A Note from the Director

Welcome back to the Methodology Center Perspective!

Our twelfth Summer Institute on Longitudinal Methods will be held this June. This year we have two excellent speakers. For the first one and one-half days Dr. Donald Hedeker from the University of Illinois, Chicago, will present on mixed models for longitudinal continuous and dichotomous data. Each attendee will receive a copy of Don’s new book, titled Longitudinal Data Analysis. For the second one and one-half days our own Joe Schafer will present on his very recent work on approaches to causal inference in quasi-experimental and nonexperimental research. As those of you who have attended our Summer Institute know, the goal is to impart both innovative methodological ideas and practical data analysis skills to attendees. Both speakers will weave real-world applications into their sessions, and attendees will get hands-on experience applying the lecture material to data.

Stephanie Lanza, grad student Bethany Bray, and I gave a week-long workshop on latent class and latent transition analysis at the University of Cologne, Germany in late February. The workshop was part of the 36th annual Spring Seminar run by the Central Archive for Empirical Social Research at the University of Cologne. This workshop gave us a chance to put our new SAS Procs, PROC LCA and the just-released PROC LTA (see details on Page 2), through their paces. Workshop attendees were kind enough to test a beta release of PROC LTA for us, and gave us useful feedback on both procedures. Attendees came from many parts of Germany as well as England, Finland, Romania, Belgium, and Croatia.

This semester we are pilot-testing a new teaching format in which the Center organizes a series of one-credit courses on focused methodological topics. Each course meets for a designated five-week portion of the semester and all meet on the same day and time, so it is possible for a Penn State student to enroll in some or all of the courses. This semester the topics are latent class and latent transition analysis, taught by Stephanie Lanza and Linda Collins, and missing data, taught by John Graham. In future semesters we hope to draw on the considerable methodological expertise at Penn State both in and outside the Center for these courses. We also may broadcast some of the courses over the Internet, so stay tuned!

Do you know a talented recent or soon-to-be PhD who is interested in a career focusing on research that advances both prevention science and quantitative methodology? If so, please encourage him or her to look into the possibility of a postdoctoral fellowship in the NIDA-funded Prevention and Methodology Training Program. This training program is a collaborative effort involving Penn State’s Prevention Research Center and The Methodology Center. Each postdoctoral fellow works with both Prevention Research Center and Methodology Center scientists. For details, look at our web site or contact Ed Smith at eas8@psu.edu.

Finally, I will be on sabbatical next year. During that time The Methodology Center will be in the capable hands of Stephanie Lanza. See you in a year!

Best regards,

Linda M. Collins, Ph.D.  
Director, The Methodology Center  
Professor, Human Development and Family Studies  
Professor, Statistics  
Penn State University

Successful Research Network on Adaptive Treatment Strategies

In 2004, Susan Murphy, H.E. Robbins Professor of Statistics and Research Professor at the Institute for Social Research at the University of Michigan, received an R21 from NIDA that was funded as part of the NIH Roadmap Initiatives. The impetus for the Methodological Challenges in Adaptive Treatment Strategies (MCATS) Network began years ago, as Dr. Murphy realized through her involvement as an Investigator in The Methodology Center that there was little statistical expertise concerning the development of adaptive treatment strategies for drug abuse and related behavior among methodologically-inclined health scientists. Further, she learned that methodological problems similar to those that arise in adaptive treatment strategies are confronted by scientists in other disciplines, including computer science and engineering. The network was proposed in order to bring this expertise to bear on issues in adaptive treatment strategies by bringing together health scientists with quantitative scientists from other disciplines.

continued on page 3
Joseph L. Schafer is an Associate Professor of Statistics at Penn State. He has served on the faculty of the Department of Statistics since 1992 and as an Investigator in The Methodology Center since 1996. He sees himself as an applications-oriented statistician with broad interests in developing, evaluating and disseminating statistical methodology to the social, behavioral and health sciences.

Joe is perhaps best known for his work on missing data and multiple imputation (MI). His 1997 book *Analysis of Incomplete Multivariate Data* (Chapman & Hall/CRC Press) and associated software have helped to bring MI into widespread use and spawned the development of new missing-data routines in SAS, S-PLUS and other statistical packages. He does not view MI as the optimal solution to every missing-data problem, but as a quick, handy tool with statistical properties that are "good enough" for many purposes, and certainly better than the ad hoc alternatives (e.g., case deletion) to which researchers often have turned in the past. He attributes the good performance of MI to its Bayesian origins. Bayesian statistics differs from the classical frequentist approach in that it treats known quantities (e.g., observed data) as fixed and unknown quantities (e.g., parameters and missing values) as random variables. When the latter are not of interest, they are eliminated from inferential statements by a process of averaging, which in MI is done through Monte Carlo simulation.

