A joint newsletter of the Statistical Computing & Statistical Graphics Sections of the American Statistical Association.

December 93 Vol.4 No.3

Statistical COMPUTING & GRAPHICS

A WORD FROM OUR CHAIRS

Statistical Computing

Sandy Weisberg writes his last column as the Statistical Computing Section Chair. The editors want to thank him for his prompt delivery of this column and insightful leadership of the section over the past year.

The Statistical Computing section has a large membership, and is financially stable. Member's dues are used primarily to pay for this newsletter, and to reimburse ASA for section services. The other main functions of the section are assembling a program for the annual meeting, publishing the *Proceedings*, and our annual joint business meeting/mixer with the Graphics section. These activities are self supporting or inexpensive, and the *Proceedings* even make a profit for the section.

CONTINUED ON PAGE 4

Statistical Graphics

Rick Becker writes his last column as the Statistical Graphics Section Chair. The editors also thank him for his leadership of the section over the past year, and for his helpful suggestions for improving the Newsletter.

As you read this, the holiday season will be well underway and the start of a new year will be fast approaching. That means this is my final column as chair of the graphics section and it seems like the right time to review where we are. As the story goes, there is good news and there is bad news.

First, the bad news. Actually, it isn't all that bad. The difficulty I see is a lack of diversity in the participants of section activities.

CONTINUED ON PAGE 4

FEATURE ARTICLE

Easy Access to Census Data

by Matthew Schall

The uniquely complicated structure of census data, and the difficulties involved in creating a useful extract of it, can be a barrier to its use. DB2 (a relational database system), SAS (a suite of software products, some of which may be used to perform statistical analyses), and an IBM supercomputer are now being used to put census data online with a point and click interface that requires no special expertise to use, and gives our campus community both performance and cost benefits.

> ***A researcher may spend weeks running to the codebook and the computer center, debugging code, finding storage space for large data sets, and addressing the myriad other details*** . . .

Transforming a large data set into a statistical package system file for analysis and graphics is a tedious task. For census users, this process is especially difficult because the PUMS (Public Use Microdata Sample) data is hierarchical, with both household and person records. Researchers with GIS applications, epidemiological studies, questions regarding insurance red-lining, and other users of census data, routinely go through the arduous process of obtaining raw census data on tape or CD; going to a codebook; identifying the variables and their locations in the data set; writing package code to associate data, variable names, and attributes; all just to prepare the data set to work with. Many researchers spend weeks running to the codebook and the computer center, debugging code, finding storage space for large data sets, and addressing the myriad other details required to extract a useful set of census data.

CONTINUED ON PAGE 7

EDITORIAL

This, the third and final issue of the Newsletter for 1993, includes not only the usual mix of articles and notes, but also photographs, and cartoons. We hope this lightens up the presentation and encourages you to send us material of a varied flavor for inclusion. *Anything* of interest to the members of the sections is welcome.

This issue contains a feature article by Matthew Schall of the UCLA Office of Academic Computing, describing their successful user-friendly front-end for accessing census data. The census data arrives on tapes, is converted to a database, and is accessed using standard statistical software. The design of the system saves users' time and system resources by avoiding duplication. By publishing this article we hope to hear from others, perhaps even within the Bureau of the Census, to learn how this process can be simplified even further. Surely other efforts to distribute this information, by CD ROM or via the Internet, are in process and can be compared with the UCLA approach. Letters are welcome.

Sallie Keller-McNulty, Statistical Computing Program Chair for 1994, has written a primer for new members who wish to participate in the program at the Annual Joint Statistical Meetings. Given the size of these meetings, this information should be helpful. The Continuing Education activities are described in a short note by Tom Devlin, the Statistical Computing Continuing Education Chair. An unusual statistical anomaly arising in the generation of random data is presented by Tom Ryan. The cartoons were contributed by Andrejs Dunkels, of the Departments of Mathematics & Teacher Education, Luleå University, Sweden.

Our regular columnists have contributed a spectrum of interesting articles— Mike Conlon describes client server strategies under Departmental Computing; Phil Spector continues his tutorial on functions of the UNIX Shell; Dan Carr explores the boundaries of map legends under Scientific Visualization; Mark Monmonier raises troublesome questions on the use of Zip Codes in his column on GIS; and Mike Meyer reports on an improved interface tool for exploring the network, Mosaic, which may put the gopher on the endangered species list. We hope you enjoy them all!

The executive committees of both the Statistical Computing and Statistical Graphics sections of the ASA have strongly endorsed the actions of the governing board of the *Journal of Computational and Graphical Statistics (JCGS)*, by providing a new investment of funds and launching a subscription drive to gain new individual and institutional library subscribers. In support of these actions we have enclosed a letter from the respective chairs of the sections, and subscription forms for convenient subscription to the journal, with a discount coupon. Though university libraries are seriously curtailing their journal subscriptions, most librarians of special collections, such as mathematics, statistics, and computer science, are pleased to have faculty members make recommendations of useful journals—especially when the journals are published by professional societies rather than commercial publishers. Call your librarian today, and ask what is needed or whose recommendation is required to subscribe. You may find that your voice is sufficient. In addition, the offer of Volume 1 for free with a subscription beginning this year (Volume 2) is almost irresistible. If your voice is not sufficient, ask your colleagues to join in requesting a subscription. See the December issue contents on page 23 of the Newsletter for a brief taste of the journal.

Submissions should be sent by e-mail to either of the editors. If you can prepare your article in TEXor LATEX, that will make our lives just a little easier. Otherwise plain old ASCII format is fine.

The Newsletter editors have decided three issues per year is about the right frequency of publication. We will accept submissions, letters, or other announcements and news clippings until Feb 15 for the Spring, June 15 for the Summer, and October 15 for the Winter issue. We look forward to your letters and comments.

Mike, his wife Cindy, and brand new baby daughter, Abrial, would like to take this opportunity to thank Jim for his heroic efforts in producing this issue of the Newsletter. Without Jim, this issue would still be one of the million things in Mike's inbox.

Jim, in turn, would like to thank Joe Broniszewski, our departmental system administrator, for his assistance in scanning and embedding the photographs and cartoons into this document, and to Diane Paules and the rest of our staff for their assistance and for tolerating me at deadline.

James L. Rosenberger
Editor, Statistical Computing Section
JLR@stat.psu.edu

Mike M. Meyer
Editor, Statistical Graphics Section
mikem@stat.cmu.edu

Support for JCGS

To the Editors:

I am writing this letter because I think our sections' members ought to know more about the financial realities of *The Journal of Computational and Graphical Statistics (JCGS)*. *JCGS* was launched with $75,000 of seed money which will be exhausted at the end of 1993. The sponsors (ASA,IMS, and IFNA) have recently agreed to provide additional funds which will allow the journal to continue for a few more years. Nevertheless, given the financial pressures that all organizations face these days, many people, including myself, are very concerned about the situation. Do we need this journal? Do we want it? Will we support it?

There are about 1,500 individual and 200 library subscriptions to *JCGS*. This translates into a budget deficit of about $20,000 per year. The initial budget for the journal figured on 2,000 individuals and 600 libraries signing up. To close the budget gap, we either need 500 more individual subscriptions, 400 new library subscriptions, or some combination of these. The recession has hurt, no doubt, and getting a library to sign up for anything new is a struggle.

I have felt from the beginning that this journal will not succeed without strong support from the Statistical Computing and the Statistical Graphics Sections. What can we do to help? At present fewer than one in four of us are subscribers. If this could be increased to one in two, without any other changes, then *JCGS* would show a $12,000 yearly profit instead of a deficit.

You can subscribe today using the form on page 23 of this newsletter. Another way to help would be for members to encourage their departments or libraries to subscribe. Right now there is a special promotion for such subscriptions. See the form on page 23 for details.

Apart from the financial problems, I do detect a lot of enthusiasm for *JCGS*. The editor, Bill Eddy, reports growing submissions of high quality. Bill Kennedy, the editor-elect, is putting plans in place to assure that *JCGS* establishes its niche firmly in the marketplace. I hope we can pull together and send a strong message to the sponsors: *JCGS* is not only viable but vital to the future of statistics.

Jon Kettenring
JCGS Management Committee
jon@bellcore.com

Copyright Issues Revisited

To the Editors:

I enjoyed Sandy Weisberg's article in the April 1993 issue of the *Statistical Computing & Statistical Graphics Newsletter* (Vol. 4 No. 1). I believe that there is a misinterpretation in one paragraph, in which 'license law' and 'copyright law' are combined. It's the one that starts "The owner of a copyright has authority to sell or otherwise..."

Under copyright law, use of a piece of software is analogous to that of a book – one person at a time. Copyright law does not constrain me from lending or giving away the software, so long as I do not keep a copy. Nor does copyright law constrain me to use the software on one and only one computer (as the Systat 'license' would have it). Copyright law is violated, in my opinion, when the software is used on more than one computer at a time, or when the programs are stored on the hard drives of multiple computers simultaneously.

The license that comes with Systat is not a part of the copyright and has no standing under copyright law. It is not clear that it has any standing at all. None have been tested in court. The reason that a court test is probably necessary is that License Law generally assumes that a contract is negotiated between equals. With 'shrink wrap' there is no opportunity to negotiate.

This becomes a major issue with the use of PC software on networks. My office has 300 PCs on a network, but only 20 statisticians. The business office doesn't use SAS*, and I, in turn, do not use Lotus 123*. To maintain the software on the individual PCs is an administrative nightmare and defeats the ability of the network to provide access to tools irrespective of location. For software that is not network aware, our solution has been to use metering software to insure that we do not use more copies of a program than we have purchased.

Further, some of these 'licenses' say that you have to buy the software for every PC that has access through the network, and not just 'simultaneous use'. Potentially then, any network that is attached to the Internet requires licensing for all machines that are on the Internet (this is a reducto ad absurdum argument, but is one interpretation of some software 'licenses'.).

I hope that this letter sheds some light on this rather vexing problem. I am a statistician too, so my concepts may have serious legal flaws... Most of my knowledge

comes from publications from the US Copyright Office and from reading the computer press. The Copyright Office has a very informative publication: Publication R1, "Copyright basics", Copyright Office, LM 455, Library of Congress, Washington, DC 20559.

* Systat is a registered trademark of Systat, Inc.
SAS is a registered trademark of SAS Institute, Inc.
Lotus 123 is a registered trademark of Lotus Development Corporation, Inc.

Lawrence H. ('Doc') Muhlbaier
Assistant Research Professor
Duke University Medical Center
DUMC 3865
Durham, NC 27710-7510
919-286-8830
muhlb001@mc.duke.edu

FROM OUR CHAIRS (Cont.). . .

Statistical Computing

CONTINUED FROM PAGE 1

This leaves the section in the enviable position of finding new activities that could be of service to section members, and having the money to support the new activities. One area of activity is continuing education. At the San Francisco annual meetings, Mary Lindstrom and Doug Bates offered a very successful and well attended short course on "Nonlinear mixed effects models for clustered data", sponsored by the computing section. I am pleased that the section has decided to use the income from this course to fund scholarships for students to attend short courses at future ASA meetings. You can read more about continuing education activities of our section on page 7 of this newsletter.

