HARVARD UNIVERSITY
DEPARTMENT OF STATISTICS

Tel. 617-495-5496
Fax. 617-496-8057
statdept@stat.harvard.edu
http://www.stat.harvard.edu/

Science Center 7th Fl.
One Oxford Street
Cambridge, MA 02138

STATISTICS COLLOQUIUM

MONDAY, APRIL 18, 2016
TALK: 4:15 PM — SCIENCE CENTER 300H
RECEPTION: 3:50 PM

"A Theoretical Framework for Prediction"

Herman Chernoff
Department of Statistics
Harvard University

ABSTRACT

As was pointed out in a previous article, the underlying theory and practice of *prediction* or *classification* for large biological data sets leaves much to be desired. Current practice with data sets where there are very many explanatory variables consists of a variable selection (*VS*) part where a few of the many variables are selected. Often these are selected by how well they correlate with the dependent variable. Then pattern recognition methods are used to produce an algorithm for classification. The prediction rate may be estimated by how well this works on an independent sample or by cross validation methods. Little is known about how well it is possible to achieve with those selected variables nor how well one could do with other subsets of the variables. In a previous article it was pointed out that depending on the significance of individual variables for *VS* has two major problems. Treating one variable at a time ignores the possible interactions of the effects of two or more variables. Significance can be a poor indicator of value for prediction, of one, or of a group of variables. In artificial examples we found that a *statistic*, *I*, that we had used previously for *VS* in a method called *Partition Retention*, was correlated with a *parameter* θ_C which measures ideal predictivity or $\theta_e = 1 - \theta_C$, the ideal error rate. We explain here why this is the case, and that *I* may be used to estimate an upper bound on θ_e. We show that the naive *training estimate* of θ_e is biased negatively and that a correction reduces the bias. Another estimate, the *out of sample estimate* is biased positively. We discuss some of the advantages of using *I* for the *VS* stage of prediction.

Joint work with S. H. Lo, T. Zheng and A. Lo.