

December 2006  
Volume 2, Issue 5

## Inside this Issue

### *Special Interest Articles:*

- Contemporary issues in causal inference
- High performance computing at MSU

### *Individual Highlights*

|  |   |
|--|---|
| MSU Faculty Profiles                           | 2 |
| Profiles of MSU Consulting and Interest Groups | 3 |
| Finding a Statistical Collaborator             | 4 |
| Meet the CSTAT Team                            | 4 |
| Upcoming Events                                | 4 |

Center for Statistical  
Training & Consulting  
178 Giltner Hall  
Michigan State University  
East Lansing, MI 48824  
PHONE: (517) 353-9288  
FAX: (517) 353-9307  
E-MAIL: [info@cstat.msu.edu](mailto:info@cstat.msu.edu)  
WEBSITE: [www.cstat.msu.edu](http://www.cstat.msu.edu)

Editor: *Becky Scott*  
*PS Publications*

# Contemporary issues in causal inference

*by Dr. Ken Frank*

*Counseling, Educational Psychology, & Special Education; Fisheries & Wildlife*

The fundamental problem of causal inference is often defined by the counterfactual. For example: I have a headache; I take an aspirin; my headache goes away. Is it because I took the aspirin? It is impossible to know. We could be certain only if we observed what happened if I had not taken the aspirin. This control condition is impossible to observe for a single individual. It is counterfactual.

Recognizing our inability to observe the counterfactual at the individual level, scientists typically infer causality by comparing sets of units. A classic example would be a chemist who divides a single solution into two parts, one exposed to a treatment and one as a control. Resulting differences are then attributed to the treatment. This example is considered the gold standard for inferring cause.

But even the gold standard requires certain assumptions. Namely, that the treatment and control units are homogeneous. The units could be heterogeneous if the original solution is not properly stirred, if the two solutions are exposed to differences in conditions (such as temperature), or if they have different levels of purity.

Social scientists often analyze sets of units to make causal inferences. But causality is uncertain when analyzing an aggregate of individuals because there may be baseline differences between the treatment and control groups, or the treatment may affect the treatment group

differently than the control group.

To reduce baseline differences, social scientists often randomly assign subjects to treatment and control conditions. As sample sizes increase, randomization reduces differences between treatment and control groups. Therefore randomized control trials (RCTs) are considered the gold standard for inference in the social sciences.

But RCTs have limitations. First, the experiments on which they are based often differ from normal treatment conditions. To insure uniform treatment, experimenters may educate or engage those who implement the treatment in ways that would not occur outside the treatment. In education, the treatment might be implemented by 70% or 80% of the teachers, a level that might not occur for most reforms.

Second, randomization requires that people be treated and respond independently. In education, many reforms are implemented through the coordinated activity of teachers. Therefore teachers within a school are not independent. Many sources of dependencies are accounted for by carefully defining the units that can be considered independent (e.g., schools) but this can dramatically increase the cost of RCTs.

Given the above limitations of RCTs, social scientists often make causal inferences from observational data or quasi-experimental designs. Examples include the relationship between smoking and lung cancer, between job training and employment, and between socioeconomic

*Please see Causal inference, page 3*

## MSU Faculty Members w/ Statistical Expertise

In each CSTAT Newsletter, MSU faculty members with statistical expertise will be profiled.

### Dr. Christina DeJong

Dr. Christina DeJong is an Associate Professor and Director of Undergraduate Studies in the School of Criminal Justice. She received her Ph.D. in Criminology and Criminal Justice from the University of Maryland at College Park. As a graduate student, she served as the teaching assistant for the required graduate statistics courses and used a wide variety of statistical



Dr. Christina DeJong

programs during her graduate education, including SPSS, SAS, LIMDEP, Gauss, Mathematica and LISREL. Her dissertation used split population survival models to identify the factors related to the probability and timing of recidivism.

In addition to her work with survival models, Dr. DeJong has used hierarchical linear modeling to investigate how community-level factors affect juvenile justice processing, and structural equation modeling to determine how police officer gender and attitude affects officer treatment of citizens. Dr. DeJong's future research projects will focus on the use of geographic information systems to map and identify neighborhood characteristics associated with domestic violence in communities.

Dr. DeJong routinely teaches classes in quantitative methods to graduate students in the School of Criminal Justice, including regression analysis, categorical regression models, and structural equation modeling. In addition, she frequently assists students in the forensic science program with their statistical analysis. These research projects have ranged in topic from DNA extraction in ancient skeletal remains to identification of suspects through DNA on detonated pipe bombs. She is currently working with professionals in the forensic

science field to create a method for assigning probabilities of positive identification to human remains.

### Dr. Anil Jain

Dr. Anil Jain is a University Distinguished Professor in the Departments of Computer Science & Engineering, Electrical & Computer Engineering, and Statistics & Probability. He received his B.Tech. degree from Indian Institute of Technology, Kanpur, and M.S. and Ph.D. degrees from Ohio State University. At MSU, he has served as the chairman of the Computer Science Department.