Although Joe would not consider himself to be a Bayesian statistician, he often uses Bayesian thinking to devise pragmatic solutions to thorny statistical problems. In the area of latent class modeling, for example, he has applied Bayesian ideas to help The Methodology Center create new SAS-based routines for latent-class (PROC LCA) and latent transition (PROC LTA) analysis. He is also applying Bayesian thinking in two areas that have been traditionally dominated by frequentist methods: the analysis of data from complex surveys, and causal inference from non-randomized observational studies. Frequentist approaches to these problems rely heavily on the individuals’ probabilities of sample selection; these probabilities are incorporated into analyses through the use of sampling weights and propensity scores. Bayesian thinking, however, suggests that the selection probabilities are irrelevant if the outcomes themselves can be modeled well. In the years ahead, he hopes to work with his collaborators and students to develop new methods for population modeling and causal analysis that are highly efficient and robust to misspecification of the population models.

Joe also maintains an interest in the development of statistical software. Statisticians are rarely trained to be good programmers, and programmers rarely have much experience in statistics and data analysis. Creating a good statistical application, however, usually requires knowledge of both. In the 2005 book *Developing Statistical Software in Fortran 95* (Springer), he and his coauthor, Dr. David Lemmon of The Methodology Center, present strategies, tools and advice for methodologists who want to create statistical software that will be robust and easy to maintain.

Joe’s easy communication style has brought him many opportunities to lecture at conferences and universities throughout the world. Over the last ten years, he has taught more than twenty professional short courses and workshops on a variety of topics related to missing data, multiple imputation, survey data and causal inference. He also has been successful in training and advising Ph.D. students; his former advisees now hold faculty positions at Duke University Medical Center, Michigan State, University of Massachusetts (Amherst), University of Connecticut and University of Illinois at Chicago.

**Featured Scientist**

**Joseph Schafer**

Joseph L. Schafer is an Associate Professor of Statistics at Penn State. He has served on the faculty of the Department of Statistics since 1992 and as an Investigator in The Methodology Center since 1996. He sees himself as an applications-oriented statistician with broad interests in developing, evaluating and disseminating statistical methodology to the social, behavioral and health sciences.

Joe is perhaps best known for his work on missing data and multiple imputation (MI). His 1997 book *Analysis of Incomplete Multivariate Data* (Chapman & Hall/CRC Press) and associated software have helped to bring MI into widespread use and spawned the development of new missing-data routines in SAS, S-PLUS and other statistical packages. He does not view MI as the optimal solution to every missing-data problem, but as a quick, handy tool with statistical properties that are "good enough" for many purposes, and certainly better than the ad hoc alternatives (e.g., case deletion) to which researchers often have turned in the past. He attributes the good performance of MI to its Bayesian origins. Bayesian statistics differs from the classical frequentist approach in that it treats known quantities (e.g., observed data) as fixed and unknown quantities (e.g., parameters and missing values) as random variables. When the latter are not of interest, they are eliminated from inferential statements by a process of averaging, which in MI is done through Monte Carlo simulation.

Although Joe would not consider himself to be a Bayesian statistician, he often uses Bayesian thinking to devise pragmatic solutions to thorny statistical problems. In the area of latent class modeling, for example, he has applied Bayesian ideas to help The Methodology Center create new SAS-based routines for latent-class (PROC LCA) and latent transition (PROC LTA) analysis. He is also applying Bayesian thinking in two areas that have been traditionally dominated by frequentist methods: the analysis of data from complex surveys, and causal inference from non-randomized observational studies. Frequentist approaches to these problems rely heavily on the individuals’ probabilities of sample selection; these probabilities are incorporated into analyses through the use of sampling weights and propensity scores. Bayesian thinking, however, suggests that the selection probabilities are irrelevant if the outcomes themselves can be modeled well. In the years ahead, he hopes to work with his collaborators and students to develop new methods for population modeling and causal analysis that are highly efficient and robust to misspecification of the population models.

