A New Year

2001

Did Your Vote Count?

www.csss.washington.edu
The CSSS Seed Grants Program is a small grants program designed to promote interdisciplinary research among UW social scientists and statisticians. The goal is to stimulate scholarly initiatives by encouraging faculty to explore new directions in research and scholarship that contributes to the development of statistical methods for social scientific problems. The seed grants are used to pursue pilot studies, feasibility studies, or preliminary research that initiates a larger project that has high promise of extramural funding, which would be administered through the Center. Awards are in the range of $10,000-20,000 and typically include one month of summer salary for the principal investigator and one quarter salary for a research assistant. The length of the project period is usually about one calendar year.

This year, we funded three outstanding proposals, which include faculty from Economics, Sociology, Statistics/CSSS, and RAND. Next year, the deadline for the Program will be around the first of November 2001. We encourage submissions from faculty who are interested in pursuing social scientific research that also contributes to the development of statistical methods in the social sciences. Although we seek interdisciplinary collaborations – particularly between social scientists and statisticians – such collaboration is not a formal requirement. For more information, see the CSSS web site, www.csss.washington.edu, and go to Seed Grants page. For specific questions, contact Ross L. Matsueda, CSSS Associate Director (matsueda@u.washington.edu and 616-2432).

The following are descriptions of funded Seed Grant proposals for 2000-01.

**Paul C. LePore, Sociology & Greg Ridgeway, Statistics Department, RAND Corporation**

**Predicting Dropouts: Boosting Algorithms and the Identification of Academically At-Risk Youth.** Although the percentage of youngsters graduating from high school has increased since the 1970s, a significant proportion of America’s high school youth continue to drop out of school. Numerous studies have identified important individual- and school-level factors associated with dropping out. But the models used to identify youth at risk of school failure have largely proved inadequate. This study applies the method of “boosting”—a new class of learning algorithms more precisely known as “gradient machines”—to improve the accuracy of classification models predicting the likelihood of high school failure. While relatively new to the field of statistics, boosting has become a popular tool in the machine learning community and shows promise in the area of applied problems. This study will apply boosting algorithms to data from the National Education and Longitudinal Study of 1988, a nationally representative sample of adolescents within schools currently in its fourth wave. NELS provides data on students’ attitudes and behaviors as well as data on family background, and other demographics. By applying boosting to the problem of school dropout, this study seeks to help teachers and administrators identify students most susceptible to school failure, and enable them to direct resources for dropout prevention to students in greatest need.

**Elaina Rose, Economics**

**The Determinants of Union Status and Partner Choice.** The role of marriage has undergone profound change in recent decades. Divorce and cohabitation have become commonplace, and age at first marriage has increased for both men and women. These shifts in marriage patterns have been accompanied by changes in assortative mating—that is, the types of partners individuals choose when they do form unions. This study will contribute to the literature on marriage and assortative mating in four ways. First, it will estimate assortative mating based on multiple outcomes. Previous research has examined assortative mating on education, testing the extent to which mating is homogamous (partners select partners with similar education), and examining trends in such patterns. This study will examine assortative mating on income and other outcomes. Second, it will develop a methodology for estimating the determinants of multiple union types, generalizing the binary outcome for marriage to an ordered polytomous outcome for cohabiting, married, or neither. Third, it will allow for non-linearity and non-monotonicity in estimating the relationship between individual characteristics and union status on the one hand, and degree of assortative mating on the other. Fourth, it will compare assortative mating patterns for cohabiting and married couples and for early and later cohorts. Lam’s model predicts that cohabiting couples will be more similar than married couples, and that assortative mating has increased over time. Fifth, it will estimate a model of partner choice that allows for selection into marital or union status. Because married (and cohabiting) couples are not a random sample of the population, estimates of one partner’s characteristics on those of the other is subject to sample selection bias. This can be addressed within the selection framework proposed.

