Benford analysis of Covid-19 data

Andrew Jackson (ETH Zürich, Switzerland) and Malcolm Sambridge (RSES, ANU, Australia)

1 Summary

In the effort to react appropriately to the spread of Covid-19 it is essential to have access to data concerning the effects of the virus; the two most commonly reported statistics are persons known to have tested positive for the virus (called “Cases”) and persons who have succumbed to the virus (“Deaths”). It is important that the data are timely and accurate; in the case of the latter, how can this be checked? Here we show that datasets such as the cumulative case and death data from different countries can be analysed by comparing the reported integers to the expected distribution for such rapidly-growing statistics, which are expected to have a Benford-law type distribution. Only the leading digit is of importance in such an analysis. We show that there is an expected good agreement between theory and reported values, although there are occasional strange deviations.

2 Analysis

Data for all countries analysed were obtained from https://github.com/mhandley/COVID19/ on 9/4/20 (courtesy Prof. M. Handley, UCL, UK), who has distributed data from national authorities and occasionally from the WHO daily briefing. These data tend to be more up-to-date than the definitive Johns Hopkins data. We take as data for each country the cumulative cases or cumulative deaths as reported in the country files and neglect the date, so that the only numbers are the growing cumulative totals. We create two “All countries” datasets for cases and for deaths by concatenating the country data. These dataset consist of 2185 and 2070 integers respectively. The latter is slightly smaller than the former because on some days when there are cases, no death information is reported. The data we acquired contained no death information for China.

We begin by analysing the two “All counties” datasets for cases and deaths in Figures 1a and 1b. In all figures the dashed curve represents the expected Benford 1st digit distribution, in which there are around 30% leading digits that are “1”s and around 5% that are “9”s.

We see a clear adherence from all the data to the expected distribution. We can see this also at a country level (Figure 2). We take as examples the European countries France, Germany, Spain, UK, Switzerland, Italy. All these countries have experienced large growths in numbers. Since the datasets are smaller, we expect to see more variability around the expected distribution. Such variability would diminish as the datasets increase in size. We have examined the plots of 53 countries, and find general agreement in their behaviour, similar to in Figure 2.

There are, however, very rare outliers that signal a behaviour far from expected. An example is given in Figure 3. The likelihood of having so many data with “9”s is very low, and one can now trace this back to the original data. Subsequent days reported 9 deaths, and then two weeks later, subsequent days reported 90 and 99 deaths.1 Figure 4 shows the correlation coefficient for all countries’ case data when compared to the theoretical curve.

It is generally accepted that compliance of data sets to Benford’s prediction improves as both numbers of data and their dynamic range increase. However, with Covid-19 data, as cumulative cases numbers begin to flatten in each country, Benford compliance is likely to decrease because first digits tend to all become the same, significantly distorting relative digit frequencies. This is already the case with China and South Korea in the datasets examined here, and will likely become the case with many countries over time.

---

1Evidence of the reluctance to report figures passing a particular threshold can be found in statements to the media: Iran’s daily death toll exceeded 100 for the first time since 13 April, taking the overall number of deaths to more than 8,800. “It was very painful for us to announce the triple-digit figure, the health ministry spokeswoman, Sima Sadat Lari, said. This is an unpredictable and wild virus and may surprise us at any time.” (Guardian online 15/6/20, article entitled “Global report: China detects highest daily case rise in months”.)
More information is at https://github.com/anu-ilab/Covid-19-Benford-Analysis where updated data sets, figures and analysis can be found.
Figure 1: Comparison of data from all countries together with the expected distribution (dashed). a) Cases. b) Deaths.

Figure 2: Comparison of data from selected European countries together with the expected distribution (dashed). Countries are France (blue), Germany (red), Spain (green), UK (cyan), Switzerland (Orange) and Italy (black). a) Cases. b) Deaths.

Figure 3: Death data reported by the Czech Republic as of 9/4/20.
Figure 4: Correlation coefficient for each of the countries’ case data that was analysed.
Cases by country
Benford austria Cumulative data
Benford canada Cumulative data
Benford data Cumulative data
Benford Greece Cumulative data
Benford hungary Cumulative data
Benford Iran Cumulative data
Benford Ireland Cumulative data
Benford japan Cumulative data

Frequency (%) vs. First Digit
Benford Malaysia Cumulative data
Benford Netherlands Cumulative data
Benford Pakistan Cumulative data
Benford poland Cumulative data
Benford portugal Cumulative data
Benford Russia Cumulative data
Benford sweden Cumulative data
Benford Switzerland Cumulative data

Frequency (%) vs First Digit
Benfords Thailand Cumulative data

Frequency (%) vs First Digit
Benford turkey Cumulative data
Deaths by country
Benford australia.dat Cumulative data
Benford austria.dat Cumulative data
Benford brazil.dat Cumulative data
Benford czechrepublic.dat Cumulative data
Benford denmark.dat Cumulative data
Benford \textit{estonia.dat} Cumulative data
Benford france.dat Cumulative data
Benford greece.dat Cumulative data
Benford hungary.dat Cumulative data
Benford india.dat Cumulative data
Benford Iran.dat Cumulative data
Benford Ireland.dat Cumulative data
Benford italy.dat Cumulative data

![Graph showing Benford's Law for cumulative data for Italy. The graph plots the frequency of leading digits against their occurrence count, demonstrating the expected distribution according to Benford's Law. The x-axis represents the leading digit (1 to 9), and the y-axis represents the frequency. The graph shows a decreasing trend from 1 to 9, with the first digit 1 occurring more frequently than the second digit 2, and so on, as expected by Benford's Law.](image-url)
Benford malaysia.dat Cumulative data
Benford NewZealand.dat Cumulative data
Benford norway.dat Cumulative data
Benford qatar.dat Cumulative data
Benford Singapore.dat Cumulative data
Benford uk.dat Cumulative data
Benford us.dat Cumulative data