> . . . ***the section has decided to use the income from this course to fund scholarships for students to attend short courses at future ASA meetings.***

We plan to offer short courses at other meetings, such as the Interface or the Fall Technical Conference. Tom Devlin devlin@mozart.montclair.edu, the statistical computing section's continuing education chair, is taking the lead both on finding presenters and on finding locations. I'm sure he would be happy to hear from section members with ideas for short courses.

Another area of opportunity for the section is the *Journal of Computational and Graphical Statistics*. This ASA journal is now completing its second year of publication. It is an outlet for new and innovative ideas in computing and graphics, and is of particular interest to our general membership. A new editor, Bill Kennedy from Iowa State, has been appointed for 1995. I urge all members of the section to subscribe, get their libraries to subscribe, and submit interesting articles to the journal.

The roles of the sections in ASA will be changing over the next few years. Large sections like the Computing and Graphics Sections can play an important part in the society and in the future of our profession. This will require the active participation of members of the section. Get involved. Volunteer to be a candidate for section office (send email to the chair-elect hastie@research.att.com as soon as possible). Or send your ideas to the current officers.

Sanford Weisberg
Chair, Statistical Computing Section
sandy@umnstat.stat.umn.edu

FROM OUR CHAIRS (Cont.). . .

Statistical Graphics

CONTINUED FROM PAGE 1

Although we have 2,585 members, activities tend to be dominated by a small group of members. Of course, not everyone can be part of the executive committee, but there are now numerous chances for everyone to be involved in something.

Graphics is an exciting field and we need to share our experiences to help the discipline grow. If you are one of the people who have been on the sidelines, try to do some activity next year: contribute a graphics poster to the annual meetings; volunteer to help with a section committee (send email to the chair); participate in the data analysis challenge; publish your work so that others can learn from you; send in a letter or article to the newsletter.

Now, the good news. The section is financially healthy and is planning more activities to promote statistical graphics.

As you will read elsewhere in the newsletter, the computing and graphics sections are working together to

support the *Journal of Computational and Graphical Statistics (JCGS)* and have instituted a special (subsidized) section-member subscription rate. *JCGS*, owned jointly by ASA, IMS and the Interface Foundation of North America (IFNA), is an important way for us to communicate with one another. It gives section members a place to publish and share interesting and relevant articles. To encourage authors to publish their best work on statistical graphics in *JCGS*, the section is preparing to institute a prize for excellence. (More about that next time.)

Plans are already underway for the graphics program at the annual meetings in Toronto. Bill DuMouchel, program chair, has organized a solid group of invited sessions. The section has sponsored several continuing education courses that have been submitted to the ASA CE committee and we plan to institute a scholarship program to help students attend section-sponsored courses.

Finally, a strong slate of candidates has been assembled for next year's elections. With your participation and support, the section will continue to serve your interests and to present you with numerous opportunities to learn about graphics.

Rick Becker
Chair, Statistical Graphics Section
rab@research.att.com

How to Participate in the Annual Joint Meetings

by Sallie Keller-McNulty
Program Chair, Statistical Computing Section

Many of you may wonder how to get on the program at our annual meetings. For those of you that have attended a recent meeting, you may wonder what the various distinctions are between the different types of sessions on the program. The purpose of this article is to bring everyone up-to-date on how to participate in the annual Joint Statistical Meetings.

A Joint Statistical Meeting is organized by a program committee. This committee is made up of Program Chairs from each of the ASA Sections as well as members from ENAR, WNAR, and IMS. The theme of the meeting as well as the Program Committee Chair is selected by the President of ASA. The Joint Statistical Meetings are made up of two main types of sessions, Invited Paper Sessions and Contributed Paper Sessions.

Invited Paper Sessions are organized by members of the program committee nearly one year in advance. The format of the sessions range from a single lecture to a panel discussion. The speakers in these sessions participate by invitation only. Each of the Sections within the ASA are allocated a certain number of invited sessions for the meetings. The Program Chairs of the Sections are given the responsibility to invite people to participate in the sessions. For the meetings in Toronto next August, the Statistical Computing Section has been allocated six invited sessions and the Statistical Graphics Section has been allocated three invited sessions.

Anyone, without invitation, can participate in the Contributed Paper Session portion of the meetings provided they meet the abstract/registration and draft manuscript deadlines of February 1st and June 1st, respectively.

The number of sessions a particular section is allocated is a function of the section size and the popularity of the sessions organized by that section in past years. Section Program Chairs depend on members to provide ideas for the invited sessions. Since Invited Paper Sessions need to be organized almost one year in advance, any ideas one might have for an invited session needs to be communicated to the Section Program Chair by the summer of the preceding year.

An equally important part of the Joint Meetings is the Contributed Paper Session portion of the meetings. *Anyone,* without invitation, can participate in the Contributed Paper Session portion of the meetings provided they meet the abstract/registration and draft manuscript deadlines of February 1st and June 1st, respectively. There have been some exciting changes to this part of the program in recent years. There are now four types of Contributed Paper Sessions.

Regular Contributed Paper Session. These sessions consist or five of six 15 minute oral presentations followed by a floor discussion. Any ASA member can participate in these sessions by submitting an abstract to ASA on the Abstract Form provided in *AMSTAT News* and registering for the meeting by February 1st. The abstract form asks the participant to indicate which Section they would like to sponsor their presentation. The participant must then send a draft manuscript of the work to be presented by June 1st to the corresponding Section Program Chair. Many people wonder how their particular contributed talk ends up in one session ver-

sus another. Once all the Regular Contributed Paper Session abstracts arrive at ASA, the contributed paper abstracts designated to a particular Section are given to that Section's Program Chair who then attempts to organize the talks into the individual contributed sessions with some common theme.

Regular Contributed Poster Session. Participants in poster sessions display prepared materials on bulletin boards during an allotted two hour time period. During that time, meeting participants can browse the poster sessions and stop for informal discussions with the poster session authors. The deadline and procedures for submitting Contributed Poster Session abstracts and manuscripts is the same as for the Regular Contributed Papers.

Special Contributed Paper Session. This format is designed for those of you who would like to organize a session with your colleagues on some common topic. An example of such a session was the highly successful "Tutorial on Smoothing" at the meetings last August. A Special Contributed Paper Session is a collection of five 20 minute presentations. The fifth time-slot in these sessions can be a discussant or paper. The key differences between these sessions and Regular Contributed Paper Sessions is the length of time allowed for the presentations and the fact that the session is organized before abstracts are sent to ASA. This way the session can focus more directly on a common theme. The person organizing the session must coordinate her/his efforts through the Program Chair of the Section that is to sponsor the session. Anyone can organize such a session. For example, if you had an idea for a Special Contributed Paper Session that you would like sponsored by Statistical Computing, you would need to find the five participants and communicate your intentions to the Program Chair for Statistical Computing. Then you would need to be sure all of your participants submitted their abstracts and paid their registration fee by February 1st. Participants are also required to submit a draft of their manuscript to the Section Program Chair by June 1st.

Roundtable Discussions. This is the newest contributed session format for the Joint Meetings. These sessions are organized along the lines of the current Roundtable Luncheon Discussions, but without food. These sessions are intended for topics and issues that lend themselves to individualized discussion rather than platform presentations. Each discussion must have a minimum of two and a maximum of three leaders. The discussion leaders remain available for discussion throughout the entire session (110 minutes). Unlike the Roundtable Luncheon Discussions these discussions do not have restricted attendance. Meeting participants may (and will) wander from table to table during the session. To organize such a session you must first find one or two colleagues that will also act as discussion leaders on your topic. Then, you must notify the Program Chair for the Section that you wish to sponsor the discussion. Finally, a single abstract for the discussion must be submitted to ASA by February 1st with registration fees for all the discussion leaders. No draft manuscripts need to be submitted for these sessions.

In addition to presenting an invited paper, a contributed paper, or leading a discussion, members are needed for Session Chairs. A Session Chair introduces all session speakers, monitors the session by keeping all participants on schedule, and mediates the floor discussions. If you would like to be a session chair, you should let your Section Program Chair know by February 1st. In order to allow the maximum number of people to participate in the Annual Joint meetings, each ASA member can present one invited paper, one contributed paper or contributed poster, participate once as a discussant, and chair one session.

Many of the ASA Sections which sponsor your presentations produce a proceedings. Both the Statistical Computing and Statistical Graphics Sections produce an annual proceedings based on invited and contributed papers and discussions from the Joint Meetings. If your presentation is with one of these Sections, you would receive manuscript preparation materials following the meetings. The final manuscript deadline is typically mid-October.

As you can see there are many ways to participate in the annual Joint Statistical Meetings. Following the meetings, the proceedings provide an excellent opportunity for you to publish your presentation. If you have any questions or want to plan a Special Contributed Paper Session or Roundtable Discussion for the Statistical Computing Section at the Joint Meetings in Toronto, August 14–18, 1994, please let me know.

Sallie Keller-McNulty
Department of Statistics
Kansas State University
(913) 532-6883
sallie@cecil.stat.ksu.edu

Continuing Education in Statistical Computing

by Thomas F. Devlin
Statistical Computing CE Chair

Two new initiatives concerning continuing education (CE) have been undertaken by the Statistical Computing Section: Co-sponsorship of continuing education events apart from the annual Joint Statistical Meetings and a CE scholarship program.

Expanded CE Activities

The section will co-sponsor the following short courses at Interface 94:

- "Nonparametric Regression and Classification" presented by Trevor Hastie and Robert Tibshirani,
- "Algorithms and Estimation and Visualization of Multivariate Density Functions with Application to Clustering" by David W. Scott,
- "Resampling-Based Multiple Testing" by Stanley Young and Peter Westfall, and
- "Data Analysis with XGobi" by Deborah Swayne, Martin Koschat and Dianne Cook.

This cooperative effort resulted from a recommendation by the Executive Committee of the Statistical Computing Section to consider expansion of continuing education activities beyond the annual Joint Statistical Meetings and the ASA Winter Conference. Consequently, the section CE Committee is pursuing collaboration with professional societies to offer statistical computing-related continuing education at their meetings. The long standing cooperative relationship between the section and IFNA made collaboration at Interface 94 a natural first step. The CE committee invites your recommendations and suggestions regarding both CE activities and professional societies with whom we might collaborate.

CE Scholarships

The section will award scholarships to CE events for 1994. Funding will come from income earned from the highly successful course on "Nonlinear Mixed Effects Models for Clustered Data" presented by Doug Bates and Mary Lindstrom, which the section co-sponsored at the 1993 Joint Statistical Meetings. A scholarship subcommittee consisting of John Miller, John Sall, and myself has been established to implement the program. More information will follow in the next newsletter.

