one poll while the opposite may be true for other channel.

This underscores the point that just having a large set of data is not enough – it must represent the population in question for the inference to be valid.

Thus, if someone gives a large chunk of data on voter preferences to an analyst and wants her to analyse, the same and predict the next elections, she must start by asking as to how the data were collected, only then can she decide if they represent the Indian electorate or not. For example, suppose an agency proposes to the analyst, a week before an upcoming state election, that she would be given data from the social media on posts and messages regarding political questions during, say the previous few weeks and on this basis she is expected to predict the election outcome[5]. The analyst should not take up the project, as in this case, it is clear that the data would not be representative of the electorate of the state. Indeed, the uneducated, rural, economically weaker sections are highly under-represented on social media and thus the conclusions drawn based on the opinion of such a group (of social media users) will not be able to give insight into how the Indian electorate will vote or has voted. The social media data can be used to quickly assess market potential of a high-end item, such as an iPhone or Pixel smartphone – for their target market is precisely those who are active on social media.

Likewise, any data on likes and dislikes or preferences from a survey conducted in a metro city in India will not give an insight into the preferences from across the country. This aspect has to be foremost in an analyst's mind when given a dataset for analysis.

Perils of blind use of tools without proper understanding

The next example is not one incident but a theme that is recurrent – that of trying to evaluate the efficacy of an entrance test for admissions, such as IIT-JEE for admission to the IITs, CAT for admission to the IIMs or SAT/GRE for admission to top universities in the USA. Let us consider these as benchmark tests which are open to all candidates, and those who perform well in these tests are shortlisted for further screening (an interview or another test may follow), or are admitted to the targeted programme. The analysis consists of computing correlation between the score on the benchmark test and the performance of the candidate in the programme. Often it is found that the correlation is rather poor and this leads to a discussion on the quality of the benchmark test, with questions being raised if the test is even necessary. What is forgotten or ignored is that the performance data are available only for the candidates selected for admission. This phenomenon is known as selection bias – where the dataset consists of only a subset of the whole group under consideration, selected based on some criterion.

Coming back to the benchmark tests, the observed low correlation between score in the benchmark test and performance among the selected candidates does not indicate that the correlation would necessarily be low in the population, if the data were available. The following simulation exercise illustrates that this is possible even if the true correlation in the population is high. Suppose for a population of say 100,000 students, X denotes the score in a selection test and Y denotes the score at the end of the programme (had the student been selected), and suppose the two are highly correlated. Assuming that only 500 (top 0.5%) students are selected, Table 1 gives the mean, standard deviation and quantiles based on 100,000 simulations of the correlation ρ^* of the benchmark

Table 1. Mean, standard deviation and quantiles of tail correlation ρ^*

ρ	Mean	SD	5%	25%	50%	75%	95%
0.6	0.21	0.01	0.20	0.21	0.21	0.22	0.23
0.65	0.24	0.01	0.23	0.24	0.24	0.25	0.26
0.7	0.28	0.01	0.26	0.27	0.28	0.28	0.29
0.75	0.32	0.01	0.30	0.31	0.32	0.32	0.33
0.8	0.36	0.01	0.35	0.36	0.36	0.37	0.38

test score X and score during the programme Y of the selected candidates, assuming that (X, Y) are bivariate normal with correlation ρ. As we can see, the high values of ρ need not translate to high ρ^*.

This illustrates the phenomenon known as absence of tail dependence for joint normal distribution. This property is inherited by models built using Gaussian copulas and use of these models for risk management has been considered as an important reason for the collapse of global financial markets in 2008. A search for the formula that killed Wall Street or the formula that killed the financial market throws up over a million hits on Google – as if it is a single formula that was responsible for the market collapse. In the words of Paul Embrechts – an eminent statistician and an expert on risk management, ... 'this is akin to blaming Einstein's $E = mc^2$ formula for the destruction wreaked by the atomic bomb'[6–8]. The indiscriminate use of mathematical and statistical formulae without understanding – coupled with human greed were responsible for the collapse of global financial markets and not the formula.

Similar bias occurs in studies related to health, where for reasons beyond the control of the team undertaking the study, some patients are no longer available for observation. The bias it introduces is called censoring bias and how to account for it in the analysis is a major theme in an area known as survival analysis in statistics.

