



**SACRAMENTO
STATISTICAL
ASSOCIATION**

A local chapter of the American Statistical Association

**PROGRAM
Institute on Research and Statistics**

Wednesday, March 23, 2005
California State University, Sacramento

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

President: Matt Facer 449-5835; Vice-President: Linda Gage 327-0103 x2549, Secretary: Kathleen Gallagher 552-9642; Treasurer: Charles Chan 552-9694; ASA Representative: Linda Gage 327-0103 x2549; Past President: Julie Yee 379-3750; Councilors: Tam Clay 669-3573; Shannon Conroy 449-5280; Christiana Drake (530) 752-8170; Kirsten Knutson 449-5305; Lois Lowe 722-3310; Gloria Robertson 654-1837; Lisette Walker 449-5330

All events will take place at the California State University Union

Registration and Continental Breakfast	Redwood Room	8:00-9:00
---	---------------------	------------------

Morning Plenary Sessions	Redwood Room	9:00-10:55
---------------------------------	---------------------	-------------------

9:00-9:15

Welcoming Remarks

Matt Facer, President, Sacramento Statistical Association
Doraiswamy Ramachandran, Chair, Department of Mathematics and Statistics, CSUS

9:15-10:05

Special Guest Speaker

Fritz Scheuren, President, American Statistical Association
Role of Statistical Science in U.S. Presidential Elections

10:05-10:55

Featured Speaker

Howard Roth, Chief Economist, California Department of Finance
The California Economy: Is This as Good as it Gets?

I. Concurrent Sessions	11:00-11:55
-------------------------------	--------------------

Delta Suite

(11:00-11:50)

Christiana Drake, Department of Statistics, University of California, Davis
Some Thoughts on the Relative Survival Rate

The Relative Survival Rate (RSR) is defined as the ratio of the observed survival in a population defined by a disease such as cancer, to the survival rate expected if that population with the same values of other factors related to survival but free of disease (Ederer et al). Hakulinen uses a similar definition but also shows that under the assumption of independence of other causes of death and death from the disease under study, the RSR can be shown to equal the net (or cause-specific) survival rate. Thus, one would expect the relative survival to be a non-increasing function of time.

However, all authors who have written on relative survival acknowledge that the estimated relative survival can exceed one. There are various reasons for this phenomenon. Among these reasons are that the population rates used to calculate expected survival may not adequately reflect the true survival that would be observed in the population of interest if disease were absent. Hakulinen shows that one main reason lies in the age structure and withdrawal patterns of the disease and reference groups. This talk will elaborate on some of these issues and also present an application of relative survival.

Orchard Suite

(11:00-11:25)

Cyndee Singer, California Department of Finance, Demographic Research Unit, State Data Center
American Community Survey: An Overview

The American Community Survey (ACS) replaces the decennial census long-form questionnaire gathering the same data but on an annual basis. Testing began in 1996 with full implementation this year with the exception of group quarters and Puerto Rico. Data are collected throughout the year from a sample of about 3 million households in the United States with annual results tabulated and scheduled for release the summer following the data collection year. Because data are collected all year long, a permanent, well-trained staff should mean improved reliability. The ACS will provide more timely data than the census and can respond to changing needs while still providing the in-depth data usually only available decennially.

(11:30-11:55)

Martha Jones, California Research Bureau, California State Library

*Estimating Use Tax Revenue Losses from Remote Sales Using Census E-Stats and Mail Order Data:
A Comparison of Methodologies*

Use tax applies when a person or business in California purchases tangible merchandise from a retailer outside of this state that will be used, consumed, given away, or stored in this state. The use tax applies to remote sales, which are goods bought over the Internet or through a mail-order catalog and shipped to California from an out-of-state address by a company that does not have a physical presence (nexus) in California. The California State Board of Equalization encourages purchasers to voluntarily register and pay their use tax obligations. Attempts by various states to require remote sellers with no physical presence to collect and remit use tax on merchandise sold to a state's residents have been restricted by U.S. Supreme Court decisions.