As Statistical Computing CE chair, I ask your help. Please send me your suggestions regarding professional societies with whom we might cooperate, and topics and presenters for continuing education activities.

Thomas F. Devlin
Montclair State College
201-655-7244
devlin@mozart.montclair.edu

FEATURE ARTICLE (Cont.) . . .

Easy Access To Census Data

CONTINUED FROM PAGE 1

Simplification

To reduce the effort involved in using census data, the Office of Academic Computing (OAC) did three things in conjunction with IBM and others at the University of California at Los Angeles (UCLA): (a) moved the California 1980 and 1990 five percent PUMS and Summary Tape 3A and 3B (STF3A and STF3B) files off of tape and on to disk in tables created using IBM's DB2 relational data base management system, (b) installed the SAS/ACCESS interface to DB2, and (c) wrote SAS macros which enable people to invoke the census data interface by just entering the command `census`, look at codebooks online, and read help documentation online. The result is a very easy to use system. To create an extract of the data, a researcher simply:

- Invokes SAS interactively on an IBM ES/9000 mainframe by typing SAS.
- Issues the `census` command.
- Selects a codebook and data set by placing an "s" next to the desired data set on a list of the available census files.
- Enters the name of the data set that will be created to store the extracted data.
- Types an "s" next to the names of the variables to be retained in the extracted data set.
- Subsets the data with a WHERE clause to select particular cases (i.e. child bearing aged females).
- Presses enter one last time.

The whole process takes less than 60 minutes for a novice to learn, and about five minutes to execute. Experienced users benefit from the convenience of an online system for data extraction and by having an online codebook at their fingertips.

The SAS System was an obvious choice for UCLA's census data application because it allows users to easily extract and subset data from tables in a database under DB2. Performance is gained by storing the data presorted and linked in DB2. SAS is a statistical package that easily handles gigabyte sized analytic problems on an IBM mainframe and provides a very easy to use interface to DB2.

By combining SAS and DB2 on a large IBM mainframe, researchers are well positioned to run complex analyses in SAS requiring up to a gigabyte of memory. Since the mainframe is on the campus backbone network, any researcher can `telnet` to OAC from anywhere and perform an extraction. Students and faculty save valuable time and computer resources by extracting data when they need it. Disk and tape charges are eliminated because the user no longer has to store large census data sets, the data can easily be re-extracted when additional analyses are required. For example, it takes less than five minutes to subset the data set for blacks, and merge the 1.5 million person records with almost 600,000 household records to create a final extraction of 300,000 cases.

Behind The Scenes

The entire process from designing the database to making new census data available in a production version can take up to nine months, depending upon the size and complexity of the data. In the approach used at OAC, the Database Administrator receives the U.S. Census data on flat file tapes. In several phases, a design for both the logical and physical structure of the database is developed in coordination with interested parties. After the database is loaded into DB2 tables, DB2 views are created and the SAS/ACCESS descriptor library for each database is developed for the census interface in SAS. Then each census database goes through an extensive internal review and validity testing check before it is promoted from test to production status.

Conclusion

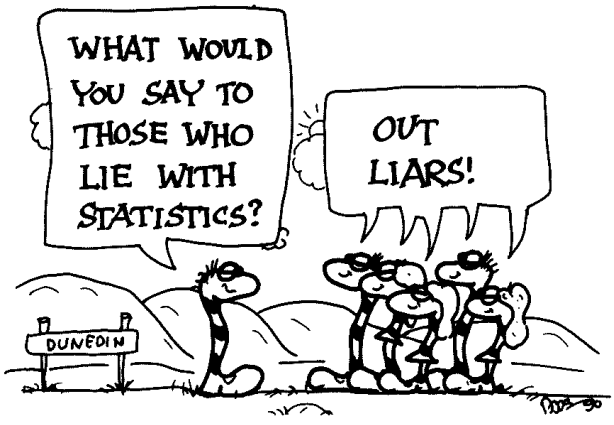
By implementing these databases, researchers save time and expense. They no longer have to individually work with the raw Census Bureau tapes or debug the archaic job control language (JCL). This work is done one time and then made available through menus to the entire research community. By centrally maintaining underlying programming, implementing the latest revisions of the data as the Bureau makes them available, providing disk space so researchers do not have to maintain duplicate copies of the data, preparation time and costs of using census data are greatly reduced for everyone.

As an additional benefit, this approach extends the use of supercomputing to non-traditional supercomputing applications and users. Researchers in the social sciences have not been traditional users of supercomputers. By making the census data so easy to use on a supercomputer, researchers are greatly enabled in an environment where large amounts of data can be analyzed efficiently. The researcher has the freedom to focus on problems related to the research question, and avoid the pitfalls caused by the size of the computer.

Acknowledgements

I want to thank Noel Taylor at the UCLA Office of Academic Computing for her substantial assistance with this article.

Matthew Schall
Office of Academic Computing
UCLA
CUSGMUS@MVS.OAC.UCLA.EDU



by Andrejs Dunkels

Pseudo-Random Data Generators: A Statistical Example

by Thomas A. Ryan, Jr.

The Problem

Problems with common pseudo-random uniform data generators are well known, such as the tendency of successive k-tuples to lie in relatively few hyperplanes in the k-dimensional unit cube (Marsaglia 1968). However, examples of the **statistical** consequences of these problems are difficult to find. In this note I give an example where an obvious application of a published pseudo-random generator gives unexpected statistical behavior.

The Laplace Generator

The uniform generator is used to produce Laplace data (with mean 0 and variance 1) using the following Fortran subroutine:

```
      SUBROUTINE LRAN(X,N,DUM)
C
C     Generate Laplace random data.
C
      REAL X(N), XRND, U1, U2, A
      INTEGER I,N,DUM
      DO 100 I=1,N
          U1 = XRND(DUM)
          A  = ALOG(U1)
          U2 = XRND(DUM)
          IF ( U2 .LT. 0.5 ) A = -A
          X(I) = A * 0.7071068
  100 CONTINUE
      RETURN
      END
```

U1 is used to produce minus an exponential variate A. Then U2 is used to assign a random sign to the variate. The standardized value is stored in array X. Clearly if U2 were truly independent of U1, the distribution of X(I) will be symmetric, and hence have a zero third moment.

In fact, the third moment produced is very non-zero. In 100 runs, each of 25,000 data points, the observed third moments ranged from .15 to .48, with a mean of .315, indicating a definitely skewed distribution.

The Uniform Generator

The uniform data generator used is a multiplicative congruential generator given by Hansson (1966). A Fortran implementation of this generator is:

```
      FUNCTION XRND(X)
C
C     Return uniform (0,1)
C     pseudo-random value,
C     based on ACM Algorithm 266 (1965),
C     modified by L. Hansson (1966).
C
      INTEGER I1,I2,I3
      INTEGER IY
      COMMON /RNDM/ IY
      DATA I1,I2 /125,2796203/
      IY = I1*IY
      I3 = IY/I2
      IY = IY - I3*I2
      XRND = FLOAT(IY)/FLOAT(I2)
      RETURN
      END
```

The first-order autocorrelation of the entire sequence produce by XRND is .004, so autocorrelation does not explain our problem. If we look at a plot of pairs of successive values, we see that the values fall in lines, as is common with congruential generators. The lower left corner of this plot (values with both coordinates between 0 and .1) is seen in Figure 1.

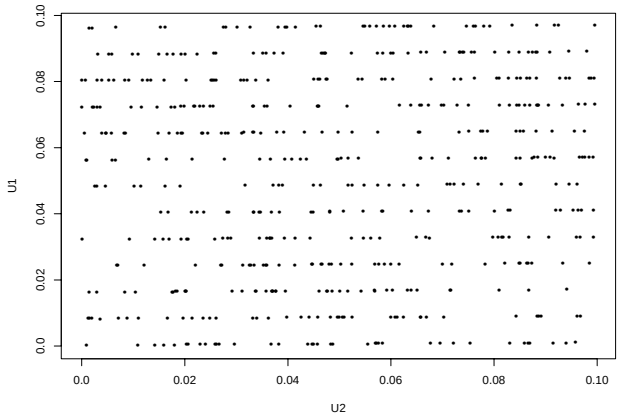


Figure 1: Lower left corner of a plot of successive pairs generated by XRND.

While this shows a less than ideal pattern, it isn't obvious from this plot why we are observing the skewness in our Laplace data.

The Solution

The key observation is this: Suppose we observe a value U produced by XRND which is less than .004. Now let's look at the next value of U produced. A value of U less than .004 corresponds to a value of IY less than 11,184,

which in turn means that the next value of IY will be less than 125 times 11,184, or 1,398,000.

This example shows the statistical impact of a choice of a pseudo-random number generator.

This value of IY will be transformed to a value of U of less than 1,398,000/2,796,203 = .499964 < .5. Now let's see what this means for the Laplace generator. If we observe a value of U1 less than .004, the corresponding value of A will be less than - 5.5, and the value of U2 will be less than .5, meaning that the value of X will be positive. Thus the most extreme 0.4% of the data will all be positive, giving rise to the skewness we observed.

This example shows the statistical impact of a choice of a pseudo-random number generator. While the problem observed in this example is a function of the small multiplier of the generator, other patterns can also cause problems. Small multipliers make it easier to implement portable random data generators, so less dramatic versions of the problem observed here may be present in other generators. This suggests using either a combined generator such as that proposed by L'Ecuyer (1988) or shuffled generators.

References

L'Ecuyer, Pierre. June 1988, "Efficient and Portable Combined Random Number Generators." *Commun. ACM* 31, 6, 742-749,774.

Hansson, L. 1966, "Remark on algorithm 266." *Collected Algorithms from CACM.*

Marsaglia, G. Sept. 1968, "Random numbers fall mainly in the planes." *Proc. Nat. Acad. Sci.* 61, 25-28.

Thomas A. Ryan, Jr.
Department of Statistics
The Pennsylvania State University
TAR@psuvm.psu.edu

DEPARTMENTAL COMPUTING

The Agony and Ecstasy of Client/Server Computing

by Michael Conlon

What is Client/Server computing?

Client/Server computing (we'll call it C/S) for the purposes of this article is a departmental computing model in which one computer (the file server) provides a central repository for software and possibly user files on a network. Other computers (called clients) get files from the server as needed. If done properly, this is all transparent to the user. On a PC, one of the user's "drives" may be located across the net on a file server. The user uses the network drive as if it were on the client computer. On a Mac, the file server is accessed using AppleShare and a public area can be available to each user's machine at startup. Clicking on programs from the server is identical to starting them from one's hard drive. C/S is a natural way to operate in the UNIX environment, where file systems are mounted with NFS (Network File System) and parts of the user's directory hierarchy may be located on various machines on the network.