Here is another example where blind use of a tool led to the wrong conclusion. In the game of cricket, all historical data on one day international (ODI) cricket matches are available in electronic form and various attempts have been made to use the data to predict outcomes of upcoming matches. In 2007, during the ODI Cricket World Cup, a techie team at one of the analytics companies in Mumbai predicted outcomes of the matches as well as (exact) scores of several batsmen based on a model with these data. In the opening match on 13 March 2007, between hosts West Indies and Pakistan, the data-driven model developed by the analytics company predicted victory for West Indies and also the exact scores of the two captains: Inzamam-ul-Haq – 36 and Brian Lara – 37. And as it turned out, their predictions were spot on. The two captains got out at exactly the predicted scores and West Indies won the match. The company got a lot of attention in the media the next day with the team leader talking about their efforts, expertise, model, etc. However, the model was way off the mark in subsequent matches. It can be seen that nothing is wrong with the models, but the best that any statistical model can do in predicting a stochastic variable is to estimate its conditional distribution given all the observed information. In the case of predicting the score of Lara in an upcoming match, the model can take into account all the information and estimate the conditional probabilities associated with various possible outcomes. Some models may show as output the outcome with highest probability (mode of this conditional distribution), median or mean of this distribution. For example, multiple linear regression (with squared error loss) yields the conditional mean, while the least absolute deviation regression would yield the median of the conditional distribution. But, of course, these predictions come with an estimate of the error, thus acknowledging the fact that the model is not forecasting the exact score of the batsman on the next outing.

Correlation does not imply causation

Most data-driven analysis can be summarized as trying to discover relationships among different variables – and this is what correlation and regression are all about. These were introduced by Galton about 150 years ago and have been a source of intense debate. One of the points that has been stressed over and over again is that correlation does not imply causation. While correlation (and in turn regression) are techniques to discover linear relationships, one needs to use transformations to get to more complex relationships. Likewise, AIML techniques try to find relations, linear or otherwise, among the variables.

It is tempting to interpret the observed relationship between various variables in the data as causation. The statistics literature has many examples where such relationships occur. See the presidential address by Yule[9] to the Royal Statistical Society in 1925 on nonsense correlation.

Often the relationship between variables, say X and Y can be explained by a third variable, say Z that influences both X and Y. One example often cited is where X is the monthly expenditure on sale of ice-cream in a coastal town in Europe and Y is the number of deaths due to drowning (while swimming in the sea, in that town) in a given month. One sees a strong correlation. While there is no reason as to why eating more ice-cream would lead to more deaths due to drowning, one can see that they are strongly correlated to a variable Z – average of the daily maximum temperature during the month – in summer months more people eat ice-cream and more people go for a swim. In such instances, the variable Z is called a confounding variable. Another instance of such observed spurious correlation was talked about in the Indian media several times during the later part of Sachin Tendulkar's career, with X being the total runs scored by him in international cricket matches and Y being the SENSEX – the index of stock prices on the Bombay Stock Exchange. The confounding variable here is Z – the time. Tendulkar arrived on the scene about the same time as the economic liberalization in India and both his runs and the SENSEX rapidly increased, yielding correlation. Perhaps, even die-hard fans of Tendulkar would not mention that there was a causal relationship.

In today's world, countrywide data would be available for a large number of socio-economic variables, variables related to health, nutrition, hygiene, pollution, economic

variables, and so on – one can list about 500 variables where data on over 100 countries are available. One is likely to observe correlations among several pairs of these 500 variables – one such recent observation is that gross domestic product (GDP) of a country and the number of deaths per million population due to COVID-19 are strongly correlated. Likewise, the rating of a country on human development index (HDI) is strongly correlated with the number of deaths per million population due to COVID-19. Of course, there is no reason why richer or more developed countries should have more deaths. Indeed, this can be explained by various other factors[10]. See Hassler and Thadewald[11] for another example where we can see that spurious correlation shows up due to pooling of heterogeneous samples. It also lists several other articles on this theme.

For more examples, one can do a search on the web with the phrase 'spurious correlation' or 'nonsense correlation' and obtain a large number of examples. There are several instances where one can observe correlation just as coincidence (without a confounding variable lurking in the background).

So the point is that just as linear relationships may be spurious, the relationships discovered by AIML algorithms may also be so. Thus, this learning from the statistical literature going back a century is relevant – rather more relevant today simply because we have a lot more data and processing power.

Simpson's paradox and the omitted variable bias

Simpson's paradox, also known as amalgamation paradox, reversal paradox or Yule–Simpson effect is an effect where ignoring an important variable may reverse the conclusion[12–15]. For example, it is possible that while the CFR – case fatality ratio (number of deaths divided by the total reported cases) due to COVID-19 for a country *A* is more than a country *B*, the CFR for each age group for country *A* may be less than that for country *B*. The apparent paradox may occur because of the differences in age profiles of the two countries along with higher mortality rates among the older population. While we have not seen the data, we have come across references that the statement above is true with *A* being Italy and *B* being China.