His main research interests include statistical pattern recognition, data clustering, image processing and computer vision. In addition to developing new algorithms for feature extraction, classifier design, cluster analysis, object recognition and texture modeling, he has also been involved in a number of applications. These include document image understanding, remote sensing, medical image analysis, and biometric recognition. His current research projects deal with fingerprint matching, face recognition, data and classifier fusion, semi-supervised learning, and dimensionality reduction. These research projects are supported by NSF, ONR, ARO, and a number of private companies.



Dr. Anil Jain

He is the author of a number of books, including *Handbook of Multibiometrics*, *Handbook of Face Recognition*, *Handbook of Fingerprint Recognition* (which received the PSP award from the Association of American Publishers), *3D Object Recognition Systems*, *Markov Random Fields: Theory and Applications*, *Neural Networks and Statistical Pattern Recognition*, and *Algorithms For Clustering Data*, which is ranked #93 of the most cited articles in computer science of all time. He is holder of six patents in the area of fingerprint matching.

Awards received include best papers from the Pattern Recognition Society,

IEEE Transactions on Neural Networks Outstanding Paper Award, IEEE Computer Society Technical Achievement Award, a Fulbright Research Award, a Guggenheim fellowship, and the Alexander von Humboldt Research Award.

### Dr. Tenko Raykov

Dr. Tenko Raykov is a Professor of Measurement and Quantitative Methods at the College of Education. He holds a master's degree in probability and statistics, and a Ph. D. in statistical psychology from Humboldt University, Berlin, Germany.

His methodological interests center on applications of statistics, and in particular of latent variable and structural equation modeling, to measurement related problems in the behavioral and social sciences. These include methods for the evaluation of 'precision' (reliability) and the validity of measurement, especially for instrument construction, revision, and development. In addition, he is interested in longitudinal research using latent variable modeling, including issues pertaining to missing data.

Dr. Raykov has published extensively in the area of measurement reliability and validity estimation, repeated measure analysis, as well as fit assessment in structural equation and latent variable modeling. He is involved in research on enhanced feedback for advanced placement tests, as well as in longitudinal studies of depression and control, diabetes care, and arthritis. He is part of the Measurement and Quantitative Methods faculty in the College of Education, and teaches courses in structural equation modeling and latent variable modeling, applied multivariate statistics, and psychometric theory. With Dr. Marcoulides, he has recently published the second edition of their introductory text on structural equation and latent variable modeling.



Dr. Tenko Raykov

## *Profiles of MSU Consulting and Interest Groups*

# HPCC: high performance computing at michigan state university

At MSU's High Performance Computing Center (HPCC), researchers from across campus are increasing our understanding of the world and the universe by using advanced computing and simulation techniques. Twenty-four departments are currently making use of the facility including Management, Geography, Oncology, Plant Biology, and Education; fields that have not traditionally taken advantage of resources like this.

The HPCC provides, without charge, computer systems beyond



the capabilities of any single department on campus, allowing faculty and students to focus their time and resources on research rather than system administration. Analysis that previously took months or years now is now being completed in days and even hours.

HPCC director William Punch observes, "The HPCC has two goals. The first is to provide the computing infrastructure for researchers from across MSU to do work that they simply cannot do otherwise. That is, researchers can now address bigger, harder problems than they could before and quickly enough so the results can be useful. The second is to provide the expertise so researchers can take advantage of this infrastructure."

Located on the 3<sup>rd</sup> floor of the Engineering building, the HPCC combines the power of two types of "supercomputers." The first is an SGI Altix; a single computer with 128 processors, 576 GB RAM, and 6.4 TB of disk. This system is easy to use and is perfect for analyzing large data sets that won't fit on your desktop. Additionally, the HPCC houses a cluster; 128, four-processor computers wired together. This system is a little more challenging to use, but can easily handle tasks that be segmented into smaller pieces.

Matlab is a popular software package in use at the Center. The

statistics toolbox may be of particular interest to CSTAT users. R, which is used among CSTAT consultants, is another package recently installed. Please note, Yongfang Zhu, CSTAT consultant, has experience with HPCC systems and welcomes questions.

"We are very pleased by the breadth of research that the HPCC is serving and promises to serve, across a broad span of University disciplines. More rapid analysis not only speeds the delivery of analytical results, but also facilitates more complete and precise analysis, deeper explorations, and generally more productive scientific work. And when MSU's research faculty use the best tools, their students also gain experience with the best tools and with the kinds of leading-edge approaches to scholarship and problem-solving those tools enable," said David Gift, Vice-Provost for Libraries, Computing & Technology.

Visit <http://hpc.msu.edu> to learn more.

*HPCC is a collaboration of the College of Engineering, College of Natural Science, National Superconducting Cyclotron Laboratory, and Libraries, Computing & Technology, with substantial financial support from the Vice President for Research and Graduate Studies.*

### *Causal inference, from page 1*

status and achievement.