Ecstasy

C/S is all the rage in computing. Why? There are several highly touted benefits. Having a common software collection on a server simplifies software maintenance — everyone is using the same version which is installed just once. This reduces system management requirements and reduces training requirements. In turn, large software collections can be maintained. There are economies of scale in purchasing large drives for a server which can then hold single copies of many different programs. Departments gain in copyright protection, since servers can be equipped with metering software to insure that only as many copies of the software as are licensed are actually used. Departments gain file sharing capability on the network. Users can save files to the server which can then be accessed by other users. It is simple to back up the file server (one machine) which then protects the software collection and any user files stored there.

Agony

So if C/S is so wonderful, why isn't everyone doing it? Well, just about everyone is doing it. Are there problems with this approach? Sure.

Dependency on the server.

If your server goes down, and your software and files are on the server, you can't work. Sounds like a 1975 mainframe system, doesn't it? Turns out not to be a major problem. Today's servers are quite reliable. Our Sun 670MP server had more uptime last year than the University IBM 3090 mainframe.

Dependency on the network.

If the network hubs or routers have problems, remote users may be disconnected from the server, or experience significant slowdowns. This is still a problem (at least for us) and servers should be located "near" clients in terms of network topology to improve reliability.

Dependency on the system manager.

Users don't install software on the file server. The system manager does. If the users are accustomed to installing their own software and tinkering with it, they will experience a significant loss of control and possible frustration with having to deal with what might appear to be a bureaucracy. A good system manager should be able to install software more efficiently than a user freeing the user to do what ever they were actually hired to do (which wasn't installing software).

Security.

With files on a server and accessible via a network, the potential exists for unauthorized access. This is a serious concern and a system manager must spend time and effort to make sure file permissions are assigned carefully and that loopholes to unauthorized access are closed. Each type of system has its own peculiar security problems.

Cost of acquisition and maintenance.

Servers can be expensive. System managers are expensive. However, the benefits (ecstasy) of C/S have typically far outweighed the costs (agony).

Adopting common software.

While the benefits of adopting common software are easy to see, there is a downside. In a perfect world we'd have universal acceptance of a word processor for a department. Such consensus does not typically exist. Users who have had their own machines and have installed their own software are typically comfortable with their choices and unwilling to change to a common software standard. This may be the largest single hurdle in adopting C/S solutions.

Coexistence

I have seen two approaches to handling the dilemmas involved with adopting common software in a C/S environment. My department has adopted a strong department server model in which we work to build consensus on common tools, install chosen software and defend our choices when asked to install other tools. We have several brands of software for each purpose. For word processing we have WordPerfect and LaTeX . We install free software without much consideration of commonality. Mail systems are all smtp based, users can have a variety of front-ends. Printing is PostScript based.

My college operates cross-department servers. Here the issues are different. No attempt is made to provide a common software environment. If the software is licensed, it can be installed on the server. So if some faculty want to use WordPerfect, others AmiPro, others Word and still others LaTeX , that's fine as long as the software is bought and paid for. The departments may encourage their users to use particular tools to foster interchange, but a wide variety of tools are available.

Both approaches are born of expediency. In the strong server model, the needs of the department for document interchange and streamlined training outweigh the requests of users to have alternate tools. Some of these decisions are very difficult. We work hard to evaluate alternatives and build consensus. We do not dictate.

So if some faculty want to use WordPerfect, others AmiPro, others Word and still others LaTeX , that's fine as long as the software is bought and paid for.

The open server model fosters the diversity of work styles that characterize a college of liberal arts and sciences. Our computing environments are very diverse (PC/XTs in some departments, DEC Alphas in others). Adopting a common software environment is not possible. But individual departments must then cope with "too many tools." This is both a blessing (diversity) and a curse (support).

C/S is the new model for department computing. Reliability of servers and networks has improved dramatically. The human issues of system managers and choice of software will remain with us.

Michael Conlon
Department of Statistics
University of Florida
mconlon@stat.ufl.edu

What is a Shell?

by Phil Spector

As mentioned in the first article of this series, the heart of the UNIX operating system is the kernel. Users generally don't access the kernel directly, but rather communicate with the kernel through a program known as a shell or command interpreter. For example, the prompt displayed on your screen is displayed by the shell; and when you type a command such as `ls` or `mail`, it is the shell that sees the command and responds.

A shell is just a program like any other program, but it performs a variety of functions. There are many different shells used in the UNIX world. Most users' needs will be satisfied by whatever default shell is provided by the system they use, but each shell has certain features that may make it more attractive to some users. Some of the shells which are widely used include the Bourne shell (`sh`), the C-shell (`csh`), the modified C-shell (`tcsh`) and the Korn shell (`ksh`). Although specifics differ, there are some general services which all shells provide as well as special features which exist in only some of the shells. In the next few sections, I'll present an overview of these features. As with every command in UNIX, the UNIX `man` command is the place to go for more details.

A shell is just a program like any other program, but it performs a variety of functions.

To find out the name of the shell you are using, you can use the `finger` command, giving your login id as an argument. There should be a line which displays what shell you will get by default when you log into your system. You can change shells by typing the name of the shell you desire. If it is available on your system, subsequent commands will then be interpreted by that shell, and, for example, you may find that the prompt you see has changed. To change back to your original shell, type Control-d.

Search Path

One of the basic functions the shell performs is to find the commands you ask it to execute. Almost every command in UNIX is stored in a file, and the shell keeps a list, known as a search path, of directories in which to search for commands. You can see the directories your shell is using by typing `echo $PATH` or `echo $path`, depending on which shell you are using. When you type a command like `ls` (which displays the names of files in your current directory), the shell looks in each of the directories contained in the search path, and executes the first occurrence of a file called `ls` which it encounters. There is one exception to this rule. Some commands are built into the shell, that is, the shell recognizes them and executes them without resorting to the search path. To determine exactly what is being executed when you type a particular command name, you can use the UNIX command `which`. You can override the search path by giving a fully qualified file name for a command, that is, one which contains one or more slashes (`/`), to specify the directory in which it resides.

Filename Abbreviations

Certain characters, sometimes known as magic characters, have special meanings to all the shells, and are expanded to these special meanings before being passed to the program which you invoke through the shell. A filename containing a magic character is sometimes called a wildcard, and the expansion process is referred to as wildcard expansion. It is important to realize that the shell does wildcard expansion, not individual programs, so that wildcard expansion is available for all programs regardless of their source. In addition, no program ever sees the magic characters you type, unless they are preceded by a backslash (\), or surrounded in single quotes, in which case they lose their "magic" properties. The table below shows some of the magic characters and how they are expanded by the shell.

Character	Meaning
`*`	anything
`?`	a single character
`[`c_1`-`c_2`]`	range of characters
`[`$c_1c_2\ldots$`]`	characters within brackets

When you submit a command containing a wildcard to the shell, the magic characters are expanded to match the names of existing files which contain the patterns specified by the wildcard. For example, suppose the following files exist in your current directory:

```
data.old     data3        goodprogram.c
data1        file.c       program.c
data2        firstdata    prog1.c
```

If you wished to use the UNIX program `wc` to count the lines, words and characters in all your C programs (that is, files with the suffix `.c`), you could use the command `wc *.c`. What the `wc` program will actually see in this case is the list of files which match the wildcard,

namely `file.c`, `goodprogram.c program.c`, and `prog1.c`. To match files with the word data followed by a number, you could use `data[0-9]`; this would match `data1`, `data2` and `data3`, but not `data.old`. To match all files with the word data anywhere in the name, you could use `*data*`, and so on. It bears repeating that the shell expands the wildcard, not the individual programs, so no program ever knows that it was called with a wildcard; the programs simply receive the list of file names which match the wildcard you submit to the shell.

It is important to realize that the shell does wildcard expansion, not individual programs, so that wildcard expansion is available for all programs regardless of their source.

One other special filename character deserves mention. In the C-shell, the tilde (~) is recognized as referring to a home directory. If you use the tilde alone, it refers to your home directory; following the tilde by a username refers to that user's home directory. So, regardless of your current directory, a filename like `~/myfile` refers to the file `myfile` in your home directory, and a filename like `~fred/somefile` refers to the file `somefile` in the home directory of the user named fred.

Redirection

By default, most UNIX programs receive their input from the keyboard, and write their output to the terminal screen. Actually, there are three so-called streams which most programs use: standard input (often called `stdin`), standard output (`stdout`), and standard error (`stderr`). One of the services of the shell is assigning other locations for these three streams; this process is known as redirection. A single greater-than sign (`>`) will send standard output to the file whose name follows it, potentially destroying the file if it already exists. Two greater-than signs (`>>`) will have a similar effect, but will append to the output file rather than destroying it. A less-than sign (`<`) will redirect standard input, so that input to a program will come from a named file, instead of the terminal. Like wildcard expansion, redirection is performed by the shell, so it will work for any program which obeys the UNIX convention of using the three streams described above.

You should realize that the redirection described above does not modify the destination of `stderr`; in particular, error messages will still be displayed to the screen even if `stdout` is redirected. While this is clearly desirable for interactive work, when you use redirection to save a non-interactive job's output to a file, you should always make sure that `stderr` is redirected to the file as well. Methods for redirecting `stderr` vary among the different shells — check your online manual pages for the details for the shell you use.

A pipe allows you to take the standard output of one command, and use it as the standard input to another command.

Another valuable form of redirection is a pipe. A pipe allows you to take the standard output of one command, and use it as the standard input to another command. For example, suppose we wished to use the UNIX `ls` command to put the names of the files in the current directory into a file called `myfiles`. The command to use would be `ls > myfiles`. Now suppose we wish to count the number of files using the UNIX command `wc`. One solution would be to call `wc myfile`. But using pipes, there is no need to even create a file to hold the output; `ls`'s standard output can used directly as standard input to `wc` by using a vertical bar (`|`), which is the symbol for a pipe, as in `ls | wc`. Like wildcard expansion, redirection is carried out by the shell, not the individual programs, but there is one difference which is illustrated by the previous example. If a program wants to, it can tell whether or not input has been redirected, unlike wildcard expansion which is completely transparent. It turns out that `ls` has been written to take special action when its output is redirected to a file or through a pipe, namely that it lists files in a single column, instead of using its multicolumn default. So while redirection will work for any program which obeys the standard input/output conventions of UNIX, redirected output is not guaranteed to be identical to the output you would see displayed on the screen.

In the next article, I'll discuss additional services which the shell provides.

Phil Spector
Applications Manager
Department of Statistics
UC at Berkeley
`spector@stat.Berkeley.EDU`

Zip Codes, Data Compatibility, and Environmental Racism

by Mark Monmonier

An ability to link differently structured databases poses an intriguing dilemma for the fuller use of geographic information systems (GIS). On one hand, GIS encourages heretofore impracticable analyses. And on the other hand, some analyses raise troublesome questions about data compatibility, convoluted errors, and uncertain results. These issues can be politically troublesome when a conceptually weak geographic analysis suggests a plausible basis for conventional wisdom and public policy.

