DOI: http://doi.org/10.4038/sljastats.v21i3.8063

Sri Lankan Journal of Applied Statistics, Vol (21-3)

A Review on Misleading COVID-19 Statistics

D. Y. Jayasinghe[1], P. Dias[2] and C. L. Jayasinghe[3]

[1,2,3]Department of Statistics, University of Sri Jayewardenepura, Sri Lanka.
[*]Corresponding Author: dovinij@gmail.com

Received: 05th November 2020 / Revised: 27th December 2020 / Published: 31st December 2020

ABSTRACT

Coronavirus (COVID-19) is an infectious disease caused by a newly discovered virus (WHO, 2021) which has established worldwide spread and hence has become the most discussed topic. In explaining the severity of the disease, various statistics are utilized by various governing/non-governing bodies as well as by the general public. Intentionally or unintentionally a plethora of misleading statistics including count data comparisons, graphical representations, indices etc. have been released daily to media all around the world, irrespective of the computational complexity. This paper presents a collection of such instances along with possible rectifications to mitigate the adverse effect of misguided decisions made.

Keywords: Misleading statistics, COVID-19, Data representation, Indices, Count data

1 Introduction

Statistics is a platform for the data to speak out. In other words, it provides information for decision making based on data. Nonetheless misleading information conveyed through misleading statistics used purposely or without knowledge can lead to miscommunication and ultimately poor/incorrect decisions. Simply researchers arrive at conclusions with respect to the summary measures they get, inferences they make, so on and so forth. The year 2019 coronavirus outbreak in Wuhan, China (Huang et al., 2020) has sparked a global pandemic. Among all the situations where statistics have misled the viewers/readers, COVID- 19 outbreak is the latest instance where numerous statistical misrepresentations are utilized and concepts are misinterpreted by various governing and non-governing bodies such as media institutions. According to Pennycook et al. (2020), giving regular prompts by various mass media platforms on the idea of accuracy might be enough to enhance people's sharing decisions related to COVID-19 information and ease the volume of misinformation on the virus. This article attempts to review and provide constructive criticism on such instances of misleading statistics established/published throughout the COVID-19 pandemic to date.

2 Misleading COVID-19 Statistics

In this section, various types of misleading statistics used all around the world are discussed. It includes both numerical and graphical misleading components such as count data, charts, indices and proportions.

2.1 Count data

One of the most primitive measures utilized to compare the severity of COVID-19 is the total of some variable of interest such as number of deaths, number of active cases, number of total confirmed etc. Counts can be useful to show when incidence is starting to recede as public health measures take effect in a particular population. As per the findings of Pearce et al. (2020), the shape of the portrayed trends in case counts enabled to see at the time of the writing of that paper that the United Kingdom, France, Italy, and Spain are on similar trajectories, whereas Korea and other Asian countries have been "flattening the curve". However, use of these types of count data should be directly exempted when making comparisons, due to several reasons.

Any mathematician or even any non-mathematician can undoubtedly agree with the fact that a comparison would be fair enough only when the compared individuals are kept under similar background conditions. The variable utilized in the comparison should be the only varied factor to make the comparison fair. In simple language, it is not fair to compare apples and oranges/dollars and coconuts. Similarly, the same theory is applied here. If we need to perform a country wise comparison, it is incorrect to use only count data, since different countries have different land areas, populations with different age distributions, various lifestyles, different dietary habits and different compositions with respect to gender, ethnicity etc. Simply, if we compare the count of confirmed cases in USA with Sri Lanka, it is obvious that USA reports a higher number with respect to the geographical region of country's spread and there-by the total population in the country. Hence, a more valid comparison can be conducted by considering not the total number, but the total deaths per million. This eliminates effect of total size of the populations of the countries being compared. Figure 1 (Source: our world in data, 5 May 2020 morning update) reveals how the country rankings change when the death counts and death per million counts are considered. The USA, for example, moves from being ranked number one in total number of deaths, to number nine when adjusted per capita (with the caveat that US deaths do not yet seem to have stabilized). Belgium makes the reverse move, going from number six in overall deaths to being the country with the most deaths per million population. On this measure, the UK (following the recent inclusion of care home deaths) now has the third highest number of deaths in the world (Krelle et al., 2020). Therefore, it is clear that using count data for comparing the severity levels of the outbreak on two different geographical regions is misleading.

In order to detect COVID-19 cases, different tests are performed. In Sri Lanka, primarily PCR tests are conducted to detect COVID-19 and more recently Anti-gent tests are utilized as well. Distinct diagnostic tests possess different testing accuracies, hence may reveal dissimilar results. Hence, comparisons related to count data (e.g. total number of cases detected) are more valid when detection has been done using the same test. Furthermore, "random testing" and "testing contacts" are two different testing strategies. It is obvious that there is a high chance of getting a positive test result for close contacts of a COVID case and hence the positive rate is higher when 'testing contacts' strategy is adopted. Conversely, the positive rate may be low for a randomly chosen sample especially if the community transmission stage has not been reached. Therefore, collectively all the aforementioned points support the idea that count data could be misleading especially when conducting comparisons.