**Katherine W. Stovel, Sociology & Peter Hoff, Statistics and CSSS**

**Hearing about a Job: A Model of Differential Information Flow and Job Matching.** While sociologists and economists recognize that employment is
fundamentally a problem of matching workers with jobs, most empirical research on occupations and labor markets has focused on either the structure of supply of workers or the structure of demand for workers. Others have recognized job searching as the mechanism by which individual jobs and workers find each other, but there are no labor market models that formally incorporate different recruitment strategies into the matching process. The goal of this study is to develop a general two-sided employment model in which the information available to workers and employers about each other may depend on characteristics of workers and of jobs (vacancies). A key innovation is to allow for the possibility that a worker may be qualified for a job, but has no information about a vacancy, and thus has reduced access to employment. This study will investigate the properties of the model using two approaches: simulation, and estimation of two-sided matching models. By using the general model to simulate a labor market, the investigators will be able to study the dynamic nature of job matching given structural constraints. They will focus particularly on the system-level properties (measured in terms such as turnover rates, over- and under-qualification of workers, duration of vacancies) resulting from various recruitment strategies. In the data analysis setting, they will develop methods for estimating parameters in the model given job market data, and thus will be able to make inferences for markets in which it is known that different workers have access to different information about vacancies. This will extend the current state-of-the-art methods available for two-sided matching data, which assume perfect access to information.

The ongoing project, which was initiated by the award of a seed grant from CSSS, has already made some comparisons between methods of estimation and has devised some new rigorous statistical procedures for goodness of fit, but much remains to be done. Network analysis focuses on the relationships between specified individuals or entities and the patterns that exist in these relationships. It may also enable predictions about future behavior of the network to be made, subject perhaps to various forms of intervention. Thus, network analysis has an enormous number of potential applications, ranging from the calling patterns between the customers of a telephone company to the interactions between a group of drug abusers. Indeed, these two examples are not only in very different fields, they also differ radically in that the first is usually concerned with complete data for many millions of subscribers, whereas the second may involve incomplete data from a relatively small number of individuals, as is all too common in the social sciences.

At the most basic level, the data from a social network can usually be described by a table, whose rows and columns are labeled by the individuals in the group and whose entries represent the frequencies or the presence/absence of the relationship or tie that is being studied. There are no entries on the diagonal. The entries may be symmetric (as in the number of calls between A and B) or asymmetric (as in the number of calls from A to B). Even in the latter case, there is almost always strong association between the A to B and the B to A entries. This may be positive (if A calls B, B is likely to call A) or negative (if A tells B a fact, then B cannot tell A). But, in addition, dependence between entries arises indirectly. As one example, if A ties to B and B ties to C, then usually this makes it more likely that A ties to C. The overall point is that network analysis is concerned primarily with dependent outcomes.

In practice, the data may include measurements on external explanatory variables or may be collected at several points in time but it is the intrinsic dependence that presents the challenge in modeling and analyzing networks from a statistical perspective, because most standard procedures assume independence. Nor is it sufficient to borrow from the time series literature, because this deals only with unidirectional dependence. Indeed, social and other networks have more in common with problems in spatial statistics and there has been some recent emphasis on modifying ideas from the literature on Markov random fields. Thus, a plausible starting point for asymmetric binary tables is to assume that the presence or absence of a tie from A to B, given all other ties, depends only on the state of the tie from B to A and the states of all other ties involving either A or B. This assumption has surprisingly strong implications about the available probabilistic models. What has been lacking has been reliable methods of parameter estimation for such models and methods of testing goodness of fit. Finally, I am glad to mention that several CSSS researchers are also active in applications and theory of social networks and, in particular, to thank Tony Rossini for introducing me to the topic.
Professor Martina Morris and students in CS&SS 320 enjoy coffee and rolls while discussing the "Evaluation of Evidence." This course will be taught every Autumn Quarter and is the first course in a three course series.

Ross Matsueda and Kevin Quinn give away Tootsie Roll Pops to students who asked question about CSSS's curriculum at the School of Arts and Science Dawg Daze.
Winter Quarter 2001 Curriculum Schedule
CS&SS 321
Case-Based Social Statistics I (5)
Instructor: Handcock
Offered: jointly with STAT321/SOC 321

CS&SS 567
Statistical Analysis of Social Networks (4)
Instructor: Hoff
Offered: jointly with STAT 567

CS&SS 560
Hierarchical Models for the Social Sciences (4)
Instructor: Quinn
Offered: jointly with STAT 560/PoliS 560

CS&SS 590
Seminar Series (1)
Instructor: Warren

Spring Quarter 2001 Curriculum Schedule
CS&SS 322
Case-Based Social Statistics II (5)
Instructor: Handcock
Offered: jointly with STAT 322/SOC 322

CS&SS 542
Social & Populations Geography (5)
Instructor: Withers
Offered: jointly with GEOG 542

CS&SS 564
Bayesian Statistics for the Social Sciences (4)
Instructor: Hoff
Offered: jointly with STAT 564

CS&SS 566
Causal Inferences in Social Sciences
Instructor: Richardson
Offered: jointly with STAT 566

CS&SS 590
Seminar Series (1)
Instructor: Warren

CS&SS 586
Network & Social Structure
Instructor: Stovel
Offered: jointly with SOC 586
Did Your Vote Count?

