Social Science & Medicine xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Social Science & Medicine

journal homepage: www.elsevier.com/locate/socscimed

The "average" treatment effect: A construct ripe for retirement. A commentary on Deaton and Cartwright

S.V. Subramanian[a,b,*], Rockli Kim[a], Nicholas A. Christakis[c,d,e]

[a] Department of Social and Behavioral Sciences, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[b] Harvard Center for Population & Department Studies, Cambridge, MA, USA
[c] Department of Sociology, Yale University, New Haven, CT, USA
[d] Department of Medicine, Yale University, New Haven, CT, USA
[e] Yale Institute for Network Science, Yale University, New Haven, CT, USA

"Don't cross a river if it is (on average) four feet deep".
-Nassim Nicholas Taleb, 2016 p.160

1. Introduction

When summarizing or analyzing a population, regardless of whether it consists of hundreds or millions of individuals, it is the norm in most social, medical, and health research to characterize it in terms of a single number: the *average*. The reliance on average is pervasive in descriptive, explanatory, or causal analyses. There is nothing inherently wrong with an "on average" view of the world. But whether such a view is actually meaningful, for populations or individuals, is another matter. The average can obscure as much as it illuminates. It is a lean summary of a distribution with no recognition of the rich variation between and within populations that is necessary to ascertain its relevance. And, on rare occasions, when summaries of variation are presented in analyses of populations in epidemiology or clinical trials, they are often simply and incorrectly labeled "error."

In this issue, Angus Deaton and Nancy Cartwright (hereafter, Deaton and Cartwright) provide a comprehensive assessment and critique of the use of Randomized Controlled Trials (RCTs) in the social sciences (Deaton and Cartwright, 2018). Their insights and critique are equally applicable to biomedical, public health, and epidemiologic research. Here, we elaborate on one aspect of the problem that Deaton and Cartwright mention in their essay, namely, that inference exclusively based on "Average Treatment Effect" (ATE) can be hazardous in the presence of excessive heterogeneity in responses. This inferential problem applies both for the study population – those with the same characteristics as the trial population, including even individuals within the trial itself – and the larger population of interest the intervention targets. While the latter (*i.e.*, the issue of external validity in RCTs) has received considerable attention, including by Deaton and Cartwright, the former remains sidelined even as it underscores the intrinsic importance of variation in any population.

Instead of expecting ATE from an RCT to work for any individual or population, Deaton and Cartwright argue that we can do better with "*judicious use of theory, reasoning by analogy, process tracing, identification of mechanisms, sub-group analysis, or recognizing various symptoms that a causal pathway is possible*" (Deaton and Cartwright, 2018). Their hypothetical example of an RCT based on a classroom innovation in two schools, St Joseph's and St Mary's, is most intuitive in this regard. Deaton and Cartwright argue that even if the innovation turns out to be successful on average, actual experiences in the school with comparable composition may be more informative when other schools decide to adopt and scale up the same innovation (Deaton and Cartwright, 2018).

Following a brief introduction to the problems of averages, we elaborate on why variation or heterogeneity matters from a substantive perspective and develop a generalized modeling framework to assessing "Treatment Effect" (TE) based on two constructs of a population distribution: the average *and* the variance. We show that existing, but woefully under-utilized, methodologies can be routinely applied to enhance the relevance and interpretation of TE in a population. We refer to treatment as a shorthand for any deliberate intervention and not just in the strict medical sense. We focus on RCT settings here because both the mean and the variance in the outcome of interest are expected to be equivalent at baseline due to randomization and any differential in the post-treatment variation clearly indicates something systematic. However, the points we raise in this commentary applies equally, and in fact more importantly, to analysis of observational data.

2. The fallacy of averages

There is nothing innately problematic about focusing only on the mean to summarize a distribution, provided it has some substantive meaning and application to the real world. The yawning gap between a statistical average and its application to the real world of individuals is well recognized (Christakis, 2014). For illustration, we present two examples from Todd Rose's thought-provoking book, "*The End of Averages*" (T. Rose, 2016).

* Corresponding author. Professor of Population Health and Geography, Harvard Center for Population & Development Studies, 9 Bow Street, Cambridge, MA 02138, USA.
E-mail address: svsubram@hsph.harvard.edu (S.V. Subramanian).

https://doi.org/10.1016/j.socscimed.2018.04.027

In 1942, in a quest to discover an "ideal" form of a woman, Dr. Robert L. Dickinson (an obstetrician) and Mr. Abram Belskie (a sculptor) decided to measure ~15,000 young adult women on 9 body dimensions (*e.g.*, height, bust, waist, hips, thigh, calf, ankle, foot, weight) and, based on the "average" across each, sculpted a female form called "Norma" (Creadick, 2010). They then decided to launch a contest, "*Are you Norma*?", encouraging women to submit their bodily dimensions. Of almost 4000 submissions received, how many resembled Norma on all 9 dimensions? Exactly zero. Indeed, Norma represented a misguided ideal that was both highly desirable yet impossible to observe. What was the impact of this exercise? Instead of confronting the individual variability around constructs of "normality", most doctors and scientists concluded that American women were physically unfit (T. Rose, 2016).