Rank	Country	Total deaths
1.	USA	68,934
2.	Italy	29,079
3.	UK	28,734
4.	Spain	25,428
5.	France	25,201
6.	Belgium	7,924
7.	Brazil	7,321
8.	Germany	6,831
9.	Iran	6,277
10.	Netherlands	5,082

Rank	Country	Deaths per million
1.	Belgium	684
2.	Spain	544
3.	Italy	481
4.	UK	423
5.	France	386
6.	Netherlands	297
7.	Sweden	274
8.	Ireland	267
9.	USA	208
10.	Switzerland	171

Figure 1: Comparing highest ranked countries by total COVID-19 deaths and by deaths per million *(Source: our world in data, 5 May 2020 morning update) - note that countries of fewer than 100,000 population have been excluded.*

2.2 Charts

Figure 2 is extracted from the COVID-19 Dashboard presented by Health Promotion Bureau (HPB) Sri Lanka on 24th December 2020. Although the computed statistics are acceptable, having these graphs presented on the same place without scaling the two graphs as a whole seems misleading. For example, if we consider France, it has a fatality rate of 2.47% while the recovery rate is 7.47%. If we compare these rates of the country France itself, it is quite straightforward that France has a recovery rate higher than the fatality rate. When these two graphs are presented simultaneously, the length of the bars tend to mislead the viewer by representing 2.47% (chart in the left) by a longer bar than the bar representing the value 7.47% (chart in the right). The aim of representing data via graphs should be letting the viewer grab the facts instantly without a thorough observance. This norm is violated here, where it has to be counted in for the list of misleading COVID statistics.

Further to this, it is clear that the fatality rate and the recovery rate have been computed as in Equations (1) and (2) (which is why the summation of the 2 rates does not equal to 100%). The rates should be calculated from the resolved cases (i.e. excluding cases which are still under treatment). Here the total number of recovered cases were defined to be those who received negative results for two consecutive PCR tests (Ministry of Health, 2020).

$$Fatality\ Rate = \frac{Total\ Number\ of\ Deaths}{Total\ Cases\ Reported} \times 100\% \quad (1)$$

$$Recovery\ Rate = \frac{Total\ Number\ of\ Recovered\ Cases}{Total\ Cases\ Reported} \times 100\% \quad (2)$$

Figure 2: Horizontal bar charts of fatality and recovery rate comparisons

Figure 3 is a bar chart containing misleading data representation, which is published in the Health Promotion Bureau (HPB) Sri Lanka website. As per the partial image it reveals that number of PCR tests conducted on 19th February 2020 is null, but there is some height indicated in the chart itself. If it was correctly represented, then there shouldn't be a vertical bar.

