abstract

Computer Science/Discrete Mathematics Seminar I
Topic:

Speaker:

Affiliation:

Date:

Time/Room:

In joint work with Paul Valiant, we consider the tasks of estimating a broad class of statistical properties, which includes support size, entropy, and various distance metrics between pairs of distributions. Our estimators are the first proposed estimators for these properties which use a sub-linear number of samples, and are based on a novel approach of approximating the portion of the distribution from which one has seen no samples. There are several implications of our results, including resolving the sample complexity of the "distinct elements problem" (given a vector of length n, how many indices must one query to accurately estimate the number of distinct elements?), and entropy estimation (given samples from a distribution of support n, how many samples are necessary to estimate the entropy to within an additive constant?); we show that for both problems, on the order of n/log n samples are sufficient.

Additionally, we show that, up to constant factors, our estimators are optimal, improving significantly upon the prior lower-bounds. The analysis of our matching lower bounds makes crucial use of two new multivariate central limit theorems that appear quite natural and general. The first is proven directly via Stein's method; the second is proven in terms of the first using a recent generalization of Newton's inequalities. The talk will include a high level overview of these techniques, and their application both in our context and more generally.