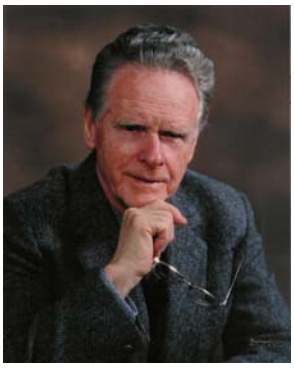
Arthur P. Dempster Award

Harvard University Statistics Department



The Arthur P. Dempster Fund "will support and recognize promising graduate students within the Department of Statistics, in particular those who have made significant contributions to theoretical or foundational research in statistics." It will be an annual award with a prize minimum of $2000. The expectation is to award one per year, though the faculty reserves the right to award two or none in any particular year depending on the quality of the submissions.

BREAKING NEWS: Anqi Zhao & David Jones have each been awarded the 2016 Dempster prize.

It was announced in March, 2015 that **Panagiotis Toulis** is the fourth Arthur P. Dempster award winner. Here is the abstract for his paper.

Implicit Stochastic Approximation for Principled Estimation with Large Datasets

The ideal estimation method needs to fulfill three requirements: (i) efficient computation, (ii) statistical efficiency, and (iii) numerical stability. The classical stochastic approximation of Robbins & Monro (1951) is an iterative estimation method, where the current iterate (parameter estimate) is updated according to some discrepancy between what is observed and what is expected, assuming the current iterate has the true parameter value. Classical stochastic approximation undoubtedly meets the computation requirement, which explains its popularity, for example, in modern applications of machine learning with large datasets, but cannot effectively combine it with efficiency and stability. Surprisingly, the stability issue can be improved substantially, if the aforementioned discrepancy is computed not using the current iterate, but using the conditional expectation of the next iterate given the current one. The computational overhead of the resulting implicit update is minimal for many statistical models, whereas statistical efficiency can be achieved through simple averaging of the iterates, as in classical stochastic approximation (Ruppert, 1988). Thus, implicit stochastic approximation is fast and principled, fulfills requirements (i-iii) for a number of popular statistical models including GLMs, GAMs, and proportional hazards, and it is poised to become the workhorse of estimation with large datasets in statistical practice.

It was announced in March, 2014 that **Peng Ding** is the third Arthur P. Dempster award winner. Here is the abstract for his paper.

A Paradox from Randomization-Based Causal Inference

Under the potential outcomes framework, causal effects are defined as comparisons between the potential outcomes under treatment and control. Based on the treatment assignment mechanism in randomized experiments, Neyman and Fisher proposed two different approaches to test the null hypothesis of zero average causal effect (Neyman's null) and the null hypothesis of zero individual causal effects (Fisher's null), respectively. Apparently, Fisher's null implies Neyman's null by logic. It is for this reason surprising that, in actual completely randomized experiments, rejection of Neyman's null does not imply rejection of Fisher's null in many realistic situations including the case with constant causal effect. Both numerical examples and asymptotic analysis support this surprising phenomenon. Although the connection between Neymanian approach and the Wald test under the linear model has been established in the literature, we provide a new connection between the Fisher Randomization Test and Rao's score test, which offers a new perspective on this paradox. Further, we show that the paradox also exists in other commonly used experiments, such as stratified experiments, matched-pair experiments and factorial experiments. (http://arxiv.org/abs/1402.0142)

It was announced in May, 2013 that **Bo Jiang** is the second Arthur P. Dempster award winner. Here is the abstract for his paper.

From SIR to SIRI: Sliced Inverse Regression with Interaction Detection

Variable selection methods play important roles in modeling high dimensional data and are keys to data-driven scientific discoveries. In this paper, we consider the problem of variable selection with interaction detection under the sliced inverse index modeling framework, in which the response is influenced by predictors through an unknown function of both linear combinations of predictors and interactions among them. Instead of building a predictive model of the response given combinations of predictors, we start by modeling the conditional distribution of predictors given responses. This inverse modeling perspective motivates us to propose a stepwise procedure based on likelihood-ratio tests that is effective and computationally efficient in detecting interaction with little assumptions on its parametric form. The proposed procedure is able to detect pairwise interactions among p predictors with a computational time of $O(p)$ instead of $O(p^2)$ under moderate conditions. Consistency of the procedure in variable selection under a diverging number of predictors and sample size is established. Its excellent empirical performance in comparison with some existing methods is demonstrated through simulation studies as well as real data examples.

It was announced in May, 2012 that **Alexander Blocker** is the inaugural Arthur P. Dempster award winner. Here is the abstract for his research presentation.

The Potential and Perils of Preprocessing: A Multiphase Investigation

Preprocessing forms an oft-neglected foundation for a wide range of statistical analyses. However, it is rife with subtleties and pitfalls. Decisions made in preprocessing constrain all later analyses and are typically irreversible. Hence, data analysis becomes a collaborative endeavor by all parties involved in data collection, preprocessing and curation, and downstream inference. Even if each party has done its best given the information and resources available to them, the final result may still fall short of the best possible when evaluated in the traditional single-phase inference framework. This is particularly relevant as we enter the era of "big data". The technologies driving this data explosion are subject to complex new forms of measurement error. Simultaneously, we are accumulating increasingly massive databases of scientific analyses. As a result, preprocessing has become more vital (and potentially more dangerous) than ever before.

In this talk, we propose a theoretical framework for the analysis of preprocessing under the banner of multiphase inference. We provide some initial theoretical foundations for this area, building upon previous work in multiple imputation. We motivate this foundation with two problems from biology and astrophysics, illustrating multiphase pitfalls and potential solutions. These examples also serve to emphasize the practical motivations behind multiphase analyses --- both technical and statistical. This work suggests several rich directions for further research into the statistical principles underlying preprocessing.