

STATISTICS COLLOQUIUM

TUESDAY, FEBRUARY 21, 2012

TALK: 4:00 P.M. — SCIENCE CENTER RM. 705

RECEPTION: 5:15 PM — SCIENCE CENTER, 7TH FLOOR

“The Intersection of Classical and Modern Statistics in the Social Sciences”

Luke Miratrix
University of California, Berkeley

ABSTRACT

With increased access to data on human activity, and increased allegiance to more statistical methods in the social sciences, there is currently a renaissance at the intersection of the social and statistical sciences. In this talk I discuss two projects, one more classical, and one of the digital age. I first present some analytical results on post-stratification, a form of covariate adjustment. In contrast to recent results showing that regression adjustment can be problematic under the Neyman-Rubin model, I show post-stratification, something that can easily be done in, e.g., natural experiments, has a similar precision to a randomized block trial as long as there are not too many strata. The difference is $O(1/n^2)$. Post-stratification thus allows for transparently exploiting predictive covariates and random mechanisms in observational data. I then present a framework for building statistical tools to extract topic-specific keyphrase summaries of large text corpora (e.g., the New York Times or other newspapers) and a human validation experiment to determine best practices for this approach. These tools, built from high-dimensional, sparse classifiers such as L1-logistic regression and the Lasso, can be used to, for example, translate essential concepts across languages, investigate the massive logs of air traffic controllers, or understand how different topics of interest are covered by various media outlets. In general, I argue that especially in the difficult domain of the social sciences we must spend extra attention on issues of validity. This motivates my using the Neyman-Rubin model for the analysis of post-stratification and my developing an approach for external, model-independent validation for the keyphrase extraction tools.