My Kingdom for a Truth Compass

Stefanos Kechagias

SAS Institute

October 2022

Traditional definitions of Statistics depict it as the science of collecting and analyzing data aiming to understand large population characteristics. For most scientists, however, Statistics is much more than that. It is a powerful and indispensable truth compass that facilitates the advancement of knowledge by supporting and validating scientific works across all disciplines.

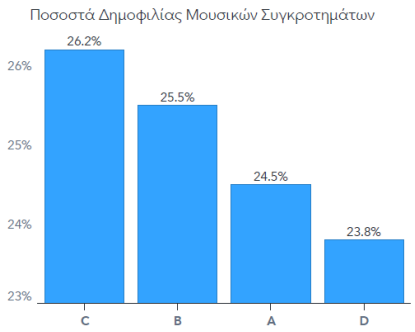
Despite its established importance in scientific progress, Statistics is still poorly understood by a large portion of our society and regularly misused by the general populace. Government officials, justice representatives, decision makers, authors, and public speakers (i.e., everyone these days) happily establish unfounded causations, and thoughtlessly (?) misinterpret ill-constructed graphs, thus transforming a truth-seeking discipline into a generator of falsehoods. Against the continuously growing menace of misinformation, we are alas found unprepared (a deficit we gruelingly felt during the pandemic) as we lack proper *data literacy* foundations, i.e., the ability to read, understand and communicate data and their stories. There is no question that education reforms and government policies aimed to reskill workforces and cultivate data-driven business landscapes are well past due. Beyond state and third-party responsibilities, however, we as citizens can (and have an inherent obligation to) modernize and bolster our critical thinking by building proper data literacy foundations that will equip us with a trustworthy truth compass. Let us take a look at some examples of data literacy principles:

1. *Who run the analysis and what methods were used?*
   It is the question we must always start with when we come across a data analysis. In clinical trials, for example, an ill-designed grouping of patients that ignores gender and age will a priori doom any hope of properly evaluating the administered drug, as any observed difference in the proportions of healed individuals between those who received a real dose and those who received a placebo cannot be safely attributed to the drug and not to the demographic differences the analyst who designed the trial failed to control for. Suppose that in an experiment of ten coin tosses we observe eight heads and two tails. Is this an aberration, one that would surely by redeemed by simply tossing the coin more times? Or does the coin have an unfair weight distribution that favors heads over tails in which case the 80% proportion of heads we observed is indeed the "truth" of the coin? Whether it is coin tosses, drug efficacy or pretty much any other theory needing to be supported by evidence, a "statistical compass" will point us to the truth via proper quantification of uncertainty.

2. *Summary measures:* Giannis Antetokounmpo's *average* three-point percentage in his last five NBA seasons is 29%, 6% lower than the NBA average. Based on this information his opponents systematically dare Giannis to shoot three-point shots hoping to deny his lethal drives to the basket (for the most part it doesn't work). Suppose now we wanted to study a country's income.

Could we also use the sample average to describe it? In contrast to shot making proportions that can only reach 100%, the top 1% of rich individuals would significantly influence our calculations and yield a large number that doesn't correctly represent the income of an average person. In this case the *median* is a far better choice of summary measure. Can you think of another quantity where the median should be preferred over the mean as a summary measure?

3. *A picture is a thousand words.* A well-constructed visualization is critical in providing insights to the truth, but a bad one can easily steer us in the opposite direction. While not all graph blunders are easy to spot there are some standard blunders to be mindful of. Axes with unnecessarily zoomed-in scales can erringly imply accentuated differences between groups. Poor color selection renders a graph unreadable for color-blinded people, and inappropriate for black and white printed copies. Unnecessary chart elements (not another 3D bar chart!) create confusion while unlabeled axes invite our imagination to interpret things that should be clearly specified by the analyst who created them. What issues can you spot in the diagram above?



4. *Correlation does not imply Causation:* European countries with larger populations of storks also exhibit large number of human births every year! In contrast to our admittedly amusing childhood stories, storks don't deliver babies! It is the country size that acts as a lurking or confounding variable causing both the number of births and storks. Have you read a recent article with phrasing that implied causation based on correlation?

5. *Ill-constructed surveys*: If you wanted to estimate the average height in your neighborhood's high school would you survey kids playing at the basketball court? If the school's basketball teams is practicing the time you gather your sample, you could very well end up with an overestimate, caused by what is known as *sampling bias*. *Response bias* is a tricker notion and comes in several forms. For example, a survey question on a politician's popularity with only yes or no as possible answers forces neutral responders to submit an opinion different than their own, thus leading to (one type of) response bias. Have you seen or taken such a survey lately?

6. *Relative Quantities.* During the first months of the pandemic the Greek National Public Health Organization reported daily the number of occupied and available COVID ICU beds across Greece. In a regional comparison of the healthcare system's stress, however, a relative quantity (say the available beds per 100,000 regional residents) would be a better suited measure to consider especially given the deruralized nature of Greek geography.

The examples we discussed may sound extreme, but they are not far from reality, as one regularly comes across similar blunders in published analyses and scientific articles appearing at poor-quality or predatory journals. And though few doubt the plethora of benefits our society has reaped from the ongoing digital revolution, the increasing impact of data in our lives is not a free lunch. In addition to the misinformation threat, more complicated issues around data such as security, protection, availability, and privacy – just

accept the cookies and show me the cat who punched the bear already – continue to surface and grow, imperatively calling for our reaction.

We hope this short exposition has prompted and inspired you to calibrate your own truth compass by researching, learning, and adopting data literacy principles. Two great resources to get you started are the excellent YouTube crash course by Jessica Pucci (see https://www.youtube.com/watch?v=yhO_t-c3yJY&t=2s) and the Data Literacy Essentials Training by SAS (see https://curiosity.sas.com/en/courses/data-literacy-essentials.html). Enjoy and keep your truth compass calibrated.