

STATISTICS COLLOQUIUM

WEDNESDAY, SEPTEMBER 02, 2009

TALK: 4:00 PM — SCIENCE CENTER, ROOM 705

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

DISCOVERING INFLUENTIAL VARIABLES: A METHOD OF PARTITIONS

Herman Chernoff
Professor Emeritus
Harvard University

ABSTRACT

A trend in all scientific disciplines, based on advances in technology, is the increasing availability of high dimensional data in which are buried important information. A current urgent challenge to statisticians is to develop effective methods of finding the useful information from the vast amounts of messy and noisy data available, most of which are noninformative. This paper presents a general computer intensive approach, based on a method pioneered by Lo and Zheng for detecting which, of many potential explanatory variables, have an influence on a dependent variable Y . This approach is suited to detect influential variables, where causal effects depend on the confluence of values of several variables. It has the advantage of avoiding a difficult direct analysis, involving possibly thousands of variables, by dealing with many randomly selected small subsets from which smaller subsets are selected, guided by a measure of influence I . The main objective is to discover the influential variables, rather than to measure their effects. Once they are detected, the problem of dealing with a much smaller group of influential variables should be vulnerable to appropriate analysis. In a sense, we are confining our attention to locating a few needles in a haystack.