Statistics and Probability Theory

Toshihiko Mukoyama

February 2019, revised April 2024

When I started learning statistics, it was difficult to connect the descriptive statistics to statistical inference and testing. When there are 40 people in the classroom and I have data on their heights, are there more to analyze than transforming these 40 numbers, computing mean, variance, and so on? I understand that *if* these numbers are actually randomly sampled from a large population and if I am interested in the property of the whole population, it is possible to draw statistical conclusions using probability theory, because the random sampling *is* a probabilistic event. But these 40 people in the classroom are already there, and we didn't sample them randomly—what is there to infer from 40 numbers using probability theory, if the "whole population" we care is already there? Why do we have to estimate an unnatural thing, such as the "population mean" and the "population variance," other than the mean and variance of the heights of these 40 people? Why do we care about the standard errors of these estimates, when we don't even do the sampling once again?

Reading Kiyoshi Ito's essay on the history of probability theory, I realized that the key historical event for this connection was Bernoulli's proof of the Law of Large Numbers. If we flip the coin 40 times, the particular combination of heads and tails is a statistic, much like the heights in the classroom. Bernoulli showed, however, *if* we keep flipping coin (even if it doesn't occur in reality), the average fraction of heads can approximate the "true" probability of heads x, which represents the (hidden) property of the coin. Now *turning around* and putting this hidden property x in the center stage, he could *represent* the coin-flipping of 40 times as the outcome of the binomial distribution. The logic is: because x can (in principle) be recovered from the descriptive statistics (with the help of the Law of the Large Numbers), why don't we use this x to characterize the process of how these statistics are generated, even though we don't really see x? This logic allows us to jump from "what we see (40 flips) is the real thing" to "x is the real thing, and the 40 flips are merely its realizations." This "turning around" is a simple form of probabilistic modeling, which seems natural in the context of coin-flipping. (It is not only natural, but also elegant and useful.) It feels natural because we *know* that the model is correct—we do know that there is an actual coin behind the $\{\text{head}, \text{tail}\}$ events that possesses this property x. x is real, because it is embodied in the coin. And the Law of Large Numbers guarantees that x is not merely a fiction that allows for an elegant mathematical representation, but something that can be seen as the reality if the number of flips is sufficiently large.

Going back to the classroom example, by (fictitiously) imagining a hidden property of an abstract concept of height (e.g. a probability distribution that it follows), these 40 numbers *can* be mathematically represented with the language of the probability theory. The key, I believe, is that it *can* be—it doesn't have to be. It is perfectly fine to stop at the descriptive statistics if we don't want to commit to any model. To have a statistical inference, however, we have to imagine a model, where the classroom people's heights are drawn from a distribution that possesses some hidden properties (like x above). The question is: can we believe in such a model, when we don't *see* something (like coin above) that embodies these properties? The reason of my initial uneasiness in the connection between descriptive statistics and the inference was probably that, when the probabilistic model is unnatural (we don't draw people randomly to put in the classroom), this "turning around" of reality and the hidden parameter feels forced and unnatural. Even worse, we can't "draw" many times and use the Law of Large Numbers to see that these distributional properties are something real—there are only limited number of students in the world! This uneasiness is also a source of issues that often is associated with the "frequentist inference"—we cannot conduct this "turning around" without a complete faith in the underlying probabilistic model. Now I realize that when I felt uneasy, it was just that I felt that the statistical model behind the inference was not natural. It is this "unrealistic" (the "population" is not always something real, you just have to "believe in" its existence) feature of the frequentist model that makes it difficult for the initial learner to absorb the concept of the statistical inference.