State governments are losing revenue due to the inability to collect use tax from remote sales (both e-commerce and mail-order). States are worried that as e-commerce grows, their revenue losses will grow as well. After reviewing trends in remote sales, this presentation will focus on methodologies for estimating state use tax losses from remote sales using Census statistics for e-commerce. Estimates for use tax losses vary widely across studies depending on the methodology used. For California, losses from e-commerce alone ranged from an estimated \$350 million to \$2.2 billion in 2003. This presentation will analyze various estimates and explain the different assumptions used.

Lunch Buffet

Redwood Room

12:00-1:00

II. Concurrent Sessions

1:05-2:00

Delta Suite

(1:05-1:30)

Julie Yee, U.S. Geological Survey, Western Ecological Research Center

Bayesian Analyses of Double-Observer Surveys for Handling Unknown Identification Error

Many research disciplines are interested in the proportion of a population having a certain attribute. However, the methods that identify the attribute are often subject to error. For example, a health researcher could want the proportion with a certain disease, yet medical screening tests can be subject to unknown rates of false positives or false negatives. An ecologist could want the proportion of sites having a rare species present, yet moving species might be undetected even when present. An avian researcher could want the proportion of wild geese belonging to a particular subspecies. These problems (or their solutions) are known by different terms in their respective fields (i.e., multiple test screening, presence-absence, and double-observer, respectively), yet they utilize similar statistical concepts and solutions tend to be re-invented for each field. This talk compares solution approaches and illustrates how a published Bayesian solution to the health problem can be extended to solve problems in other research areas.

(1:35-2:00)

Julie Yee, U.S. Geological Survey, Western Ecological Research Center

WinBUGS and PRESENCE Software Demonstration

The studies described in the previous talk can be conducted using freeware. WinBUGS software was developed as part of the BUGS (Bayesian inference Using Gibbs Sampling) project (<http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml>). Program PRESENCE was developed specifically for presence-absence data and does not use Bayesian methods (<http://www.proteus.co.nz/home.html>).

Orchard Suite

(1:05-1:30)

Ann Webb, Public Health Institute, Survey Research Group

20 Years of California Behavioral Risk Factor Survey Data: How Far have we Come?

The California Behavioral Risk Factor Survey (BRFS) is an on-going telephone survey of randomly selected adults, which collects information on a wide variety of health-related behaviors. The survey has been conducted since 1984 by the California Department of Health Services in collaboration with the Centers for Disease Control and Prevention and the Public Health Institute. The survey includes a number of questions on health related topics.

This presentation will review 20 years of the BRFS trends on a variety of health-related topics such as overweight and obesity, tobacco use, pap testing, mammography testing, sigmoidoscopy testing, gun safety and seatbelt use. In addition, these results will be compared with the US Department of Health and Human Services Healthy People 2010 Objectives where appropriate.

While California has made great strides in some areas such as cigarette smoking prevalence and seatbelt use, other areas such as overweight and obesity still need much improvement.

(1:35-2:00)

Daniel Smith, California Dept of Health Services, Environmental Health Investigations Branch
Overlapping Confidence Intervals are Not a Statistical Test

A common practice in epidemiology is to graph or tabulate several disease rates and their confidence intervals. Some authors go on to assess statistically significant differences among the rates by noting whether the confidence intervals overlap. Unfortunately, such a comparison is not valid, as even significantly different rates can have overlapping confidence intervals.

Classical ANOVA (Analysis of Variance) has an approach for comparing several means this way, known as LSD, or "Least Significant Difference." I give simple analogous procedures for person-time disease rates and proportions to construct "error bars" for several rates that have the property of illustrating significant differences when they don't overlap, and non-significant differences when they do. Such an approach can be useful for graphing rates across several areas and showing differences and similarities.