One of the examples of Simpson's paradox is a study of gender bias among graduate school admissions to the University of California, Berkeley, USA. In 1973, it was alleged that there is a gender bias in graduate school admissions – the acceptance ratio among males was 44%, while among females it was 35%. When the statisticians at Berkeley wanted to identify which department is responsible for this, they looked at Department-wise acceptance ratios and found that if anything, there was a bias against the males. See Bickel *et al.*[16] as also numerous articles on the web on the topic. The apparent bias in the pooled data appeared because a lot more women applied to departments which had lower acceptance rates. See the Wikipedia article on Simpson's paradox for more examples from the literature. The variable department in this example is called a confounding factor. In the economics literature, the same phenomenon is also called omitted variable bias.

Ignorance of this well-known phenomenon can cause statistically invalid conclusions being taken seriously. In May 2020, *Lancet* published an article about the use of hydroxychloroquine (HCQ) in treating COVID-19 infection. The paper claimed to analyse real-world evidence and seemed to show that HCQ is not useful in treating COVID-19 infection, and may actually be worsening the situation. The authors had taken data from a large number of hospitals across continents. The very next day, World Health Organization stopped clinical trials involving HCQ. There were many issues with the statistical analysis in the *Lancet* paper. The authors had not factored in as to how patients receiving HCQ treatment were chosen. It is known that fatality rates for patients with symptoms are higher than the asymptomatic patients infected with COVID-19. Thus, if patients with severe symptoms were mostly given HCQ and others were not, then the conclusions cited in the paper are invalid. The *Lancet* paper had many other issues too – the authenticity of the data could not be verified – and the paper was eventually withdrawn. But if the referees or editors had been aware of Simpson's paradox, the paper may not even have been published in the first place. To add to the puzzle, some studies[17,18] have also shown that HCQ has been effective in treating COVID-19. Perhaps, better designed tests on HCQ would have conclusively settled the issue and prevented this farcical situation.

Are statistical models still relevant in the big data era?

Let us now move to the 1940s – the World War II (WW II) years. Both sides in the War had realized the importance of air warfare though the aircraft used in WWII were primitive (by today's standard) and so were the guns used to fire at them. Some aircraft of the allied forces were being jumped down by the German forces while many were returning to the base, some without any hit and several coming back in spite of being hit. Based on a naive analysis of the data on the locations where the planes were taking a hit, it was proposed to the British Airforce that armour be added to those areas that showed the most damage.

Abraham Wald was a Professor at Columbia University, USA and a member of the Statistical Research Group (SRG) which had been consulting with the US defence establishment during the war. He came across this recommendation and it was questioned as to how much armour should be added to the vulnerable parts.

Wald looked at the problem from a different angle. He realized that there was a selection bias in the data that were presented to him – only the aircraft that did not crash returned to the base and made it to the data. He assumed that the probability of being hit in any given part of the plane was proportional to its area (since the shooters could not aim at any specific part of the plane). Also, given that there was no redundancy in aircraft at that time (like having two engines while only one working engine is required to bring it back to the ground, as has been the case for decades), the effect of hits on a given area of the aircraft was independent of the effect of hits in any other area. Once he considered these two assumptions, the conclusion was obvious – that armour be added in parts where fewer hits had been observed (among the aircraft that returned)[19,20]. So the statistical thinking led Wald to the model that gave the right frame of reference which connected the data (hits on planes that returned) and the desired conclusion (where to add the armour). If one is on a wrong frame of reference, one may end up with a wrong conclusion.

The phrase garbage in garbage out (GIGO) is often used to describe the fact that even with the best of algorithms, if the input data (to the algorithm) are garbage, then the conclusion (output) is also likely to be garbage. The discussion above can be summed up as, it is equally import to remember that garbage model may also lead to garbage output (GMGO) even with accurate data.

Conclusion

We have seen several examples where just treating data as sacrosanct and ignoring all other aspects of the domain from where the data have been taken can lead to wrong conclusions. Indeed, along with the data, giving a thought to the domain, understanding the answers that are of interest, etc. (using statistical ideas or fresh thinking) can be far better.

A good example of this is the internet search engine during the 1990s, several internet search engines appeared; some like Yahoo were extremely popular and based on the strength of its search engine, Yahoo became a major multinational. Then came Google. Using the PageRank algorithm developed by its founders, Google has captured over 90% market share, pushing everyone out of this space and has become one of the major drivers of the internet.