Joe also maintains an interest in the development of statistical software. Statisticians are rarely trained to be good programmers, and programmers rarely have much experience in statistics and data analysis. Creating a good statistical application, however, usually requires knowledge of both. In the 2005 book *Developing Statistical Software in Fortran 95* (Springer), he and his coauthor, Dr. David Lemmon of The Methodology Center, present strategies, tools and advice for methodologists who want to create statistical software that will be robust and easy to maintain.

Joe’s easy communication style has brought him many opportunities to lecture at conferences and universities throughout the world. Over the last ten years, he has taught more than twenty professional short courses and workshops on a variety of topics related to missing data, multiple imputation, survey data and causal inference. He also has been successful in training and advising Ph.D. students; his former advisees now hold faculty positions at Duke University Medical Center, Michigan State, University of Massachusetts (Amherst), University of Connecticut and University of Illinois at Chicago.

**PROC LTA: New SAS Software for Latent Transition Analysis**

The Methodology Center is pleased to announce PROC LTA, a new SAS® procedure for latent transition analysis (LTA) that is available free of charge on the Internet. This procedure adds to the set of finite mixture models that can be fit using SAS. In 2006, The Methodology Center released PROC LCA for conducting latent class analysis in SAS (see Lanza, Collins, Lemmon, & Schafer, in press, for more information). LTA is a longitudinal extension of latent class analysis used to estimate latent status membership probabilities at Time 1, along with probabilities of transitions in latent status membership over time.

PROC LCA and PROC LTA are distributed as a suite of add-on procedures for SAS. These straightforward procedures allow users to pre-process data, fit a variety of latent class and latent transition models, and post-process the results without leaving the SAS environment. Features of PROC LTA include:

- Simple model specification
- Multiple-groups LTA and LTA with covariates
- Baseline-category multinomial logit model and binary logit model options for prediction
- Posterior probabilities saved to SAS data file
- Optional stabilizing prior to handle sparseness issues in estimation
- Prediction can be modeled using a baseline-category multinomial logit model or a binary logit model, adding to the set of questions that can be addressed in LTA.

A Bayesian stabilizing prior can be invoked with the STABILIZE statement when sparseness is an issue for parameter estimation.

*continued on page 3*
Specifically, the network's goals were to:

- jump-start the development of data analysis methods and the experimental data collection methods necessary for producing evidence-based adaptive treatment strategies;
- develop interdisciplinary collaboration necessary for formulating powerful adaptive treatment strategies;
- work together to identify major methodological challenges to the development of adaptive treatment strategies; and
- form interdisciplinary groups that will work to address the identified challenges.

Participants from a variety of disciplines attended the two major Network workshops. These include chemical engineering, industrial engineering, psychiatry, psychology, infectious disease medicine, family practice medicine, computer science, statistics, biostatistics, and health policy and administration.

As the funding period draws to a close, success of the network is apparent. For example, five grant proposals on adaptive treatment strategies have been submitted as a result of the Network meetings: an RO1 (by a psychiatrist, a computer scientist, an epidemiologist, a statistician, and an external medicine doctor), a PS0 (by two psychiatrists and a statistician), an R21 (by a quantitative psychologist and a chemical engineer), and a K25 (sent to NIDA by a chemical engineer).

In 2006, Dr. Murphy, along with David Oslin, A. J. Rush and Ji Zhu, contributed a white paper for the MCATS procedure for latent class analysis. Structural Equation Modeling.


Copyright © 2002-2003, SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

PROC LCA and PROC LTA, along with a user’s guide, are available on the Web at http://methodology.psu.edu.

Christopher Barto, a leading computer scientist, presented at the August, 2005 Joint Statistical Meetings on Dynamic Treatment Regimes; Andy Butch Tsiatis and Susan Murphy presented “Recent Innovations on Dynamic Treatment Regimes” in a symposium at the March 2007 Eastern North Atlantic Region of the International Biometrics Society. John Rush organized a workshop on adaptive treatment strategies at the 2006 meeting of the National Drug Control Unit (presenters included David Oslin, Linda Collins and Susan Murphy). Susan Murphy organized an Invited Session at the August, 2005 Joint Statistical Meetings on Dynamic Treatment Regimes; Andy Barto, a leading computer scientist, presented at this session.

