



**SACRAMENTO
STATISTICAL
ASSOCIATION**

PROGRAM

23rd Annual Institute on Research and Statistics

Wednesday, March 27, 2002
California State University, Sacramento

A local chapter of the American Statistical Association

<http://www.amstat.org/chapters/sacramento/>

President: Wolfgang Polonik (530) 752-7612; Vice-President: Farzaneh Tabnak 323-4536; Secretary: Julie Yee 379-3750;
Treasurer: Charles Chan 651-9080; ASA Representative: Linda Gage 327-0103 ext 2549; Past President: Hans-Georg Müller (530) 752-1629
Councilors: Scott Bartell (530) 752-3867; Matt Facer 323-7335; Kirsten Knutson 324-7967; Xueli Liu (530) 752-3873; Michael Quinn 445-6348;
Doraiswamy Ramachandran 278-6534; Lisette Walker 323-4593

REGISTRATION AND CONTINENTAL BREAKFAST

Redwood Room

8:00 to 8:45

PLENARY SESSION

Redwood Room

8:45 to 9:00

WELCOMING REMARKS

WOLFGANG POLONIK, President Sacramento Statistical Association
WALLACE ETTERBEEK, Chair Department of Mathematics, CSUS

9:00 to 9:45

KEYNOTE ADDRESS

TED GIBSON, Former Chief Economist California Department of Finance
What Kind of Recovery?

9:50 to 10:45

FEATURED SPEAKER

DAVID A. FREEDMAN, Department of Statistics, University of California, Berkeley
On the Likelihood of Improving the Accuracy of the Census Through Statistical Adjustment

I will sketch procedures for taking the census, making adjustments, and evaluating the results. Despite what you read in the newspapers, the census is remarkably accurate. Statistical adjustment is unlikely to improve on the census, because adjustment can easily put in more error than it takes out. Indeed, error rates in the adjustment turn out to be comparable to errors in the census. The data suggest a strong geographical pattern to such errors, even after controlling for demography-- which contradicts a basic premise of adjustment. The complex demographic controls built into the adjustment mechanism turn out to be counter-productive.

Proponents of adjustment have cited "loss function analysis" to compare the accuracy of the census and adjustment, generally to the advantage of the latter. However, the chosen analyses make assumptions that are highly stylized, and quite favorable to adjustment. With more realistic assumptions, loss function analysis is neutral, or favors the census.

At the heart of the adjustment mechanism, there is a large sample survey-- the post enumeration survey. The size of the survey cannot be justified. The adjustment process now consumes too large a share of the Census Bureau's scarce resources, which should be reallocated to other Bureau programs.

The paper is available at: <http://www.stat.berkeley.edu/~census/612.pdf>

California Suite

CHRIS DRAKE, Department of Statistics, University of California, Davis
*Determining the Distribution of the Time from Diagnosis to Loss to Follow-up of
 Cancer Cases Reported to the California Cancer Registry*

Authors: James Beaumont, Chris Drake, Julie Smith

The California Cancer Registry serves as the central state registry for all cancer cases diagnosed in California. The registry is divided into ten sub-regions, which gather information on each cancer case and report it to the central registry. The information in the registry is used to study population trends, clustering of cases and environmental influences on the development of cancer. The California Health and Safety Code mandates hospitals, physicians and laboratories report all cancers except basal and squamous cell skin cancers and in situ cancer of the cervix.

Case-control studies are commonly employed in cancer research to establish associations between risk factors and a specific type of cancer. Case-control studies establish exposure retrospectively, after onset and diagnosis of the disease and interviews are often conducted with cases or next of kin to obtain the necessary information. Therefore, it is necessary to be able to locate cases and/or next of kin. To minimize the risk of loss to follow up, many studies use rapid case ascertainment (RCA). Controls are often obtained through random-digit dialing (RDD). Random digit dialing is population based for obtaining controls. It differentially excludes people without phones, such as migrant workers. Therefore some studies choose to select controls from other patients at the institutions the cases are selected from, sometimes restricting the eligible controls to other cancer cases.