We suggest that in order to enable students of statistics to take up positions involving data science, they should be exposed to various real-life examples where statistics is playing a role in data science. We need to reduce focus on theorem–proof approaches to teaching and bring in simulation-based assessment of various techniques.

1. Galton, F., *Hereditary Genius: An Inquiry into its Laws and Consequences*, Macmillan, London, UK, 1869.
2. Galton, F., Regression towards mediocrity in hereditary stature. *J. Anthropol. Inst. G.B. Ireland*, 1886, **15**, 246–263.
3. Gallon, F., Typical laws of heredity. *Proc. R. Inst.*, 1877, **8**, 282–301.
4. Galton, F., Family likeness in stature. *Proc. R. Soc*., *London*, 1886, **40**, 42–73; Includes appendix by J. D. Hamilton Dickson, *ibid.*, 63–66.
5. Karandikar, R. L., Mathematics and elections. *Proc. Indian Natl. Sci. Acad.*, 2020, **86**, 1461–1479.
6. Demoulin, C. and Embrechts, P., Revisiting the edge, ten years on. *Commun. Stat. – Theory Meth.*, 2010, **39**, 1674–1688.
7. Donnelly, C. and Embrechts, P., The devil is in the tails: actuarial mathematics and the subprime mortgage crisis. *ASTIN Bull.*, 2010, **40**, 1–33.
8. Embrechts, P., Did a mathematical formula really blow up wall street? https://www.actuaries.org/ASTIN/Colloquia/Helsinki/Presentations/Embrechts.pdf (accessed on 9 February 2021).
9. Yule, G. U., Why do we sometimes get nonsense-correlations between time-series? – a study in sampling and the nature of time-series. *J. R. Stat. Soc.*, 1926, **89**, 1–63.
10. Chatterjee, B., Karandikar, R. L. and Mande, S. C. Mortality due to COVID-19 in different countries is associated with their demographic character and prevalence of autoimmunity. *Curr. Sci.*, 2021, **120**, 501–508.
11. Hassler, U. and Thadewald, T., Nonsensical and biased correlation due to pooling heterogeneous samples. *J. R. Stat. Soc.*, *Ser. D*, 2003, **52**, 367–379.
12. Good, I. J. and Mittal, Y., The amalgamation and geometry of two-by-two contingency tables. *Ann. Stat*., 1987, **15**, 694–711.
13. Pearson, K., Lee, A. and Bramley-Moore, L., Genetic (reproductive) selection: inheritance of fertility in man and of fecundity in thoroughbred racehorses. *Philos. Trans. R. Soc. A*, 1899, **192**, 257–330.
14. Simpson, E. H., The interpretation of interaction in contingency tables. *J. R. Stat. Soc.*, 1951, **13**, 238–241.
15. Yule, G. U., Notes on the theory of association of attributes in statistics. *Biometrika*, 1903, **2**, 121–134.
16. Bickel, P. J., Hammel, E. A. and O'Connell, J. W., Sex bias in graduate admissions: data from Berkeley. *Science*, 1975, **187**(4175), 398–404.
17. Badyal, D. *et al.*, Hydroxychloroquine for SARS CoV2 prophylaxis in healthcare workers – a multicentric cohort study assessing effectiveness and safety. *J. Assoc. Physicians India*, 2021, **69(**6), 11–12; https://www.japi.org/x284d434/hydroxychloroquine-for-sars-cov2-prophylaxis-in-healthcare-workers-ndash-a-multicentric-cohort-study-assessing-effectiveness-and-safety
18. Smith, L. G. *et al.*, Observational Study on 255 Mechanically Ventilated Covid Patients at the Beginning of the USA Pandemic, 2021; medRxiv preprint doi:https://doi.org/10.1101/2021.05.28.21258012 (accessed on 31 May 2021).
19. Ellenberg, J., *How not to be Wrong: The Power of Mathematical Thinking*, Penguin Press, 2014.
20. Wallis, W. A., The statistical research group, 1942–1945. *J. Am. Stat. Assoc*., 1980, **75(**370), 320–330.

ACKNOWLEDGEMENTS. I thank Prof. Tapen Sinha (University of Nottingham) and Dr Srinivas Bhogle (Honorary Scientist, CSIR-Fourth Paradigm Institute, Bengaluru) for help while preparing the manuscript and the referees for providing useful suggestion that helped improve this manuscript.

Received 2 July 2021; revised accepted 7 September 2021

doi: 10.18520/cs/v121/i8/1016-1021