Sailing through Data: Discoveries and Mirages

Emmanuel Candes

NOTE. An earlier version of this text appeared in the *IMS Bulletin*. See bulletin.imstat.org/2017/05/wald-lectures-emmanuel-candes.

For a long time, science has operated as follows: a scientific theory can only be tested empirically, 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 independently by Karl Popper and 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 irreproducibility. 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 and adapt to this new reality in which scientific hypotheses/theories are generated by data snooping. In my lecture, I will make the case that statistical science is taking on this great challenge and discuss exciting achievements.

An example of how these dramatic changes in data acquisition have informed a new way of carrying out scientific investigation is provided by genome-wide association studies (GWAS). Nowadays we routinely collect information on an exhaustive collection of possible explanatory variables to predict an outcome or understand what determines an outcome. For instance, certain diseases have a genetic basis and an important biological problem is to find which genetic features (e.g., gene expressions or single nucleotide polymorphisms) are important for determining a given disease. Even though we believe that a disease status depends on a comparably small set of genetic variations, we have a priori no idea about which ones are relevant and therefore must include them all in our search. In statistical terms, we have an outcome variable and a potentially gigantic collection of explanatory variables, and we would like to know which of the many variables the response depends on. In fact, we would like to do this while controlling the false discovery rate (FDR) or other error measures so that the results of our investigation do not run into the problem of irreproducibility. The lecture will discuss problems of this kind. We introduce "knockoffs," an entirely new framework for finding dependent variables while provably controlling the FDR in finite samples and complicated models. The key idea is to make up fake variables—knockoffs—that are created from the knowledge of the dependent variables alone (not requiring new data or knowledge of the response variable) and can be used as a kind of negative control to estimate the FDR (or any other error of type 1). We explain how one can leverage haplotype models and genotype imputation strategies about the distribution of alleles at consecutive markers to design a full multivariate knockoff processing pipeline for GWAS!

Some of the work I will be presenting is joint with many great young researchers including Rina Foygel Barber, Lucas Janson, Jinchi Lv, Yingying Fan, Matteo Sesia, as well as many other graduate students and post-docs, and also with Professor Chiara Sabatti, who played an important role in educating me about pressing contemporary problems in genetics. I am especially grateful to Yoav Benjamini: Yoav visited Stanford in the winter of 2011 and taught a course titled "Simultaneous and Selective Inference." These lectures inspired me to contribute to the enormously important enterprise of developing statistical theory and tools adapted to the new scientific paradigm—*collect data first, ask questions later*.

Emmanuel Candes

Credits

Author photo is courtesy of the John D. and Catherine T. MacArthur Foundation.

Emmanuel Candes holds the Barnum-Simons Chair in Mathematics and Statistics at Stanford University. His email address is candes@stanford.edu.

DOI: https://dx.doi.org/10.1090/noti1770