Unequal Protection

A case in point is a journalistic tour de force published in the September 21, 1992 issue of The National Law Journal (NLJ). A series of related articles under the general title "Unequal Protection: The Racial Divide in Environmental Law" concluded that environmental regulations were less rigorously enforced in areas occupied largely by African-Americans and other ethnic minorities. Among other awards for this work, journalists Marcia Coyle, Marianne Lavelle, and Claudia MacLachlan received the Scripps Howard Foundation's National Journalism Award for environmental reporting.

An important part of their work was a geographic analysis that linked the performance of the U.S. Environmental Protection Agency (EPA) to the racial composition of areas surrounding toxic waste sites. To complement a series of case studies (vignettes) and identify particularly egregious examples of slow remediation and weak enforcement, NLJ reporters assembled an enormous amount of data on the progress of cleanup at Superfund sites and the EPA's collection of civil penalties for environmental violations. Staff identified each location by Zip Code, and weeded out sites for which no Zip Code could be discerned as well as sites in Zip Code areas without residential populations. The result was a Superfund data set representing 1,177 of the EPA's final list of 1,206 Superfund sites (as of March 1992) and an enforcement data set representing 929 cases (concluded between 1985 and March 1991) in which authorities assessed a civil penalty. Analysts linked these sites and cases to racial and income data (estimated for 1989) for their respective Zip Code areas. For each data set, they ranked locations according to the white percentage of the population, used these ranks to divide the data into quartiles, and compared the highest quartile (identified as "white areas") with the lowest quartile (identified as "minority areas").

Means computed for the upper and lower quartiles provided a basis for a bar graph showing that "In certain of the 10 autonomous regions that administer EPA programs, the pace of cleanup at Superfund sites is far slower for minority communities." In the EPA's Midwest region, for example, average cleanup time was more rapid in white areas (9.7 years) than in minority areas (13.8 years), and two other sets of horizontal bars revealed similar differentials for the West (9.3 years for white areas and 12.3 years for minority areas) and the Great Plains (9.6 years for white areas and 12.3 years for minority areas). Although the graph did not address the EPA's other seven administrative regions, another graph indicated that for the nation as a whole, the differential in cleanup time was greater for race (with averages of 4.39 and 5.63 years for white and minority areas, respectively) than for income (with averages of 4.79 and 5.31 years for the highest and lowest income quartiles).

Although environmental racism is real and reprehensible, five-digit Zip Codes are a poor basis for a broad, systematic investigation of racist practices . . .

Selective use of favorable comparisons in the report's text and graphics suggests that the analysis itself was largely rhetorical. It would appear that the journalists highlighted the numerical results only where the data strongly confirmed their hypothesis. Because the series contains no maps or systematic tables, it is impossible to tell whether other results revealed weaker differentials, inconclusive differences, or differentials that contradicted the hypothesis of environmental racism.

Flawed Definitions

Whatever the strength or direction of these unpublished results, conceptual flaws undermine the entire analysis. A prominent limitation is the fact that some Zip Code areas in the "minority areas" quartile are only 15.9 percent non-white for the Superfund data and only 20.8 percent non-white for the enforcement data. Simple arithmetic would describe these populations as 84.1 and 79.2 percent white, respectively—slightly integrated perhaps, but hardly "minority areas" in a nation that is about 83 percent white overall. The NLJ reporters might at least have either used deciles rather than quartiles or contrasted areas less than 50 percent white with areas more

than 90 or 95 percent white. Of course, estimates for the somewhat smaller, more narrowly defined white, non-Hispanic segment of the population would have yielded more meaningful minority clusters, based on ethnicity as well as race.

> . . . ***marketing firms and the producers of "Beverly Hills 90210" conveniently equate postal codes with income and status,*** . . .

But better demographic measures alone would not have made the results meaningful because Zip Code areas are neither census tracts nor neighborhoods. Although marketing firms and the producers of "Beverly Hills 90210" conveniently equate postal codes with income and status, Zip Code boundaries reflect the local geographic organization and operation of the U.S. Postal Service, not the boundaries of homogeneous socioeconomic communities. And even if the Postal Service had deliberately sought racially distinct postal zones, Zip Code areas generally are much larger than census tracts and inherently more diverse. (Although an accident of postal geography might let some Zip Code areas reflect comparatively large minority neighborhoods in major cities, postal zones are more likely to be racially mixed than segregated. Additional information is needed to tell whether a racially mixed Zip Code reflects, for example, a uniform zone with segregated housing or the juxtaposition of a white ethnic enclave and an impoverished black neighborhood.) Moreover, the comparatively large size of Zip Code areas allows a substantial separation between residential and industrial neighborhoods, and between homes and toxic dumps. While residents of rural areas and small cities relying on well water might be highly apprehensive about groundwater contamination anywhere within their Zip Code areas, the water supply systems of large metropolitan areas with substantial minority populations typically rely on aqueducts and reservoirs, not on local aquifers vulnerable to a toxic landfill a block or even a mile away.

Zip Code Convenience

Why then is Zip Code information used so widely in marketing studies and advertising campaigns? Because of the obvious and straightforward link between demographic data and potential buyers or voters. If a study reveals, for example, that well-heeled Republicans account for 70 percent of the households in a Zip Code area, campaign literature mailed to all addresses in the zone will most certainly reach a high proportion of potential voters likely to support a candidate advocating traditional family values and lower top-bracket tax rates. But in the NLJ's analysis, this link was missing. Simply put, aggregated demographic data reported by five-digit ZIP Codes reveal little about the people (or the land use, for that matter) in the immediate vicinity of point locations.

> . . . ***the comparatively large size of Zip Code areas allows a substantial separation between residential and industrial neighborhoods, and between homes and toxic dumps.***

Although environmental racism is real and reprehensible, five-digit Zip Codes are a poor basis for a broad, systematic investigation of racist practices in either environmental enforcement or the cleanup of Superfund sites. A credibly thorough geographic analysis requires more detailed information based on smaller spatial units and identified links between toxic dumps and drinking water.

Mark Monmonier
Syracuse University
mon2ier@mailbox.syr.edu

TOPICS IN SCIENTIFIC VISUALIZATION

Constructing Legends For Classed Choropleth Maps

by Dan Carr

Constructing map legends is a statistical graphics topic worthy of attention. Map legends can provide a distributional summary for a variable represented on a map. A distributional summary augments the map's spatial information and provides the reader with a few key numbers to remember. As an example, researchers and politicians may be concerned if one of "their" counties had a high death rate due to a specific type of cancer that is environmentally or life-style related. The high rate becomes more deserving of study (or more useful for funding leverage) if the rate for the people in this county is above the 95 percentile for the nation. While death rates have a direct interpretation, a population-based comparison provides a useful standard of reference. Thus distributional summaries should be considered as part of the map construction process.

A great deal can be learned about maps legends by examining the options available in GIS packages and by looking at publications. For example Goldman

(1991) provides numerous county-based choropleth maps showing death rates and environmental hazards. The maps show counties with high death rates or high potentials for exposure and include both density and percentile legends that are based on the number of counties involved. The use of distributional summaries based on the number of political regions is common and often convenient. The distributional summaries promoted here answer questions about percentage of people or percentage area. These differ for summaries that count the number of political regions. Figure 1 provides an example.

The 802 regions represented in the map are health service areas. Health service areas (HSAs) are either counties or aggregates of counties as discussed in the last newsletter article (Carr and Pickle 1993). The spatial patterns of mortality rates are clearly of interest with the higher rates in the Northeast. The legend provides a table lookup capability for the classes of mortality rates. The legend also provides a distributional summary for the percent of the white male population. For example the legend indicates that 50% of white males live in HSAs with rates at or below 22.4 deaths per 100,000.

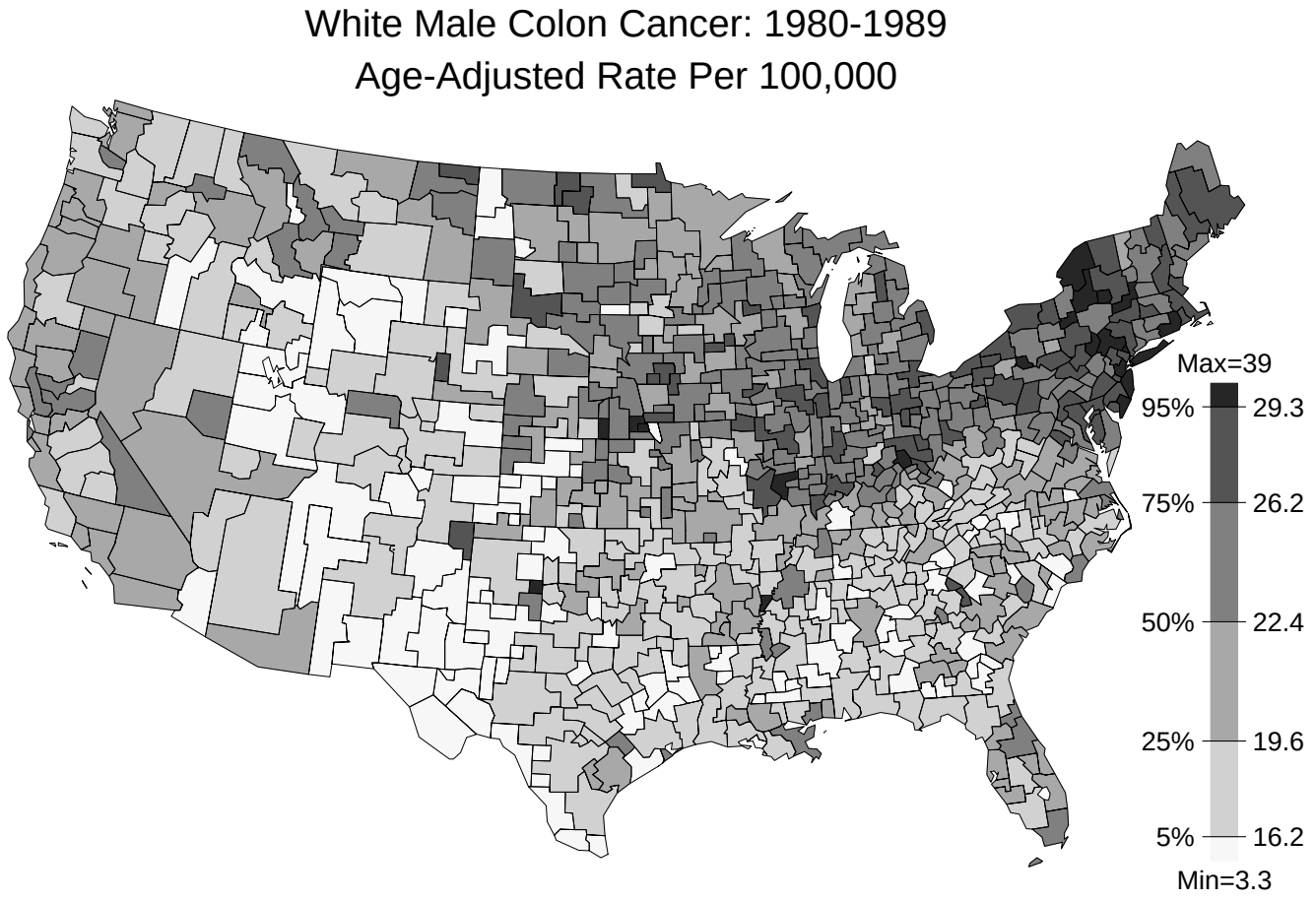


Figure 1: Example of a county-based choropleth map.