By
Kevin Quinn & Mark S. Handcock

The 2000 US Presidential Election raised many issues of great political importance. These issues include the role of the electoral college, the design of ballots, the local control of balloting procedures, the machine and manual counting of votes cast, and the process of vote certification.

Many of the questions have a statistical aspect. Our interest here is in the statistical issues and how they relate to political science, rather than the political issues themselves. The validity of an election result is a legal issue, rather than one determined by statistical evidence.

Anomalies in Palm Beach County

One of the issues immediately raised was the design of the ballot in Palm Beach County, the second most populous in Florida. It was claimed that the rare and outdated layout of the ballot led people intending to vote for the Democratic Party candidate, Al Gore, to vote instead for the Reform Party candidate, Patrick Buchanan. Prima facie evidence for this was the large vote for Buchanan in the county (0.8%) compared to the statewide average (0.3%). The natural statistical question here is how to assess the evidence for voting anomalies in Palm Beach County.

There have been many analyses of the Palm Beach County vote since the election day (Smith 2000 presents an excellent analysis and review). Many of these looked for evidence that Gore supporters mistakenly voted for Buchanan. Often these analyses consider the relationship between the Buchanan vote and the Bush vote across Florida’s 67 counties. Figure 1 is a typical graphical comparison that plots the total number of votes for Buchanan in each county against the corresponding number of votes for Bush in the same county. The area of the circles surrounding each point is proportional to the total vote in the county. Hence the largest circle (on the right) corresponds to Miami-Dade County. Figure 1 hits you between the eyes – the point corresponding to Palm Beach County (marked by PB) is far from the “scatter” of other counties. Given the total vote count, the direction of deviation appears to be an excess of Buchanan votes relative to the norm. Visually, we might have expected Buchanan to have received about 300-1000 votes in that county instead of the 3407 he actually receive. Many have claimed the cause of this excess was the design of the ballot, and these voters intended to vote for Gore (Brady 2000).

Figure 1. Votes for Buchanan and Bush for each county in Florida

Total number of votes

Does this result stand up to closer statistical scrutiny? Many factors should be borne in mind. A primary one is to quantify how unusual this county is compared to counties in other places and times.

Another approach is to consider more detailed data, such as that at the level of the individual precincts in Florida (Wand et.al., 2000). A further perspective can be obtained by comparing the patterns in non-absentee votes to those in absentee votes (which did not use the suspect ballot) by precinct (Dolar 2000). All of these analyses reach similar conclusions about the degree to which Palm Beach County was anomalous.

These simple graphical displays can be deceiving, though. We expect that variation in the vote totals will increase with the population size. Indeed, in Figure 1 we see a “fanning out” of the points as the vote total increases. Figure 2 is an alternative to Figure 1 that considers the percentage of the vote rather than the total vote. We see that Palm Beach is still anomalous, but not visually as strong as Figure 1. In addition a few other counties (e.g., Calhoun, Liberty and...
Gulf) are also quite different. It should be noted, though, that these other counties are small in size and Figure 2 suffers from similar differing variation as Figure 1. Figure 3 plots the cube root of the vote totals. This transformation was used by Smith (2000) to stabilize the variances of the Buchanan vote. Here we see that Palm Beach County is still an outlier, albeit less visually striking.

If an analysis is based on the county level information we can still adjust for other sources of variation by building a model incorporating them. It is natural to include demographic information about the counties readily available from the U.S. census. Smith (2000) undertook such an analysis and found that the predicted vote for Buchanan in Palm Beach County should be roughly between 181 and 534 votes.

**Ecological Inference for the voting patterns**

The analyses illustrate the difficulties of making inferences about individual behavior with aggregate data. In the social science literature, such inferences are referred to as *ecological inferences*.

For example, in addition to the claim that some Gore voters mistakenly voted for Buchanan, there have been claims that spoiled Presidential ballots (ballots in which no choice of candidate or more than one choice of candidate was registered) in Palm Beach County were cast disproportionately by citizens who intended to vote for Gore. Evaluating this claim is hard based solely on the available data. Given the secret ballot, one will never be able to match individual demographic characteristics and issue preferences (factors that are extremely strong predictors of voting behavior) to actual individual voting behavior.