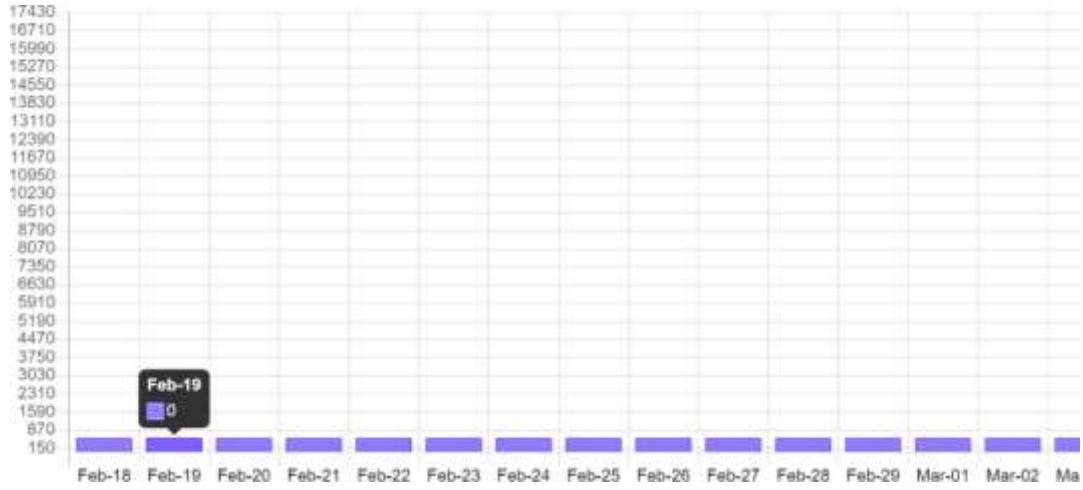

Figure 3: Partial bar chart of daily PCR test conducted in Sri Lanka

2.3 Indices

Koetsier (2020) has introduced a safety index to rank the 100 safest countries during pandemic. They have allocated scores for each and every country in the world with respect to 6 factors namely quarantine efficiency, government efficiency of risk management, monitoring and detection, healthcare readiness, regional resiliency and emergency preparedness. These factors were obtained by using several other variables as shown in Figure 4 and the weighting schemes used are given in Figure 5. However a country becomes safer to live when the number of positive cases is reduced and number of deaths is comparatively less. The factors considered here is not unimportant, but it omits the basic consideration. Most importantly, it is doubtful whether it is fair enough to generalize a single scoring system to the entire world, since the countries are different in many aspects. The variables considered for scoring depend on country's situation and hence the weighting scheme is not consistent. For example, countries where there is war, cannot implement on efficient quarantine system. There can be countries which naturally did not face any severe emergencies and if that is the case, allocating an equal weight on emergency preparedness for all the countries is not fair.

Most importantly, different weighting systems cannot be incorporated, if the measurement is used for comparison.

Not only these factors, level of health education should be a major consideration in this scenario. According to Buckles et al. (2016) when level of health education is increased, death rates decrease. Further to this, geographical positioning, type(s) of geographic boundaries/borders a country possess may also be important in such an analysis. It could also be argued that the number of access points to the country (legally or illegally) should be considered. Anyhow, potential illegal migration definitely possesses a huge COVID-19 threat while risk entitled with legal migration can be controlled with appropriate safety measures. Therefore, it is clear that this scoring system requires an improvement considering the aforementioned points.

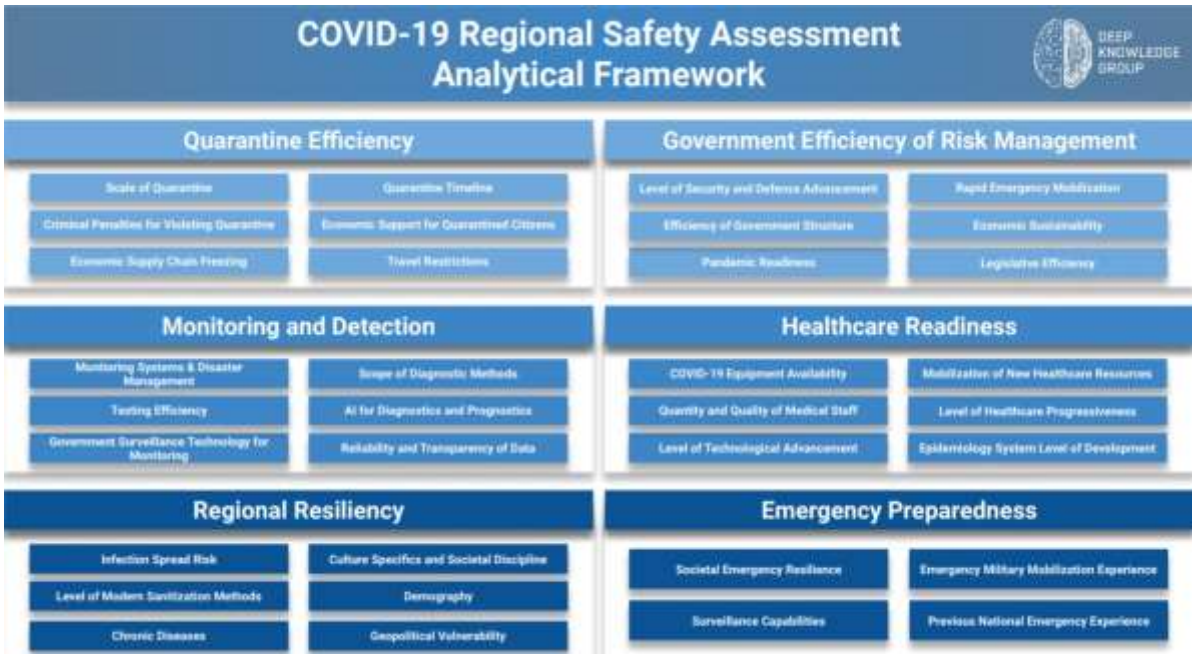

Figure 4: Variables used to obtain country scores
(Source: http://analytics.dkv.global/covid-regional-assessment-200-regions/methodology.pdf)

Figure 5: Weights used to obtain country scores
(Source: *http://analytics.dkv.global/covid-regional-assessment-200-regions/methodology.pdf*)

2.4 Proportions

To the place where count data is misleading, proportions take the place. It could be argued that the number of PCR tests conducted should be taken into account when comparing active cases on a daily basis. That speaks out when it is clear that none of the countries perform similar number of PCR tests per day - not only among the countries in the world, but also within the country itself. Say for example, in the first day, a country performed 100 tests to figure out new cases and found 5 active cases. In the same country, say the next day, 1000 PCR test were performed and found only 10 positive cases. Just because 10 is a larger value than 5, it is unfair to conclude that the number of positive cases of that country has doubled its value. To address this issue, proportions were taken in to account. See Equation (3) for the formula to compute the proportion of active cases.