To plan a case-control study carefully, it is desirable to know how soon cancer cases are lost to follow-up, either through death or leaving no forwarding address. We undertook a study obtaining all cancer cases reported between January 1, 1996 and December 31, 1997 from the California Cancer Registry regions 2, 3 and 6, which comprise the Central Valley of California. We studied the distribution of the report time (defined as the time from diagnosis to availability in the central registry) and factors related to this report time. We furthermore randomly selected 8000 cases from the 14 most common cancers and established a place and date of last known residence using Vital Statistics and DMV. For a subset of 5000 we also obtained credit records to verify place and date of residence. This talk will present some of the major findings and discuss the possibility of carrying out registry based studies or using data from the registry to control for selection and other biases.

Delta Suite

HONGZHE LI, Medical School, University of California, Davis
*Multivariate Survival Models Induced by Genetic Frailties
 with Application to Linkage Analysis*

Many complex human diseases are due to multiple disease genes and both genetic and environmental risk factors. These diseases often also show variable age of disease onset. In order to incorporate both covariates and age of onset information into genetic linkage analysis, we define an additive genetic gamma frailty model constructed based on the inheritance vectors. Based on this model, we derive the joint survival and density functions for age of onset data of a sibship and propose a retrospective likelihood ratio test for linkage using sibship data. This test is an allele-sharing-based test and does not require specification of genetic models or the penetrance functions. This new approach can incorporate both affected and unaffected sibs, environmental covariates and age of onset or age at censoring information, and therefore provides a practical solution to mapping genes for complex diseases with variable age of onset. Simulation studies indicate that the proposed method has correct type I error rate and performs better than the commonly used allele sharing based methods for linkage analysis, especially when the population disease rate is high. We demonstrate the methods using data sets of affected sib pairs of prostate cancer and type I diabetes. We are currently extending the methods to linkage analysis for two-locus disease models and to test of genetic association in the presence of linkage.

Forest Suite SUE GELLER, Center for Image Processing and Integrated Computing, University of California, Davis
Variance Stabilization and Normalization in Microarray Data

Authors: Sue Geller, Jeff Gregg, Paul Hagerman, David Rocke

Microarray data can be thought of as an $n \times m$ array, in which n is the number of genes, m is the number of slides or replicates, and $n \gg m$. It is the last fact, $n \gg m$, that makes the analysis of microarray data challenging. Also, the size of the arrays, e.g., 12625×4 or $252,500 \times 4$, raise computational issues. Many common methods of analyzing microarray data use parametric techniques, complete with the assumption that the variance is constant across all expression levels (i.e., without regard to the size of the data value). It was determined by Rocke and Durbin that the variance was not constant for spotted microarray data and instead

A planned review of basic sampling theory transitions into a formal level of stratified sampling. The theory used in formula derivations will be accompanied by an application involving property taxes.

CONCURRENT SESSIONS III

1:35 to 2:00

California Suite

LOIS LOWE, Independent Program Evaluation Consultant
Using Qualitative Methods for Program Evaluation

Over the past several years, the California Department of Corrections and substance abuse program providers have used qualitative methods to augment program evaluation information. By asking program participants one or two open-ended questions, evaluators have been able to link specific program learning to subsequent outcome in the participant's life. An example of a question is "Has something happened in your life where you used what you learned while in the treatment program?" "What happened?" Follow-up studies of adult men and women show that former treatment participants actually used what they learned while in the in-prison treatment programs.

Delta Suite

ALICE VAN OMMEREN, California Department of Mental Health
Receiver Operating Characteristic (ROC) Analysis to Evaluate the Accuracy of Violence Prediction

The management of offenders within the mental health and criminal justice system is influenced by an offender's perceived risk of recidivism. The restriction, supervision, sentencing, as well as the commitment of violent offenders is decided by the ability of clinicians to make predictions of future violent behavior. The validity of these decisions is important to mental health clinicians making these recommendations, especially those suggesting commitment of sex offenders as sexually violent predators. Receiver Operating Characteristics (ROC) is a statistical method frequently used for evaluating decisions and has only recently been introduced into the area of violence risk assessment.

