EASY-QUICK-CLEAN-EFFECTIVE SCREENING TOOL FOR IMPROVING THE INTEGRITY OF DATA

BY
DR. MAI ZURWATUL AHLAM BINTI MOHD JAFFAR
SENIOR LECTURER, FACULTY OF SCIENCE
UNIVERSITI PUTRA MALAYSIA

Data is essential to serve an intended purpose including planning, decision-making, policy - making and operation for any organization. High data quality provides more insights which leads to more effective decision-making. However, data quality problems are widespread in practice. Factors that safeguard the quality of data includes reliability. Reliable data refers to absence of error or fraudulence for instance manipulation and alteration that breach consistency of the data. One thing to note is that reliability is not validity, but it is an essential prerequisite for validity. Data can be reliable without being valid. While data integrity refers to characteristics that determine data reliability throughout its lifecycle describing its quality such as valid or invalid or the process of ensuring and preserving the validity and accuracy of the data.

In the light of the above, the integrity of any data acts as the foundation upon which plan, decision, policy, and operation are built. Imagine making an extremely important healthcare decision depending on the data that is entirely, or even partially, inaccurate. If healthcare data collected and recorded by pharmaceutical manufacturers is altered, then the worst impact is patient safety and the loss of lives. Another example of an extremely important data-driven decision that demands thorough data integrity is related to policies of global warming and climate change. This is because such data offers a range of perspectives for insightful actions needed to formulate appropriate responses to combat global warming as well as climate change globally. Indeed, issues of data integrity pose such a high risk.

A polar opposite of data integrity is data corruption involving fraudulence that basically means intention of changing the data. It is important to realize that the key difference between misconduct or fraud and honest error is intent. Bear in mind that fraud refers to intent to cheat. For instance, forgetting to key in a value is honest error, deliberately not to key in it is fraud. Another example is incorrectly jotting down a value is honest error, but purposely changing its value is fraud. Data corruption occurred in accounting, banking, economics, corporate finance, insurance, tax and many more.

Perhaps, one solid motive that led to data corruption is closely related to general expectation such that many people often looking up the leading digits in the numbers they encounter daily are all numbers are equally likely to be the leading digit. In practice, this is

translated as the leading digits are distributed uniformly with a similar expected frequency of 19≈11.11% each. However, in 1881, American polymath Simon Newcomb proposed that such expectation is actually a pitfall of elementary statistics. Newcomb first observed the pattern of the first few pages in logarithmic tables. Initially, he noticed the pages were much dirtier than the last pages. This indicated that researchers spent a lot of time dealing with numbers beginning with 1, less time with numbers beginning with 2 and so forth. This prompted him to suggest that the pages of numbers whose leading digit was 1 were more worn than the pages of numbers whose leading digit was 9.

Newcomb's article published by the *American Journal of Mathematics entitled by Note on the Frequency of Use of the Different Digits in Natural Numbers* basically states that digits in fact appeared to be heavily skewed towards low numbers obeying a well-defined uniform logarithmic distribution. Also, he proposed the probability that a number begins with the leading digit d is equal to d+1d. As a result, the most frequent leading digit is 1 with expected frequency of 30.1%, digit 2 with expected frequency of 17.6%, digit 3 with expected frequency of 12.5%, digit 4 with expected frequency of 9.7%, digit 5 with expected frequency of 7.9%, digit 6 with expected frequency of 6.7%, digit 7 with expected frequency of 5.8%, digit 8 with expected frequency of 5.1% and the least common digit is 9, with an expected frequency of 4.6%. Newcomb's observation did not attract researchers for 50 years. In 1939, Frank Benford rediscovered the same pattern that Newcomb had observed repeatedly as he looked at a variety of different datasets. Benford observed at a huge sample size with over 20 000 numbers from various data-collections and analysed all of them. As a result, Benford came to the same conclusion with what Newcomb had mentioned in 1881. Furthermore, Benford managed to show that such a distribution applies to a wide range of phenomena and this is when the first real explosion of interest concerning pattern of digits in numbers started to attract attention among researchers. Therefore, the law is named after Benford though Newcomb discovered it. In 1995, Hill proposed a thorough proof of Benford's law based on probability theory which can be summarized as scale invariance implies base invariance and hence, base invariance implies the Benford's Law.

Natural law is associated by scale invariance and so does Benford's law. This means that the law is independent of man-made measurement systems or concepts. Sets of numbers that satisfy the law are usually naturally occurring random numbers that originate from multiple different distributions and expand many orders of magnitude. To clarify, naturally occurring numbers are numbers that are not sequential or man-made. Important to note that sets of numbers such as serial numbers or license car plates are not random.

The Benford's law is an efficient digital analysis technology, for example, in applicability verification, anomaly data detection and cross application with other methods. Note that the law is frequently used for surveillance and detection of fraud and money laundering. The use of the law is regarded as being superior to other traditional approaches because it is time-economical (easy-quick-clean-relative effective) overpowering some of the key concerns surrounding these approaches. This is a significant benefit of applying the law. More importantly, the law helps improve data quality which results in improving the integrity of data. The law is deemed as a useful tool for reference, where, in practice, it has been proven as an effective screening tool to indicate if a dataset deserves a deeper analysis or not.