Special Session. An Invited Tutorial Sponsored by the Sacramento Valley SAS Users Group
--

2:05-4:20

Redwood Room

(2:05-4:20)

Anthony An, SAS Institute
Analyzing Sample Survey Data Using SAS

Survey statistical analysis calls for sophisticated computational tools. In order to make statistically valid inferences about a study population based on probability-based sample surveys, statisticians require computational support that takes the features of complex sample designs into account, such as stratification, clustering, and unequal weighting. A suite of SAS procedures has been developed that incorporates the design of a survey sample into its analysis.

This short course will illustrate how to use these procedures to:

- Select probability samples according to various designs using PROC SURVEYSELECT
- Compute descriptive statistics (such as means, totals, ratios, domain means, and their standard errors) using PROC SURVEYMEANS
- Fit linear regression models, perform hypothesis tests, and compute estimates using PROC SURVEYREG
- Cross-tabulate survey data using PROC SURVEYFREQ
- Perform logistic regression analysis using PROC SURVEYLOGISTIC for categorical responses in survey data

III. Concurrent Sessions

2:05-3:00

Delta Suite

(2:05-2:55)

Jimin Ding, Department of Statistics, University of California, Davis
Joint Modeling of Longitudinal and Survival Data

This talk intends to provide an overview for an emerging area in biostatistics. Often in medical studies longitudinal covariates, such as blood pressure or CD4 counts in AIDS study, are collected together with survival time for each subject. Relationship between a failure time process and some longitudinal covariates is of key interest and so is the understanding of the pattern of longitudinal process to learn more about health status of patients, or to get some insight into the progression of disease. However, difficulty arises in situations when the longitudinal covariates are collected only periodically or measured with error as most existing survival models, such as the Cox proportional hazards model and accelerated failure time model, require knowledge of the entire history of the longitudinal covariates. An ad hoc and common practice is to employ the Last Value Carry Forward method available in standard software. It is now well known that such a method may lead to serious bias, and several remedial measures have been proposed including a two-stage approach, regression calibration, conditional score approach among others. These remedies work to some extent but are not efficient. A further complication is a bias issue which occurs when the observation of the longitudinal process was interrupted when a patient dies. Such a truncation causes informative missing data and needs to be handled differently. Joint modeling of longitudinal and survival data provide an effective way to overcome these problems. We will introduce such an approach, illuminate on several key issues including some difficulties, and provide some answers to overcome these difficulties.

Orchard Suite
(2:05-2:30)

Gloria J. Robertson, Office of Statewide Health Planning and Development
Looking at Statistical Illustrations' in the Evaluation of a Pilot Project

This talk looks at the Postsurgical Recovery Care Demonstration Project. The purpose of the project was to evaluate the safety and effectiveness of accommodating patients recovering from surgery procedures in a health facility setting other than the required general acute care hospital. The talk will begin with a brief introduction and background information related to the establishment of the program. Then, it will give the program data design, illustrations, and non-tested statistical ending results. This talk has been described by the speaker to be at the novice level.

(2:35-3:00)

Lois Lowe, Ph.D.
Program Evaluation of the Walden House In-Prison Substance Abuse Treatment Program
California Department of Corrections Substance Abuse Treatment Facility

Walden House, a private non-profit organization, has provided in-prison substance abuse treatment services to felon male inmates at the California Department of Corrections (CDC) Substance Abuse Treatment Facility, Corcoran, California since 1997. At any point in time, Walden House provides treatment for over 700 inmates. This evaluation study was based on data that is routinely collected by Walden House at intake and during treatment. These data were merged with CDC's return-to-custody data and analyzed.

A total of 899 Inmates entering treatment from 1997 through 1999 were tracked through February 2001 for at least 12-months following release from prison. Basic program and inmate characteristics data were correlated with return-to-custody data. Analyzed variables included participant personal characteristics, length of time in treatment, and psycho-social indicators of mental health status and alcohol/drug use/dependence. Lower return-to-custody rates were found for participants who were older, completed high school, had a first arrest after age 18 and spent a longer time in post-parole follow-on care. As expected, those who reported higher levels of mental health, alcohol/drug problems were returned to custody noticeably more frequently. Detailed findings are contained in the 2002 final report.