Figure 2 shows three candidate distributional summaries. All three plots involve sorting the HSA values in increasing mortality rate order before calculating the cumulative values. The left plot is an approximate cumulative distribution for the described population. The center plot gives cumulative percentages for the number of HSAs and the right plot provides the cumulative percentages of map area. The second summary is mostly of interest because it is often used. The area-based summary is often relevant for maps of environmental variables. The area-based summary is interesting here because it provides a rudimentary characterization of what we see on the map. (While the map projection is area preserving, the characterization would be more exact if our visual response were linear with area and unaffected by color interactions.)

A well-known problem with choropleth maps is that the large regions draw visual attention that is not nec-

essarily proportional to the described population. Figure 3 shows the difference in percentages between the area-based and population-based cumulative estimates in Figure 2. The striking result demonstrates once again the importance of plotting the difference between curves rather than visually estimating the difference. The positive percents suggest that after putting the values in mortality-rate order, the large area HSAs are encountered sooner than the high population HSAs. This suggests a relationship between mortality rate and population density. When the population percents define the class intervals, visual attention drawn to the relative areas with different shading can have a population density based interpretation.

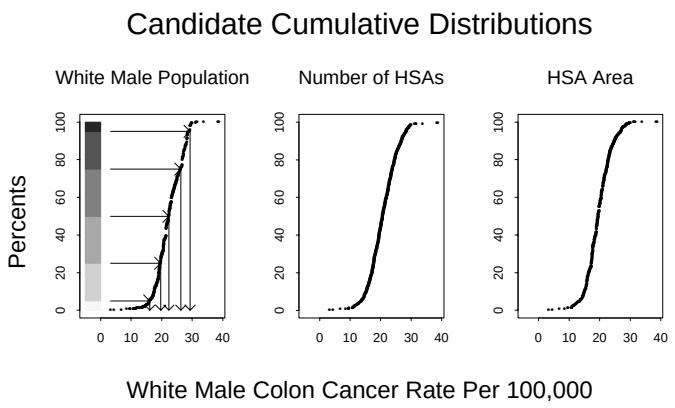


Figure 2: Example cumulative summary distributions.

Consider viewing a series of maps that use the same percents to define class intervals. A few maps may have an unusually small total area devoted to the high rate class and an usually large total area devoted to the low rate class. For gray scale maps this changes total amount of reflected light and acts as a cue. Defining the class intervals symmetrically with respect to population percents allows the meaningful class area comparisons within a map. This comparison can be done for the pair classes in gray scale maps such as Figure 1.

The striking result demonstrates once again the importance of plotting the difference between curves rather than visually estimating the difference.

For example, the lowest rate class covers a much larger area than the highest rate class. The comparisons are easier when a symmetric color scheme is used like the one described below. Of course directly plotting mortality rate versus population density is better than relying on visual estimates of class area differences and other variables need be considered than surrogates like population density and spatial position. The point is that the area devoted to the classes can be suggestive.

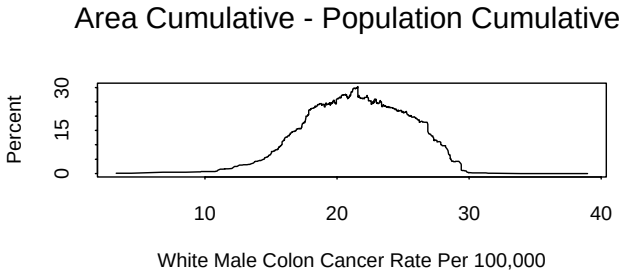


Figure 3: Difference in percentages between area-based and population-based cumulative estimates.

The legend in Figure 1 shows both population percents and mortality rates as class boundaries. The mortality rates are quantiles determined from the cumulative distribution. The arrows in the left plot of Figure 2 illustrates the process of going from selected percents to the corresponding quantiles. A convenient approach is to linearly interpolate between the points on the cumulative distribution rather than treat the distribution as a right continuous step function. Thus the estimated quantiles may not produce the intended percentages when used as class boundaries. However, using more accurate but non-standard percentages complicates the description. Unless there is big discrepancy, simplicity suggests emphasizing that the quantiles are approximate.

Simple Design and Boxplots

When designing map legends it helps to pay attention to graphics methods that have been successful in the past. The boxplot has been one of the few modern statistical plots that has worked its way into elementary statistical texts and into use by applied scientists. The boxplot with its emphasis on approximate .25, .5 and .75 quantiles recognizes some important factors in its design. First it uses a simple standard (.25, .5 and .75) that is easy to accept. Second, the standard relates strongly to important concepts of central tendency, distributional spread, and assessment of symmetry. Third, the summary focuses mental attention on a few values that can be used for making comparisons.

Emphasizing a few values for mental comparison is an important principle. The human short term memory can only handle about 7 ± 2 units of information at one time. (Depth of thought may relate more to what people use as

units of information than to the ± 2). Ehrenberg (1981) argues persuasively that short term memory considerations should be used in the design of tables. For example an ordinary person can divide a two digit number by a two digit number and have room to store an approximate two digit answer. Most people have difficulty when they try to ratio two three-digit numbers. What happens in the graphical environment is somewhat different because the graphic can be used for rapid mental refresh. However one might conjecture that people will withdraw from map reading if there are more than seven or so obvious and equally important classes or layers of information.

Color Scales

The legend in Figure 1 uses six classes with internal boundaries determined approximately by 5, 25, 50, 75, and 95 percentiles. While humans can easily distinguish many more than six gray levels, Figure 1 appears complicated. Part of difficulty relates to the dot-based representation of gray, part relates to the spatial variability that provides a changing background against which to judge color and part relates to using as many as six "equal" classes. When full color is available the map can be made to appear simpler by using shades of red for high rates and shades of blue for low rates. This "grouping" of information has only two equal classes at the top and three ordered classes nested inside. The red and blue colors can be ordered both in terms of saturation and value. Using low-saturation near-white colors in the middle of the scale eases the transition from blue to red. Putting the saturated and dark colors at the extremes follows the advice of Eduard Imhof (see Tufte 1990) by devoting relatively little area to saturated colors.

Of course other color scales can be considered, but the above scale is a reasonable start for those who are not red color blind. The verbal description above does not do the color selection justice. For those interested, compressed color postscript files are available by anonymous `ftp` to `galaxy.gmu.edu` and stored under `submissions/eda/maps`. The directory also contains the Splus commands files and the data used to produce the maps. Splus users can easily modify the colors and experiment with legend variations.

Variations

Legend variations are worth considering. Those who are intensely studying the phenomena may want to see the full cumulative distribution so they can read the percentage for any mortality rate. The Figure 1 legend shows only selected values both to save space and keep the legend simple. The legend scaling is linear in percentages rather than mortality rates to provide programming convenience and generality. If the legend scaling were based on mortality rates, the quantiles corresponding to the standard percentiles could be so close that they would overplot. For visual communication a linear scale in terms of mortality rates would also be helpful.

When full color is available the map can be made to appear simpler by using shades of red for high rates and shades of blue for low rates.

Many map variations are worth considering. Perhaps the most suggestive and elegant map in the map directory cited above is that showing the extreme residuals from the smooth. These local discrepancies from the smooth can be very useful for hypothesis generation. However new topics like exploration of spatial residuals and disaggregation approaches to smoothing using high resolution population data deserve separate consideration.

Acknowledgements

I thank Linda Pickle of the National Center for Health Statistics for discussions concerning the data and Tony Olsen of the U.S. EPA for stressing the importance of distributional summaries. Research related to this article was supported by NSF under grant No. DMS-9107188 and by EPA under cooperative agreement No. CR820820-01-0. The article has not been subjected to the review of the EPA and thus does not necessarily reflect the view of the agency and no official endorsement should be inferred.

References

Carr, D. B. and L. W. Pickle. 1993. "Plot Production Issues and Details: Smooth Cancer Rates and Hexagon Mosaic Maps." *Statistical Computing & Statistical Graphics Newsletter,* Vol. 4, No. 2, pp. 16-20.

Ehrenberg, A. S. C. 1981. "The Problem of Numeracy." *The American Statistician,* Vol. 35 No. 2, pp. 67-71.

Goldman, Benjamin A. 1991. *The Truth About Where You Live, An Atlas For Action on Toxins and Mortality.* Times Books, New York.

Tufte, Edward R. 1990. *Envisioning Information.* Graphics Press, Cheshire, Connecticut.

Daniel B. Carr
George Mason University
`dcarr@galaxy.gmu.edu`

Mosaic and WWW

by Mike Meyer

For many months (the world of networked information changes so quickly that the longest appropriate metric is months!), gopher had been the most popular network navigation tool. Seemingly from nowhere, gopher has been supplanted by Mosiac (in X-windows, Mac, and DOS/Windows versions) as the cool tool to use. There have been articles about Mosaic in the New York Times, Washington Post, and even Science magazine.

Mosaic is a hybrid product which provides an interface to the World Wide Web (WWW), Wide Area Information Service (WAIS), and gopher. Mosaic is a very slick, professional tool reflecting its roots as a product of the National Center for Supercomputing Applications (NCSA) at the University of Illinois.

What is WWW and Mosaic?

The World Wide Web is an Internet-based hypermedia *browsing* protocol, that allows you to display documents and data from all over the Internet. You should note the emphasis on browsing. This is both a strength and weakness of the WWW and Mosaic.

Mosaic is by far the most popular client software for accessing WWW servers. It is fair to say that the WWW might have completely languished without Mosaic.

So what is it! If you have ever used Hypercard on a Macintosh, then you know something about WWW and Mosaic. When you start up any of the WWW clients, and in particular any of the Mosaic tools, you are presented with a page of information. On that page there will surely be hyperlinks and perhaps various icons. The hyperlinks are areas on the page where one can click to retrieve more information or follow a link to a related subject. For example, if this was a WWW document I could put a hyperlink here that pointed to the editorial. Merely clicking on the appropriate word or icon would bring up the referenced document, sound, picture, animation or whatever. It quickly becomes clear why it is called a web. There is no one starting place, nor any natural path through the maze of material. A document author knows what other documents are referenced but need never know what other documents reference the new material.

