Sailing Through Data: Discoveries and Mirages

Emmanuel Candès
Stanford University

Abstract

For a long time, science has operated as follows: a scientific theory can only be empirically tested, and only after it has been advanced. Predictions are deduced from the theory and compared with the results of decisive experiments so that they can be falsified or corroborated. This principle formulated by Karl Popper and operationalized by Ronald Fisher has guided the development of scientific research and statistics for nearly a century. We have, however, entered a new world where large data sets are available prior to the formulation of scientific theories. Researchers mine these data relentlessly in search of new discoveries and it has been observed that we have run into the problem of irreproducibilty. Consider the April 23, 2013 Nature editorial: "Over the past year, Nature has published a string of articles that highlight failures in the reliability and reproducibility of published research." The field of Statistics needs to re-invent itself to adapt to the new reality where scientific hypotheses/theories are generated by data snooping. We will make the case that statistical science is taking on this great challenge and discuss exciting achievements. In particular, we will introduce the method of knockoffs, which reliably selects which of the many potentially explanatory variables of interest (e.g. the absence or not of a mutation) are indeed truly associated with the response under study (e.g. the risk of getting a specific form of cancer).