The second example illustrates an even more consequential case. During World War II, the United States Air Force aircrafts were crashing at a higher-than-expected rate even though no mechanical and human errors could be detected. After much probing, the Air Force commissioned a study in 1950 to design a better fitting cockpit based on the average of more than 4000 pilots on 140 body measurements. Yet, when Lieutenant Gilbert S. Daniels did an exercise to see how many pilots fit the so called "average pilot" on 10 dimensions (*i.e.*, height, sleeve length, crotch height and length, and circumferences for chest, vertical trunk, hip, neck, waist and thigh), the answer was, yet again, zero (Daniels, 1952; T. Rose, 2016). Yes, even in such an evidently homogeneous group of airmen, it was impossible to find even one individual who fit the average on all dimensions, even when the average was generously defined as falling within the middle 30 percent of the range of values for each of the dimensions. Essentially, by designing the cockpit to fit the average airman, it was ensured that it fit no one. Daniels concluded, "*It is virtually impossible to find an "average airman" in the Air Force population […] not because of any unique traits in this group of men, but because of the great variability of bodily dimensions which is characteristic of all men*" (Daniels, 1952 p. 1).

3. The reality of variation

The above illustrative examples point to an important limitation concerning ATE even in an ideal RCT. For the ATE to be truly meaningful even within the limited trial sample population, we argue, two dimensions need to be considered.

First, there should be a systematic and a statistically significant difference in the average outcome between the Treatment and the Control groups in the expected direction (*i.e.*, treatment, on average, had the intended effect). If this occurs, the trial is considered a success and, after few repeated demonstrations of a similar ATE, is usually followed by recommendations for scaling up intervention.

A second consideration of equal importance is: of the sample population that received the treatment, what percentage actually experienced the intended effect? Stated differently, what is the regularity or predictability with which individuals in the Treatment group experienced the desired effect? In the extant literature, this dimension is completely ignored. Consider two successful RCTs, both showing systematic differences in ATEs. However, in RCT 1, 90% of the individuals in the Treatment group experience the desired effect while in RCT 2 only 10% of the individuals in the Treatment group experience any therapeutic benefit. The remaining individuals in both groups are either unaffected or experience changes in the unintended direction. Assuming these are two types of treatments intended to have a similar effect, which one of these would we consider more successful overall? Arguably, the treatment from RCT 1! The substantially higher degree of regularity and predictability with which the treatment worked in RCT 1 not only is desirable because the ATE now is more meaningful as it applies to a majority, it also suggests a better understanding of who are more susceptible to the treatment, and potentially the mechanism of "why" it works, and the judiciousness in designing the treatment.

We consider a toaster to be working if it is able to toast the bread every time it is used. One does not take solace from the claim that the bread will pop up toasted, say, 2 out of every 10 times. In clinical settings, however, if a drug works 20% of the time in RCT, compared with 5–10% for a placebo, it is often accepted to be "effective" (Christakis, 2008). For instance, among the top 10 highest-grossing drugs in the United States, Humira, Enbrel, and Remicade each works for 1 in 4 people who take them, and Nexium only works for 1 in 25 people who take it for heartburn. Statins are effective in lowering cholesterol for as few as 1 in 50 individuals (Schork, 2015). The truth, therefore, is that, most people taking RCT-validated, effective treatments derive no benefit from them; even in the study population (let alone the larger real-world population) (Christakis, 2008). As clinicians struggle in their efforts to understand low adherence to several prescribed medication regimens, it is worth considering if the low adherence is because patients realize that the medication does *not* work for them. In fact, the growing recognition that the effectiveness of different treatments are vetted for the actual individual patient has motivated "precision medicine" and N-of-1 trials (Schork, 2015).

The case for recognizing individuals and the variability that is observed between individuals in matters of health was eloquently made by Stephen Jay Gould in his classic commentary, "*The median isn't the message* " (Gould, 1985). In this personal story of statistics written after Gould was diagnosed with abdominal mesothelioma, an incurable disease with a median mortality of only eight months, he noted two important aspects about statistical distributions. First, the distribution of experiencing adverse events is more likely to be heavily skewed than normally distributed. Second, the distribution may alter when circumstances change. Gould embodied these characteristics as he lived for 20 highly productive years after the initial diagnosis (and extremely competent surgery).

Another example concerns why doctors tend to offer "Do Not Resuscitate" orders to AIDS patients at much higher rates than to patients with advanced liver cirrhosis even though these two conditions might have equal average prognoses (Wachter et al., 1989). It might be tempting to conclude that doctors are more eager to avoid resuscitation in AIDS patients, perhaps for discriminatory reasons. But the real reason might be that the *variance* in survival in the AIDS group is much higher, and there may be many more patients in that group who will die imminently. It may be to this fact (*i.e.*, the greater variance) that the doctors are more oriented rather than to the average survival of the two groups; the doctors may reason that they can wait to offer DNR orders to the cirrhosis patients (Christakis, 2014).

Most "successful" (*i.e.*, a "statistically significant ATE" in the expected direction) social, health, and medical interventions, we speculate, will be characterized by such poor regularity and certainty with which the treatment works among those who have received the treatment. Closing the gap between a robustly estimated, but mythical, "average" and its ability to say anything meaningful about the constituents of both the trial population as well as the real-world population has to be an integral part of any scientific endeavor that claims to be "useful" in its motivation and inference.

4. Why this fixation with averages?

The origins of use of average to describe a characteristic or trait in a population appears to trace back to Adolphe Quetelet's 19th century notion of "*l'homme moyen*" or the "average man" (Krieger, 2012; Porter, 1985; Quetelet, 1842). This metaphor of "average man" was derived from the fields of astronomy and meteorology where the results of observations from multiple observatories were combined to determine a star's celestial coordinates. Quetelet argued that the distribution of a population's characteristics composed of "deviations" or "errors" resulting from the imperfect variations of individuals is analogous to the data produced by each observatory in astronomy, and hence can inform a population's true (inherent) value (Krieger, 2012).