Causal inferences from observational studies are tenuous compared to RCT because of possible differences between the treatment and control groups. To account for these,

social scientists employ a range of statistical tools: first, they can control for a covariate using the general linear model, as in ANCOVA. Second, more complex controls might employ an instrument as an alternative measure of assignment to treatment condition. Third,

social scientists have begun to approximate the counterfactual by matching treatment and control subjects on propensity scores.

Even after employing statistical controls for covariates, treatment effects

*Please see Causal inference, page 4*

## Finding a “Statistical Collaborator” Made Easy!

When putting together a research project, investigators are often confronted with the problem of finding one or more collaborators with complementary expertise. CSTAT now has a new venue for finding such collaborators at MSU. Go to [www.cstat.msu.edu](http://www.cstat.msu.edu) and click on MSU Faculty with Statistical Expertise. Over 60 MSU faculty (and the number is growing) have listed their collaborative interests and expertise

on this site. The site can be searched by key words or department, or one can peruse an alphabetical list of all collaborators. For example, suppose you need a collaborator with expertise in spatial statistics. A search by keywords, “spatial statistics,” produces eight names with attendant information on email, website, and interests. If you would like to add your name to the collaborators’ list, contact us at [info@cstat.msu.edu](mailto:info@cstat.msu.edu).

## Upcoming Events

### SPRING 2007 WORKSHOPS

All workshops meet from 1-4 p.m. in B104 Wells Hall. Details and registration on [www.cstat.msu.edu/workshops](http://www.cstat.msu.edu/workshops).

No fees! They’re free. But you must register online to guarantee workshop materials and admittance. If you register and cannot attend, please let us know ASAP ([reg@cstat.msu.edu](mailto:reg@cstat.msu.edu)) so someone else can sign up.

#### Friday, January 19

##### ***Eigen Analysis***

Dr. Brian Maurer  
Dept. of Fisheries and Wildlife

#### Friday, January 26

##### ***Basic Data Analysis Using SPSS***

Dr. Sandra Herman  
Associate Director, CSTAT

#### Friday, February 9

##### ***Basic Data Analysis Using STATA***

Dr. Freda B. Lynn  
College of Education  
Dept. of Sociology

#### Friday, February 23

##### ***Measurement***

Dr. Mark Reckase  
Dept. of Counseling, Educational Psychology and Special Ed

#### Friday, March 16

##### ***Regression***

Dr. Vincent Melfi  
Dept. of Statistics and Probability

#### Friday, March 23

##### ***Intermediate Data Analysis Using SPSS***

Dr. Sandra Herman  
Associate Director, CSTAT

#### Friday, March 30

##### ***MatLab***

Dr. Dirk Colbry  
Cognitive Science  
Dept. of Computer Science and Engineering

#### Friday, April 13

##### ***Geostatistics***

Dr. Sasha Kravchenko  
Dept. of Crop and Soil Sciences

#### Friday, April 20

##### ***Propensity Scores***

Dr. Jeff Wooldridge  
Dept. of Economics

### *Causal inference, from page 3*

could still be attributed to differences between treatment and control groups. My work uses a robustness index to quantify how large the impact of an uncontrolled confounding variable would have to be to invalidate a statistical inference.

I conducted an analysis regarding the inference that attaining National Board certification affects the amount of help a teacher provides to others. Because teachers who are more inclined to be helpful may attain National Board certification, the inference may be invalid. But calculations show that the impact of an unmeasured confounder would have to be four times greater than the impact of the strongest covariate in our model to invalidate the inference. Thus, there is at least moderate robustness with respect to concerns about unmeasured confounding variables. These robustness indices are a form of sensitivity analysis, do not alter the initial inference, and quantify robustness of the inference. (For a spreadsheet and SAS software for calculating my indices of robustness, and PowerPoint and related papers, see <http://www.msu.edu/~kenfrank/research.htm#causal>).

Causal inference is one of the most rapidly growing areas for social science methodologists and many MSU faculty are working in the area. Barbara Schneider, my colleague within education, has written on issues of causal

## Meet the CSTAT Team

### **Ms. Yongfang Zhu**

Yongfang Zhu is a doctoral student in statistics under the supervision of Professor Sarat C. Dass and Professor Anil K. Jain. She joined CSTAT in the fall of 2006 and has enjoyed working with clients in various areas.

Her research interests are in the areas of spatial data analysis, Bayesian inference and computational methods, machine learning algorithms, and data mining. She worked as a research assistant in the Pattern Recognition and Image Processing Lab at MSU for two years, which gave her intensive experience in computation-based statistical applications. She has papers published in peer-reviewed journals and conference proceedings.



Yongfang Zhu

inference and scale-up, and she and I are teaching a seminar (CEP991B: section 3) in spring of 2007 on causal inference. For a starting point on the web, try <http://www.wjh.harvard.edu/~winship/cfa.html>.