Afternoon Refreshments

3:00-3:15

IV. Concurrent Sessions

3:20-4:15

Redwood Room

(3:20-4:20)

Anthony An, SAS Institute
Analyzing Sample Survey Data Using SAS (Continued)

Delta Suite

(3:20-3:45)

Enoch Haga, Ph.D.
How to Please an Editor, Print or Electronic Media

Writing is an important skill either in-house or for publication in the print or electronic media. While we all can write, writing effectively can make you more successful at work as well as help gain immediate acceptance and publication of your work. This presentation points out some of the more common writing errors that hinder acceptance, as well as explains self-editing for technical accuracy, language, and appearance. Attention to these details will please an editor and secure publication.

(3:50-4:15)

Tim Gaffney, Ph.D., California Department of Education
On the Factor Structure of Standardized Educational Achievement Tests

This research analyzed the factor structure at both the item and parcel level of California's norm and criterion referenced standardized educational achievement tests (SEAT) used in that state's high-stakes educational accountability assessments. It will be shown through full information factor analysis (e.g., NOHARM and Testfact) that at the item level, SEATs are highly unidimensional (i.e., they appear to tap a general test-taking ability) even when items representing broad content areas such as English, science, mathematics, and history are analyzed simultaneously as a single measure. These item-level factors also account for a relatively small proportion (1/4 to 1/3) of the variance. It will also be shown that, when items representing each content domain are combined into parcels, a richer bi-factor structure emerges that accounts for a larger portion (2/3) of the variance, in which the general and combined specific factors roughly split the common factor variance. The meaning of this lack of correspondence between the item and parcel level structures will be discussed (e.g., well articulated structure at the parcel level does not necessarily imply correspondingly clear structure at the item level), as well as the diagnostic and remedial implications of these tests' factor structure for educators and test forms developers.

Orchard Suite
(3:20-3:45)

Willard Hom, Chancellor's Office, California Community Colleges
Pooling Data from a Multi-site Study vs. Meta-analysis

This talk reports on an analysis of survey data from samples of students at ten community colleges to test a model of predictors for student plans to terminate enrollment before completing their desired academic agenda. Some analysts would simply pool the survey responses and treat the data as a single study of a single population. Other analysts would instead employ the meta-analysis approach, which requires much more work. This second approach treats each site as a separate study warranting a separate model estimation process. Presumably, the more modern approach of meta-analysis would provide a more accurate estimate of an overall effect size (i.e., predictors of educational plans) than the estimate from simple pooling. The presentation covers how the prediction model differs according to the method used.

The content will have an intermediate level of statistical difficulty but even novice researchers may benefit from the discussion of this topic. The talk will focus upon the differences between the two methods rather than upon the mechanics of estimating the model (through logistic regression) or upon the data sources (student surveys).

(3:50-4:15)

David A. Heiser
Tests of Significance: A Review of Some Emerging Views and New Method, in Economics, Psychology, Education Research and Medical Treatments.

Describes some of the current misconceptions and problems with the Neyman-Pearson hypothesis test (NHST). Errors in published papers reporting on NHST significance tests and the wrong conclusions made, are reaching a critical level in economics, psychology, education research, medicine and biology publications. These errors and faults arise from unthinking mechanical approaches (software programs), bad textbooks, faulty statistics teaching, weak journal editors, biased peer reviewers and poor publication guidelines. I describe some of the newer modified NHST tests, confidence interval based tests and Bayesian methods. I discuss the issues of practical significance versus statistical significance and the emerging use of Cohen's effect sizes as a sole measure of significance. The greater use of models, which require a close-fit criteria/test as a test of significance is introducing basic problems of just what is the "correct model".

Networking, Socializing and Raffle (door prizes provided by the Sacramento Valley SAS Users Group)

4:20-5:00