The data available are the total number of spoiled ballots in Palm Beach County as well as the precinct-level vote totals for all offices. Suppose one allows that more liberal voters intended to vote for Gore and actually did vote for other Democrats (such as Senator-elect Bill Nelson). Then evidence that Nelson voters tended to cast spoiled ballots disproportionately would suggest that Gore was hurt by the number of spoiled ballots in Palm Beach County.

The table presents what is known about the relationship between votes for Nelson and the number of spoiled Presidential ballots. The table makes clear the core difficulty of making ecological inferences: *while the row and column sums are observed, the interior entries of the table are unobserved.*

![Figure 3. Transformed Votes for Buchanan and Bush in Florida](image)

**Figure 2. Percentage Votes for Buchanan and Bush in Florida**

![Percentage of votes](image)

**Figure 3. Transformed Votes for Buchanan and Bush in Florida**

Variance Stabilized vote totals

10 15 20

What can we say about the fraction of Nelson voters who cast spoiled ballots? As Duncan and Davis (1953) showed, the fact that the interior cells must sum up to the row and column totals implies that the interior cells fall within certain bounds. The fact that the interior cells must sum up to the row and column totals implies that the interior cells fall within

Continued on page 8
certain bounds. In this case, the bounds for the fraction of Nelson voters and non-Nelson voters who cast spoiled ballots are completely uninformative- we only know that between 0 and 29,702 Nelson voters who cast spoiled ballots.

Table: Spoiled Presidential Ballots in Palm Beach County by Votes for US Senate Candidate Bill Nelson (Democrat).

<table>
<thead>
<tr>
<th></th>
<th>Votes for Nelson</th>
<th>Votes not for Nelson</th>
<th>Totals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spoiled Presidential</td>
<td>?</td>
<td>?</td>
<td>29,702</td>
</tr>
<tr>
<td>Ballots</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unspoiled Presidential</td>
<td>?</td>
<td>?</td>
<td>432,286</td>
</tr>
<tr>
<td>Ballots</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Totals</td>
<td>269,835</td>
<td>192,153</td>
<td>461,988</td>
</tr>
</tbody>
</table>

is equal to 29,702 minus the number of Nelson voters who cast spoiled ballots. On the other hand, we do know that the number of Nelson voters who cast unspoiled presidential ballots is between 242,080 and 268,017. The number of non-Nelson voters who cast unspoiled presidential ballots is simply 432,286 minus the previous quantity. Unfortunately, the breakdown of voters who cast unspoiled ballots is of little interest.

When the bounds on the interior cell counts are uninformative and data from multiple precincts are available researchers are often willing to make assumptions (of varying degrees) that allow them to estimate the interior cell counts based on associations between the row and column sums across the precincts. We refer the interested reader to King (1997) for a more comprehensive discussion of the ecological inference problem.

The cited references and links related research can be found at http://www.csss.washington.edu/csss-spr01.voting.html.
CSSS statisticians assist social scientists in all quantitative aspects of their research including study design, data analysis and report preparation. In addition, CSSS consultants collaborate with scientists to develop new methods to address scientific questions of interest. Since operations began in October 1999, more than 50 university faculty and graduate students, representing nearly all of the associated departments, made use of the consulting service. In one project, CSSS statisticians assisted researchers in identifying the appropriate analyses for a complicated case-control design. Dr. Jean Kruzich and Jami Bodonyi of the School of Social Work contacted CSSS for assistance in analyzing data collected from Child Protective Services (CPS) of Washington. Their study compares child neglect fatalities to a matched control group of living children with CPS involvement. It also describes the similarities and differences between the neglect fatalities and a sample of accidental deaths of children with CPS referrals. Minimal prior research has been conducted on child fatalities, and their findings will be a crucial initial step in identifying factors that may indicate increased risk for neglect fatality.

In a second project, CSSS statistical consulting assisted Sociology professor Dr. Paul Burstein with methodological questions in preparation for a grant proposal. Dr. Burstein is interested in explaining what happens to policy proposals introduced as bills into the U.S. Congress - why some win increasing support and are ultimately enacted into law, others win little support and disappear from the congressional agenda, and still others stay on the agenda but neither win enough support to be enacted nor lose enough to disappear. We are pleased to report that Dr. Burstein was awarded a NSF grant at the end of spring quarter and, with the assistance of CSSS core faculty member Dr. Kevin Quinn, is now pursuing his research with legislative experts. Statistical consultation and advice is available at no charge to university researchers in CSSS affiliated departments. More information about CSSS Statistical Consulting is available at http://www.csss.washington.edu.