Navigating within the global web is not particularly easy, but it is an awful lot of fun. If you know you want material about (say) Australia it is not easy to find out where to start, but once you get started there is a lot of information to be found. Documents are referenced by Universal Resource Locators (URLs) which give an Internet style pointer to the document. For example, one of the sources that I keep in my hotlist of interesting things is the Digital Tradition Folk Song Database. The URL is `http://web2.xerox.com/digitrad`. The `http` tells us that the document is a hypertext document (often written in the Hypertext Metalanguage–a hybrid of SGML) and the rest of the URL is the machine name and directory where the information resides. The folk song database is a good one to look at. Just browsing through the names of songs would be interesting, but much less useful than the ability to search for titles or words in songs. The server actually provides the ability to search through the database for song titles, keywords, or a full-text search on the words. Most servers that provide searching features use the WAIS software to implement the searches. (I will write more about WAIS and the Z39.50 search protocol in another issue. Z39.50 is becoming a standard within the library community so you may already be familiar with it from your own institutions electronic library). Once you have found the appropriate song you can either look at the words (which is invaluable when you are looking for the words to a song to placate the baby) or, in many cases, even play the tune. The ability to play the tune depends on the software and hardware that you are using. It works well for me on my Macintosh and HP workstations.

There have been articles about Mosaic in the New York Times, Washington Post, and even Science magazine.

Another example of a URL is `gopher://lib.stat.cmu.edu:70/1` which enables one to access the gopher server that presents StatLib information. There are many interesting places to look. One of my current favorites is the gopher server at the Library of Congress (`gopher://rs5.loc.gov:70/11/loc/events/online`) which has various online exhibits including, “1492: An Ongoing Voyage”, “Rome Reborn: The Vatican Library and Renaissance Culture”, and “Scrolls from the Dead Sea”. Each of these exhibits has text material as well as digitized copies of manuscripts and pictures.

Searching, as opposed to browsing, is one of the problems with the WWW. It is possible to search through the folk music archives and the ability to search within a particular server is very useful, especially when full-text searches are possible. Of course even with efficient software full-text searches can place a large burden on

the server. At busy times of the day any popular WWW servers are already beginning to show signs of stress. Global searches across all of the WWW world would be much more useful than local (to a server) searches, but also much less practical. How would one collect the information?

> ***One has to wonder how long this happy set of circumstances—free software, essentially free networking, and free access to servers—will continue.***

Various groups are starting to collect indexes of material, including "Joel's Hierarchical Subject Index" (URL `http://web2.xerox.com/digitrad`). This type of index will become much more important as the web continues to grow, subdivide, and become otherwise impossible for one person to follow.

How does it work?

Mosaic (WWW), WAIS, and gopher all depend on the Internet for global connectivity and on remote servers to provide the information. Beyond that there are a few relatively simple protocols that both the client and server software need to obey. At the moment the Internet is free (at least there are no usage fees for most users) and many organizations provide interesting material on their servers, for free. Other organization provide software such as Mosaic, gopher, WAIS for free. One has to wonder how long this happy set of circumstances—free software, essentially free networking, and free access to servers—will continue.

How can I try it?

Mosaic is available via anonymous FTP from NSCA `ftp.ncsa.uiuc.edu`. There are precompiled versions of X Mosiac for many popular Unix platforms as well as the full source code. NCSA also provides Macintosh and DOS/Windows implementations. If you have any of these types of computers connected to the internet, it is certainly worth spending a few hours exploring with Mosaic. Even over slow links (my home link was until recently a SLIP connection at 19.2 k-baud) the Mosaic software gives excellent performance.

Mike Meyer
Carnegie Mellon University
`mikem@stat.cmu.edu`

CONFERENCE NOTICES

Interface '94

26th Symposium on the Interface:
Computing Science and Statistics
Computationally Intensive Statistical Methods
June 15-18, 1994,
Research Triangle Park, NC,
Sheraton Imperial Hotel

The Interface Conference is the premier annual conference on the interface of computing and statistics. It is sponsored by the non-profit Interface Foundation of North America, and will be hosted in 1994 by SAS Institute, with John Sall as program chair. The 1994 conference will be in Research Triangle Park, the home of many top research centers, and anchored by three major universities.

Contact: Email: `interface94@sas.com` phone: 919-677-4499, fax: 919-677-8224, mail: Interface 94, SAS Campus Drive, Cary NC 27513 USA

Call for Papers:

The deadline for submitting a contributed paper is Jan 31. It may be a presentation, or poster session. Submit an abstract either by email (if you don't have messy formulas) or hardcopy formatted to 6.5 inches wide by at most 4 inches tall, with centered lines for title, author(s), and address.

Financial Support:

Limited funds may be available to support travel and per diem expenses of young researchers and graduate students. Preference will be given to those who will be presenting papers. Please apply to John Sall, program chair [`sall@sas.com`].

Adjoining Meetings in the Triangle:

Before the Interface is the Spring Research Conference on Statistics in Industry and Technology. After the Interface is the Third World Congress of the Bernoulli Society and the 55th Annual IMS meeting.

Hotel:

For reservations, call or write Sheraton Imperial Hotel, I40 Exit 282 at Page Rd, P O Box 13099, Research Triangle Park, NC 27709. Phone 919-941-5050.

Exhibits:

If you would like to exhibit, please contact Armistead Sapp, SAS Institute, Cary NC, 919-677-4499, sasaws@unx.sas.com.

Tours:

Several tours will be available on Thursday evening, one to SAS Institute, and one to the UNC Graphics and Image Lab (Virtual Reality Center).

Keynote:

- G.W. Stewart, Institute for Advanced Computer Studies, University of Maryland, a well-known authority in the field of numerical linear algebra.

Invited Sessions:

- *Space Filling Experimental Designs,* S. Stanley Young, Glaxo
- *Fast Implementations of Smoothers,* Steve Marron, University of North Carolina
- *Smart Monte Carlo Methods for Conditional Inference in Exponential Families,* Cyrus R. Mehta, Harvard Univ.
- *Nonparametric Regression for Edge and Peak Preserving* Alex Georgiev, Ethyl Corporation
- *Stochastic Modeling In Carcinogenesis,* Chris Portier, NIEHS
- *Convergence of Markov Chain Samplers,* Richard Smith, University of North Carolina
- *Panel of Editors of Journals for Statistical Computing,* Ed Wegman, George Mason University
- *Computational Techniques in Genetics and Molecular Biology,* Francoise Seillier-Moiseiwitsch, University of North Carolina
- *Efficient Bootstrap Computations,* Tim Hesterberg, Franklin and Marshall College
- *Neural Net Tutorials,* organized by Ron Gallant, University of North Carolina
- *Wavelets Tutorial,* Mary Ellen Bock, Purdue University
- *Gibbs Tutorial,* Adrian Smith, University of Nottingham
- *Computing for MetaAnalysis,* Bill DuMouchel, Harvard
- *Robust Regression and Multivariate Analysis,* David Rocke, University of California at Davis
- *Green Thumbs: Extensions and Applications of Tree Modeling Methods,* Sally Morton, Rand
- *Applications of Wavelets,* Iain Johnstone, Stanford
- *Bayesian Curve Fitting,* Mike West, Duke University
- *Longitudinal and Mixed Models,* Russ Wolfinger, SAS Institute
- *Statistics of Protein and Macromolecular Structures,* Peter Munson, National Institutes of Health
- *Panel: Statistics Education in the Computer Age,* Raoul LePage, Michigan State University
- *Issues in Software* Stephen Eick, AT&T Bell Labs
- Additional Tutorial Sessions. *Internet Facilities* Tim Arnold, N C State University, *Perl programming* Phil Spector, University of California at Berkeley.

John Sall
SAS Institute
sall@sas.com

Artificial Intelligence and Statistics

Fifth International Workshop on Artificial Intelligence and Statistics
January 4-7, 1995
Ft. Lauderdale, Florida, USA

This is the fifth in a series of workshops that has brought together researchers in Artificial Intelligence and in Statistics to discuss problems of mutual interest.

Format: To encourage interaction and a broad exchange of ideas, the presentations will be limited to about 18 discussion papers in single session meetings over the three days of the technical portion of the workshop (Jan. 5-7). Focussed poster sessions will provide the means for presenting and discussing the remaining research papers. Papers for poster sessions will be treated equally with papers for presentation in publications. Attendance at the workshop is not limited to presenters. The three days of research presentations will be preceded by a day of tutorials. These are intended to expose researchers in each field to the methodology used in the other field.

Language: The language will be English.

Topics Of Interest: We strongly encourage research papers in the areas of selecting models from data, integrated man-machine modelling methods, empirical discovery and statistical methods for knowledge acquisition, probability and search, uncertainty propagation, combined statistical and qualitative reasoning, inferring causation, quantitative programming tools and integrated software for data analysis and modelling, discovery in databases, meta data and design of statistical data bases, automated data analysis and knowledge representation for statistics, machine learning, and clustering and concept formation.

This list is not exhaustive and we encourage submissions in other areas at the interface of artificial intelligence and statistics as well.

Submission Requirements: Submissions will be extended abstracts (up to four pages). Submissions for discussion papers (and poster presentations) will be considered if postmarked by June 30, 1994. Abstracts postmarked after this date but before July 31, 1994, will be considered for poster presentation only. Please indicate the topic(s) addressed by your abstract and include an electronic mail address for correspondence. Acceptance notices will be mailed by September 1, 1994. Preliminary papers (up to 20 pages) must be returned by November 1, 1994. These preliminary papers will be copied and distributed at the workshop.

Submissions may be sent by air mail or email (latex documents preferred) to either chair:

Doug Fisher, General Chair
5th Int'l Workshop on AI & Stats
Department of Computer Science
Box 1679, Station B
Vanderbilt University
Nashville, TN 37235 USA
dfisher@vuse.vanderbilt.edu

or

Hans Lenz, Programme Chair
5th Int'l Workshop on AI & Stats
Free University of Berlin
Department of Economics
Institute for Statistics and Econometrics
14185 Berlin, Garystr 21 GERMANY
HJLENZ@fubvm.wiwiss.fu-berlin.de

JCGS UPDATE

Subscription Offer

Dear Section Member,

By joining other ASA members in the ASA Statistical Graphics or Statistical Computing Sections, you have expressed your interest in these fast-growing fields. The sections provide you with a broad-based program at the annual meeting, with the *Statistical Computing & Statistical Graphics Newsletter*, and much more that will help you stay abreast of these fast-moving fields.

We would like to call your attention to another way of staying current: the ASA sponsored Journal of Computational and Graphical Statistics. A joint venture with IMS and the Interface Foundation, this journal is a leader in both computing and graphics. It is a great way to learn about current work in our areas of interest.

JCGS is a journal for all of us and it is up to us to support it. You can subscribe using the coupon at a special section member's rate of $30 until February 28, 1994. If you already subscribe, the coupon can be used to renew for an additional year at the special rate.

Please also check with the library at your university or place of business and make sure they know the value of *JCGS*. We know that personal recommendations influence libraries; we would like to encourage you to talk with your librarian and describe the many benefits of access to the journal.

Forms for an individual subscription and for your institution's library subscription are included for your convenience. Please complete and send or FAX to the ASA for processing. Note the deadline for savings on individual subscriptions.