$$Proportion\ of\ Positive\ Cases\ for\ a\ Day = \frac{No.of\ Positive\ Cases\ of\ the\ Day}{No.of\ PCR\ Tests\ Performed\ for\ that\ day} \qquad (3)$$

Table 1 shows the statistics of the aforementioned example. With respect to the positive cases, day 2 is worse than day 1 due to the increase number of detected cases, but when proportion is considered, the converse is correct; day 2 is better than day 1 due to decrements in the proportion of positive cases. As per this example, it is clear that, proportion provides a better representation in such a scenario.
However, when comparing the world data, it is very important to see how countries give definitions to the terms. According to Max Roser and Hasell (2020) the informed number of tests conducted by a particular country does not refer to the same in each country - one difference is that some countries report the number of people tested, while others report the number of tests performed (which can be higher if the same person is tested more than once). PCR tests are performed not only to detect new cases, but also to detect cured cases after treatment. In such an instance, it is definitely misleading to compare countries making the proportion of positive cases as the bottom line. Therefore, if a comparison is needed, it is highly advisable to check the definitions of the variables given by countries in order to avoid misleading representations.

Table 1: Statistics of the hypothetical example

Day	Positive Cases	PCR Tests	Proportion of Positive Cases
1	5	100	0.05
2	10	1000	0.01

3 Discussion and Conclusions

Amongst the various statistics utilized to date by various governing/non-governing bodies and the general public, a plenty of misleading statistics have been released/used intentionally or not. In this paper, a review was conducted on a collection of such instances including count data, graphs, indices and proportions that have been utilized/published.

It was highlighted that use of count data for comparisons under dissimilar background conditions is misleading. Similarly, data representation via graphs should be done with care such that the norms of presenting summarized data graphically are not violated and that they communicate the real scenario. Computations of fatality rates and recovery rates should be improved to showcase the actual figures, by excluding uncertain cases such as patients still under treatment who could recover or die. It was

also found that, using a common index/scoring system to compare all the countries in the world which possess dissimilar geographic/political factors are highly misleading.

Finally, when comparing proportions of daily positive cases, considering only the results related to daily PCR tests conducted for detection of new cases is imperative and if the intention of conducting the test is disregarded (whether it is to diagnose the disease or detect recovery) it could significantly impact the actual figures and hence the decisions made. Comparisons should be made (for example when comparing daily positive cases) with counts generated through the same diagnostic test. Further to this, since the test results are not reported on the same day, the number of positive cases and the number of tests performed cannot be matched. Therefore, it is important to be mindful regarding the definitions of these measures and consequently their computations to avoid generation of misleading facts/figures and thus inaccurate decisions.

References

1. Buckles, K., Hagemann, A., Malamud, O., Morrill, M. and Wozniak, A. (2016), 'The effect of college education on mortality', *Journal of Health Economics 50*, 99 – 114.
   URL: *http://www.sciencedirect.com/science/article/pii/S0167629616301382*
2. Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., Zhang, L., Fan, G., Xu, J., Gu, X. et al. (2020), 'Clinical features of patients infected with 2019 novel coronavirus in wuhan, china', *The lancet 395(10223)*, 497–506.
3. Koetsier, J. (2020), 'The 100 safest countries for covid-19: Updated'.
   URL: *https://www.forbes.com/sites/johnkoetsier/2020/09/03/the-100-safest-countries-for-covid-19-updated/*
4. Krelle, H., Barclay, C. and Tallack, C. (2020), 'Understanding excess mortality the health foundation'.
   URL:*https://www.health.org.uk/news-and-comment/charts-and-infographics/understanding-excess-mortality-the-fairest-way-to-make-international-comparisons*
5. Max Roser, Hannah Ritchie, E. O.-O. and Hasell, J. (2020), 'Coronavirus pandemic (covid-19)', Our World in Data.
   URL: *https://ourworldindata.org/coronavirus*.
6. Ministry of Health, I. M. S. (2020), 'Covid-19 laboratory test strategy in Sri Lanka: Version 2'.
   URL: *www.epid.gov.lk/web/images/.../final˙draft˙of˙testing˙strategy˙v˙2.pdf*
7. Pearce, N., Vandenbroucke, J. P., VanderWeele, T. J. and Greenland, S. (2020), 'Accurate statistics on covid-19 are essential for policy guidance and decisions'.
8. Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G. and Rand, D. G. (2020), 'Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention', *Psychological science 31(7)*, 770–780.
9. WHO (2021), 'Coronavirus'
   URL: *https://www.who.int/health-topics/coronavirus*.