ROC statistics have been recommended to evaluate the accuracy of recidivism predictions because they are easily interpreted, not influenced by bases rates and independent of selection ratio. ROC's allow for the comparison of various thresholds on prediction measures, an overall index of accuracy for all thresholds, the identification of the optimal threshold and the comparison of two or more measures. The presentation introduces the basic ROC concepts and its usefulness for the evaluation of instruments used for the prediction of violent recidivism of sex offenders. Statistical applications, list of references and additional resources are also briefly discussed.

Forest Suite

DAMLA SENTURK, Statistics Department, University of California, Davis
Varying Coefficient Models and Mathematical Coupling

Varying Coefficient Models are an extension of regression and generalized regression models where the coefficients are allowed to vary as a smooth function of a third variable. A common application is to longitudinal data where the third variable becomes time. Some recent techniques for estimating the coefficient will be reviewed, including the application of smoothing and of binning methods.

A new application is to mathematical coupling. In the coupling situation we consider, predictors and response in the regression model are not observable directly but only a modified form is observed where the variables have been multiplied with an unknown function of a third variable. This is an instance of mathematical coupling as it usually introduces a dependence between response and predictors.

We demonstrate how coupling can be handled by fitting a varying coefficient model to the data. It is shown how this allows estimation of the original regression coefficients. This approach was used for regressing albumin synthesis on a set of predictors for 64 subjects. Coupling is an issue as the variables measured are all known to be influenced by body mass. Simulation studies demonstrate the efficacy of this method.

CONCURRENT SESSIONS IV

2:10 to 2:35

California Suite

WILLARD HOM, California Community Colleges
Grouping Colleges by Changes in Enrollment Volume

This talk reports an effort to find a typology for the enrollment change in California's community colleges. Such a typology can help researchers and planners by exploring the various types of enrollment shifts that have occurred in the state's community colleges

since 1992. This information could aid planners who must search for explanations of their enrollment trends and/or who must do enrollment projections. Analysts who must handle an array of longitudinal data for any type of population may also benefit from the presentation.

The analysis in this paper will use longitudinal enrollment data in the Chancellor's Office MIS. Various statistical tools will allow us to investigate the (1) slope of the change; (2) raw variability of the change; (3) relative variability of the change; and (4) correlation of the change per college with change in the state. Cluster analysis will provide a method for finding a typology of the colleges according to these four factors.

Delta Suite

DARYL METZ, California Energy Commission
Estimation of the Effects of Daylight Saving Time on California Electricity Use

Authors: Adrienne Kandel, Daryl Metz, Newton Wai

Daylight Saving Time (DST) is the advancement of standard time by one hour so that the solar day more closely corresponds to our normal activities. Historically, it has been assumed that electricity can be saved with DST because people have an extra hour of daylight in the evening and thereby use less electric lighting.

The presentation will report on the estimation a statistical model of aggregate hourly electric use for California in order to evaluate the claims of energy conservation from DST, proposals to extend the observance of DST and the effects of Summer-season Double Daylight Saving Time (DDST). The model is a system of 24 linear equations, one for each of the hours of the day that relates the level of electric use to the time of day: whether or not it is a workday, the hourly weather conditions, whether there is sunlight or twilight present at that time, along with an economic demographic variable and the interactions of these variables. This approach permits the estimation of the average change in electric use resulting from advancing daily schedules relative to the sun and daily weather patterns while controlling for the changing seasonal weather, length of days, holidays, and economic conditions.

We concluded that both Winter DST and Summer-season DDST would probably save marginal amounts of electricity - around 3400 MegaWatt hours (MWh) a day in winter (one half of one percent of winter electricity use) and around 1500 MWh a day during the summer season (one fifth of one percent of summer-season use). Winter DST would cut winter peak electricity use by around 1100 MW on average, or 3.4 percent. Summer Double DST would cause a smaller (220MW) and more uncertain drop in peak, but it could still save hundreds of millions of dollars because it would shift electricity use to low demand (cheaper) morning hours and decrease electricity use during higher demand hours.