Rick Becker
Chair, Statistical Graphics Section

Sandy Weisberg
Chair, Statistical Computing Section

JCGS Section Member's Subscription Form

Special offer worth $10.00 off the regular subscription price.
(Good through February 28, 1994)

☐ **YES** I would like to subscribe to the *Journal of Computational and Graphical Statistics* at the special $30 rate for members of the Statistical Computing and Statistical Graphics sections of the ASA.

- ☐ New subscriber
- ☐ Renew for an additional year

Method of Payment
(all subscriptions must be prepaid)

- ☐ Check/Money Order payable to the **American Statistical Association** (U.S. funds drawn on a U.S. bank)
- ☐ VISA ☐ MasterCard ☐ Diners Club

Name _____
Organization _____
Address _____
City _____ State/Province _____
Zip/Postal Code _____ Country _____
Telephone _____ Fax _____

Name _____
Card Number _____
Exp. Date _____
Total Payment $ _____
Signature _____

Credit Card orders may be FAXED to: (703) 684-2037
Please return this order form with payment to:
The American Statistical Association/Subscriptions • 1429 Duke Street, Alexandria, VA 22314-3402
(703) 684-1221

JCGS Library Subscription Form

☐ **YES** My library would like to subscribe to the *Journal of Computational and Graphical Statistics.*

- ☐ $95.00 (one year – Second volume year, 1st year free.)
- ☐ $180.00 (two years – includes 1st year free.)

Method of Payment
(all subscriptions must be prepaid)

- ☐ Check/Money Order payable to the **American Statistical Association** (U.S. funds drawn on a U.S. bank)
- ☐ VISA ☐ MasterCard ☐ Diners Club

Name _____
Organization _____
Address _____
City _____ State/Province _____
Zip/Postal Code _____ Country _____
Telephone _____ Fax _____

Name _____
Card Number _____
Exp. Date _____
Total Payment $ _____
Signature _____

Credit Card orders may be FAXED to: (703) 684-2037
Please return this order form with payment to:
The American Statistical Association/Subscriptions • 1429 Duke Street, Alexandria, VA 22314-3402
(703) 684-1221

December Contents of JCGS

The December issue features the Invited Article, "A Model for Studying Display Methods of Statistical Graphics," by William S. Cleveland of AT&T Bell Laboratories. This paper was initially presented at an Invited Paper Session at the 1993 Joint Statistical Meetings in San Francisco this past August. A model has been developed to provide a framework for the study of visual decoding. This model consists of three parts: (1) a two-way classification of information on displays—quantitative-scale, quantitative-physical, categorical-scale, and categorical-physical; (2) a division of the visual processing of graphical displays into pattern perception and table look-up; (3) a specification of visual operations that are employed to carry out pattern perception and table look-up.

Commentary on the Invited Article is provided by the fol-

lowing discussants: Lothar Tremmel of Bio-Pharm Clinical Services, Susan Holmes of the Biometry Unit, INRA, Montpellier, France, and Leland Wilkinson of SYSTAT, Inc. Tremmel argues that the merit of Cleveland's article is the "convincing demonstration that not 'virtually any method of display suffices' for graphing statistical data." Holmes discusses two separate brain functions which she labels left-brain and right-brain functions. Wilkinson concludes that "by placing Cleveland's model in the context of the more general information processing model favored by most psychologists today, we can help to understand how and why distortions occur in the perception of graphs." Cleveland concludes his response with, "But I have not yet succeeded in finding a convincing application of the work in cognitive psychology that conjectures a full prescription of all of the processes that lead from the retinal image to conclusions about visual scenes."

In the article, "The Plot-Data Interface in Statistical Graphics," Catherine Hurley of George Washington University states that the multiplicity problem caused by many plot varieties and many data representations is avoided by constructing a plot-data interface. The interface is a convention by which plots communicate with datasets, allowing plots to be independent of the actual data representation. This paper describes the components of such a plot-data interface. The same strategy may be used to deal with the dependence of model-fitting procedures on data.

In "Confident Search," Paul R. Rosenbaum of the University of Pennsylvania begins with an arbitrary heuristic search procedure and supplies it with a confidence statement of the following form: With specified high probability β, the output of the confidence procedure will be among the best 100 α % of the elements of P. The confidence procedure will report either the outcome of the heuristic search or a better alternative with the required properties; that is, it will either certify that the heuristic answer has the desired confidence property, or it will produce a better answer having the property. The approach involves combining a heuristic search with a form of heuristic sampling that tends to sample the better elements of P.

Xing Sam Gu, C.S. First Boston Pacific, and Paul R. Rosenbaum, University of Pennsylvania, discuss "Comparison of Multivariate Matching Methods: Structures, Distances, & Algorithms." A comparison and evaluation is made of recent proposals for multivariate matched sampling in observational studies, answering questions in the three areas of Algorithms, Structures, and Distances. Three recent proposals are compared. Practical advice is summarized in a final section.

NEWS CLIPPINGS

Interface '93 Meeting

The 25th anniversary Symposium on the Interface, Computing Science and Statistics, was held in San Diego April 14-17, 1993. David R. Brillinger, of UC Berkeley, delivered the keynote address on "Some Examples of Statistical Analysis and Computing in Science." He provided stimulating examples of graphical analysis providing insights into the data analysis and interpretation which would otherwise go unobserved.

David Scott, Bill Eddy and Jim Rosenberger participated in the Interface '93 harbor excursion.

Joint Mixer at the Annual Statistical Meetings

by Linda Clark and Jim Rosenberger

The Statistical Computing/Graphics Business Meeting and Mixer in San Francisco, August 8-12, was a rousing success. Over 150 people attended and enjoyed several hours of refreshments, door prize drawings, and visiting with other section members.

We'd like to thank the following companies for sponsoring the mixer:

- BBN
- BMDP
- Cytel Software Corp.
- Data Description Inc.
- Minitab Inc.
- SAS
- SPSS Inc.
- Statistical Graphics Corp.
- StatSci
- Systat, Inc.
- Visual Numerics Inc.

Linda Clark and Rick Becker handing out tickets for the random drawing for door prizes.

The following companies donated books and/or software for door prizes, and the grand prize was a book on wine and a bottle of California wine.

- Chapman & Hall
- Academic Press
- Addison Wesley
- American Math. Society
- Birkhaeuser
- Cambridge U. Press
- Duxbury
- Irwin
- Jandel Scientific
- J. Wiley
- Macmillan
- Marcel Dekker
- Oxford
- Power Thinking Tools
- Smith Hanley
- Springer Verlag
- W H Freeman
- SAS

The usual activities could be found at the Annual Meetings held in San Francisco, August 8-12, 1993. The photographs below show new and old members at the mixer following the joint computing/graphics sections' business meeting.

Mr. StatLib – Mike Meyer at his best!

Deborah Swayne of Bellcore, watching a comical sit-com on XGobi.

Trevor and Daryl of AT&T confer with Magdalena of NIST on colorful graphical issues with impact.

Sandy Weisberg, Statistical Computing Chair, Jim Rosenberger, Statistical Computing Newsletter Editor, David Scott, Statistical Graphics Program Chair, and Mike Meyer, Statistical Graphics Newsletter Editor, sporting the tee-shirts given to past contributors to the Newsletter.

SECTION OFFICERS

Statistical Graphics Section - 1994

Roy E. Welsch, Chair
617-253-6601
Massachusetts Institute of Technology
rwelsch@sloan.mit.edu

David W. Scott, Chair-Elect
713-527-6037
Rice University
scottdw@rice.edu

Richard A. Becker, Past-Chair
908-582-5512
AT&T Bell Laboratories
rab@research.att.com

William DuMouchel, Program Chair
617-489-2631
dumouche@jimmy.harvard.edu

Sallie Keller-McNulty, Program Chair-Elect
913-532-6883
Kansas State University
sallie@cecil.stat.ksu.edu

Michael M. Meyer, Newsletter Editor
412-268-3108
Carnegie Mellon University
mikem@stat.cmu.edu

Linda A. Clark, Secretary/Treasurer
908-582-4807
AT&T Bell Laboratories
lac@research.att.com

Howard Wainer, Publications Officer
609-734-5623
Educational Testing Service
Princeton, NJ 08541
hwainer@rosedale.org

Jane F. Gentleman, Rep. to Council of Sections
613-951-8213
Canadian Centre for Health Information
Ottawa, ON K1A OT6, CANADA
GENTLEJF.NRCVM01.bitnet

Sally C. Morton, Rep. to Council of Sections
310-393-0411
The Rand Corporation
Santa Monica, CA 90407-2138
Sally_Morton@rand.org

Statistical Computing Section - 1994

Trevor J. Hastie, Chair
908-582-5647
AT&T Bell Labs
trevor@research.att.com

Mary Ellen Bock, Chair-Elect
317-494-6053
Purdue University
mbock@stat.purdue.edu

Sanford Weisberg, Past Chair
612-625-8777
University of Minnesota
sandy@stat.umn.edu

Sallie Keller-McNulty, Program Chair
913-532-6883
Kansas State University
sallie@cecil.stat.ksu.edu

John A. Rice, Program Chair-Elect
510-642-6930
University of California at Berkeley
rice@stat.berkeley.edu

James L. Rosenberger, Newsletter Editor
814-865-1348
The Pennsylvania State University
jlr@stat.psu.edu

John Sall, Secretary-Treasurer
Sas Institute
sall@sas.com

Karen Kafadar, Publications Liaison Officer
301-496-8556
National Cancer Institute
kk@helix.nih.gov

Daryl Pregibon, Rep. to Council of Sections
908-582-3193
AT&T Bell Labs
daryl@research.att.com

Russell Lenth, Rep. to Council of Sections
319-335-0814
University of Iowa
rlenth@stat.uiowa.edu

by Andrejs Dunkels

INSIDE

Statistical COMPUTING & GRAPHICS

The *Statistical Computing & Statistical Graphics Newsletter* is a publication of the Statistical Computing and Statistical Graphics Sections of the ASA. All communications regarding this publication should be addressed to:

James L. Rosenberger
Editor, Statistical Computing Section
Department of Statistics
The Pennsylvania State University
University Park, PA 16802-2111
(814) 865-1348
JLR@stat.psu.edu

Michael M. Meyer
Editor, Statistical Graphics Section
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213-1380
(412) 268-3108
mikem@stat.cmu.edu

All communications regarding membership in the ASA and the Statistical Computing or Statistical Graphics Sections, including change of address, should be sent to:

American Statistical Association
1429 Duke Street
Alexandria, VA 22314-3402 USA
(703) 684-1221 • FAX (703) 684-2036
ASAINFO@ASA.MHS.COMPUSERVE.COM

PENNSTATE

1855

Department of Statistics
University Park, PA 16802-2111

Nonprofit Organization
U. S. POSTAGE
PAID
Permit No. 1
University Park, PA 16802

Published by the Penn State Department of Statistics
326 Classroom Building, University Park, PA 16802-2111