Recently Susan Murphy and Alena Scott (a post-doctoral researcher in The Methodology Center) have teamed up with Dr. Janet Levy, a statistician in the NIDA Clinical Trials Network, to provide an evaluation of the sequential, multiple assignment randomized trial (SMART). The SMART experimental design can be used to develop adaptive treatment strategies. As part of their collaboration, they are developing sample size formulae that scientists can use to size a SMART trial. Watch the Methodology Center website for more information!

Several conference presentations grew out of Network collaborations. For example, Daniel Rivera, Tony Chen and Linda Collins teamed up to present a symposium at the 2006 Society for Prevention Research conference in San Antonio on potential applications of engineering methods in prevention science. Also, Marie Davidian, Butch Tsiatis and Susan Murphy presented “Recent Innovations on Dynamic Treatment Regimes” in a symposium at the March 2007 Eastern North Atlantic Region of the International Biometrics Society. John Rush organized a workshop on adaptive treatment strategies at the 2006 meeting of the National Drug Control Unit (presenters included David Oslin, Linda Collins and Susan Murphy). Susan Murphy organized an Invited Session at the August, 2005 Joint Statistical Meetings on Dynamic Treatment Regimes; Andy Barto, a leading computer scientist, presented at this session.

The NIH roadmap initiatives: bringing together scientists from many disciplines to take on methodological challenges faced in adaptive treatment strategies.

The interdisciplinary program will be attended by participants from applied mathematics, engineering, operations research, statistics, biostatistics, epidemiology, computer science, and business management science. This program is funded by the National Science Foundation through the Statistical and Applied Mathematical Sciences Institute (SAMSI), an institute that brings together statistical and applied mathematical sciences with other disciplinary sciences. See http://www.samsi.info/programs/2007adaptivetreatmentprogram.shtml for more information!

The NIH roadmap initiatives: bringing together scientists from many disciplines to take on methodological challenges faced in adaptive treatment strategies.

Visit us on the Web: http://methodology.psu.edu/
I often use fit criteria like AIC and BIC to choose between models. I know that they try to balance good fit with parsimony, but beyond that I’m not sure what exactly they mean. What are they really doing? Which is better? What does it mean if they disagree? — Signed, Adrift on the IC’s

Dear Adrift, as you know, AIC and BIC are both penalized-likelihood criteria. They are sometimes used for choosing best predictor subsets in regression and often used for comparing non-nested models, which ordinary statistical tests cannot do. The AIC or BIC for a model is usually written in the form \(-2\log L + kp\), where \(L\) is the likelihood function, \(p\) is the number of parameters in the model, and \(k = 2\) for AIC and \(\log(n)\) for BIC.

AIC is an estimate of a constant plus the relative distance between the unknown true likelihood function of the data and the fitted likelihood function of the model, so that a lower AIC means a model is considered to be closer to the truth. BIC is an estimate of a function of the posterior probability of a model being true, under a certain Bayesian setup, so that a lower BIC means that a model is considered to be more likely to be the true model. Both criteria are based on various assumptions and asymptotic approximations. Each, despite its heuristic usefulness, has therefore been criticized as having questionable validity for real-world data. But despite various subtle theoretical differences, their only difference in practice is the size of the penalty; BIC penalizes model complexity more heavily. The only way they should disagree is when AIC chooses a larger model than BIC.

AIC and BIC are both approximately correct according to a different goal and a different set of asymptotic assumptions. Both sets of assumptions have been criticized as unrealistic. Understanding the difference in their practical behavior is easiest if we consider the simple case of comparing two nested models. In such a case, several authors have pointed out that IC’s become equivalent to likelihood ratio tests with different alpha levels. Checking a chi-squared table, we see that AIC becomes like a significance test at alpha=.16, and BIC becomes like a significance test with alpha depending on sample size, e.g., .13 for \(n = 10\), .032 for \(n = 100\), .0086 for \(n = 1000\), .0024 for \(n = 10000\). Remember that power for any given alpha is increasing in \(n\). Thus, AIC always has a chance of choosing too big a model, regardless of \(n\). BIC has very little chance of choosing too big a model if \(n\) is sufficient, but it has a larger chance than AIC, for any given \(n\), of choosing too small a model.

So what’s the bottom line? In general, it might be best to use AIC and BIC together in model selection. For example, in selecting the number of latent classes in a model, if BIC points to a three-class model and AIC points to a five-class model, it makes sense to select from models with 3, 4, and 5 latent classes. AIC is better in situations when a false negative finding would be considered more misleading than a false positive, and BIC is better in situations where a false positive is as misleading as, or more misleading than, a false negative.