Forest Suite

KIRSTEN KNUTSON, California Department of Health Services
Estimating the Low-Income, Underinsured/Uninsured Population in California Counties Eligible for Health Services

Authors: Kirsten Knutson, Weihong Zhang,

The California Department of Health Services, Cancer Detection Section (CDS) administers breast and cervical cancer screening programs which provide services to eligible women in "Partnership" areas, which are aggregates of California counties. Describing the low income, uninsured/underinsured eligible population by Partnership is problematic, however, because current county-level data containing income and health insurance status is not available.

To estimate the eligible population per Partnership by demographic characteristics, small area estimation methodology was used. County data on factors known to be associated with eligibility status were combined in a generalized linear mixed effect model with state data describing eligibility status. A random effect variable included in the model represents the area cluster variation that cannot be explained by the regression variables. Bootstrap methodology was applied to the model to calculate the variance.

This paper demonstrates the application of statistical methods to develop a quantitative description of a population not commonly represented by data. Though this is work in progress, methods of evaluating the results will be discussed.

CONCURRENT SESSIONS V	2:45 to 3:30
------------------------------	---------------------

California Suite

RAHMAN AZARI, Department of Statistics, University of California, Davis
Categorical Data Analysis: A Short Review

Categorical data analysis is widely used in many areas of research such as health services, transportation, and environmental studies. A short review of the most popular methods of modeling categorical data will be presented. The techniques will include log-linear

models, logistic, Poisson and negative binomial regressions, as well as, logit and probit models. The versatility of these methods will be discussed by applying them to a wide variety of data from different fields. Random and mixed effects models also will be discussed. Special attention will be given to the most widely available software for categorical data analysis.

Delta Suite

DAVID HEISER, National Technical Systems, Sacramento Division
EXCEL for Statistics, the Issues, Applications, Faults, Fixes and Workarounds

Microsoft EXCEL is extensively used in teaching introductory statistics in America and Europe. It is also extensively used in business as a spreadsheet program, using the built-in statistical functions and the Data Analysis Tool Kit. There have been many criticisms on the use of EXCE for statistics. I present some of these criticisms and the related issues in teaching and in applications. This includes criticisms on accuracies (McCullough and others), graphics (Cryer), the paired t-test (STERN School), statistical distributions (Knusel), faulty algorithms (Simon), missing data and other general limitations. I describe the complete NIST StRD test for accuracy on EXCEL. I describe fixes and methods (not dependent on Macro's) to overcome these faults and limitations. I discuss ways to test for the accuracy of results." Note that EXCEL is used in the Introductory Statistics course at Sac State, being taught by Business School professors.

Forest Suite

PAUL T. MELEVIN, Audit and Evaluation Division, State of California Employment Development Department
Personal Delivery of Mail Questionnaires for Household Surveys: a Test of Four Retrieval Methods

Authors: Paul T. Melevin, Don A. Dillman, Rodney Baxter, C. Ellen Lamiman

This paper reports the results of an experiment aimed at overcoming no-coverage error, the biggest limitation of mail surveys of the general public. The effectiveness of four different procedures and three methods of delivery were tested for retrieving 20 page mail questionnaires delivered by face-to-face interviewers to a statewide area probability sample of households. Results indicate a significant difference in the rate of response obtained from those who were provided the post-incentive, the immediate follow-up or both, versus those who were provided no special treatment. The response rates also differed on the basis of delivery method. Upon controlling for delivery method, a difference due to the four experimental treatments was also found among those respondents who had questionnaires personally delivered to them. The overall response rates attained are too low to justify widespread use of the exact methods implemented here, but with slight improvements, for which research directions are identified, they may be improved to acceptable levels.

CLOSING SESSION Redwood Room 3:40 to 4:20

FEATURED SPEAKER

CLYDE TUCKER, Bureau of Labor Statistics
Implementing the 1997 Standards for the Collection and Reporting of Data on Race and Ethnicity

This presentation will begin with a very brief review of the major changes in the new standards and some of the possible effects on demographic series. The second part of the talk will give an overview of the Federal efforts to implement the standards. The last part will focus on research on ways of bridging from the old to the new series for trend analysis.

NETWORKING AND REFRESHMENT Redwood Room 4:20 to 4:45