Professor Peter Hoff begins the new year as the center's new Consulting Director, replacing Dean Billheimer.

206-616-6982
hoff@stat.washington.edu
B5K Padelford Hall
see web site for hours
www.csss.washington.edu/Consulting
Seminar Series - Winter Quarter

The CSSS seminar series has been one of the most visible and productive of the Center’s activities in its first year. In Fall Quarter of 2000, the seminar usually attracted about 40 people from dozens of campus units. Seminars meet on Wednesdays from 12:30 until 2:00 in Savery 209. Come early, and bring your lunch!

Please note: Graduate students may obtain one credit per quarter for regularly attending and participating in seminars. Interested students should contact jrwarren@u.washington.edu for details.

The CSSS seminar series is centered around intellectual exchange and interaction between faculty and students in a wide variety of statistical and social science disciplines. The audience is encouraged to ask questions during presentations.

Speakers will end their presentations after 1 hour in order to guarantee time for discussion. The goal is a seminar that looks less like a lecture and more like a spirited discussion of issues raised in a relatively brief presentation of a statistical issue or a research project.

January 10 - Dick Startz (Economics)
“The Relationship Between Permanent and Transitory Components of Recession”

January 17 - Daphne Kuo (Sociology)
“How to Measure ‘What People Do For a Living’ in Research on the Socioeconomic Correlates of Health” (w/ Rob Warren, Sociology)

January 24 - Mark Handcock (Departments of Statistics & Sociology, CSSS)
“Models and Inference for Random Graphs”

January 31 - Paul Waddell (Department of Urban Design & Planning, Evans School of Public Affairs)
“Simulating the Effects of Urban Land Use and Transportation Policies: An Exploration of Potential Applications for Bayesian and Spatial Statistical Methods”

February 7 - TBA

February 14 - Norm Breslow (Biostatistics)
“Designs and Analyses of Case-Control Studies: Applications to Prognosis of Wilms tumor”

February 21 - Peter Hoff (Statistics & CSSS)
“Modeling information, access, and choice”

February 28 - Michael Ward (Political Science)
“Spatial Models of International Conflicts” (with Patrick Heagerty, Biostatistics; and Kristian S. Gleditsch, Faculty of Social Sciences, University of Glasgow)

March 7 - Tony Rossini (Biostatistics)
“Visualization and Analytic Goodness of Fit for Social Network Analysis”

The Executive Committee

The executive committee members serve for rotating terms of two years. This year’s members consist of:

Mary Gilmore - Social Work
Mark Handcock - Statistics, Sociology
Ross Matsueda - Sociology
Adrian Raftery - Statistics, Sociology
Pamela Mitchell - Nursing
Martina Morris - Sociology, Statistics
Elaina Rose - Economics
Gerald van Belle - Biostatistics
Michael Ward - Political Science

Left: Adrian Raftery, Mark Handcock, Thomas Richardson, Rob Warren, Mary Gilmore (hidden), Ross Matsueda, Kent Koprowsicz

see CSSS web page for most abstracts
Recruitment Opportunities

University of Washington, Seattle. Full, associate or assistant tenure-track professor of Nursing, possibly joint with the Department of Statistics or the Department of Biostatistics, as part of the new Center for Statistics and the Social Sciences (www.csss.washington.edu). We seek a faculty member contributing at the cutting edge of the development of statistical methodology relevant to the social sciences. Duties include new course development, teaching at the undergraduate and graduate level, and methodological and collaborative research. Ph.D. required by date of appointment. Send CV and 4 letters of recommendation to Gayle Gray, Administrator, CSSS, Box 354322, University of Washington, Seattle, WA 98195-4322.

The University of Washington is building a culturally diverse faculty and strongly encourages applications from women and minority candidates. EOE/AA employer.

University of Washington, Seattle. Full, associate or assistant tenure-track professor of Social Work, possibly joint with the Department of Statistics or the Department of Biostatistics, as part of the new Center for Statistics and the Social Sciences (www.csss.washington.edu). We seek a faculty member contributing at the cutting edge of the development of statistical methodology relevant to the social sciences, social work and social welfare. Duties include new course development, teaching at the undergraduate and graduate level, and methodological and collaborative research. Ph.D. required by date of appointment. Send CV and 4 letters of recommendation to Gayle Gray, Administrator, CSSS, Box 354322, University of Washington, Seattle, WA 98195-4322.

The University of Washington is building a culturally diverse faculty and strongly encourages applications from women and minority candidates. EOE/AA employer.