STATS

STATS

The Magazine for Students of Statistics

Fall 2002 • Number 35

STATS: The Magazine for Students of Statistics (ISSN 1053-8607) is published three times a year, in the winter, spring, and fall, by the American Statistical Association, 1429 Duke St., Alexandria, Virginia 22314-3415 USA; (703) 684-1221; fax: (703) 684-2036; Web site: *www.amstat.org*

STATS is published for beginning statisticians, including high school, undergraduate, and graduate students who have a special interest in statistics, and is distributed to student members of ASA as part of the annual dues. Subscription rates for others: $13.00 a year to members; $20.00 a year to nonmembers.

Ideas for feature articles and material for departments should be sent to the Editors; addresses of the Editors and Editorial Board are listed above.

Requests for membership information, advertising rates and deadlines, subscriptions, and general correspondence should be addressed to *STATS* at the ASA office.

Features

Departments

Editors' Column

Beth Chance **Allan Rossman**

Statistics often make headlines. Presidential popularity polls, unemployment rates, stock market returns, and heart attack rates for people who do and do not take aspirin regularly are but a few examples of statistics that find prominent display in the popular press. But do the headlines and popular articles do a credible job of representing the findings from research studies? In the lead article of this issue, Christopher Paul offers "how to" advice for digging behind the headlines to find the original reports on which to judge the accuracy of the media representation. He also presents a helpful list of questions that all students of statistics and all educated consumers of statistical information should ask. He illustrates his advice with a case study involving teenage consumption of alcohol that generated some media controversy.

How did you spend your summer vacation? Marilisa Gibellato and Kristin Duncan describe the experiences of a group of students who spent their summer immersed in a "Research Experience for Undergraduates" program at Ohio State University. The students participating in this program gained first-hand knowledge of the challenges, rewards, and just plain fun involved in conducting collaborative scientific research, and many have been inspired to pursue graduate study in statistics or biostatistics. The students happily refer to the program as "Stats Camp," suggesting that their learning experience was an enjoyable one. We hope that reading this article will help you to decide if such an experience might be beneficial for you.

Can you imagine spending more than fifty years in the same job? We certainly can't (we write this as we begin our fourth and second years at our new universities), and we suspect that the notion sounds especially amazing to students just thinking about beginning their careers. Bob Hogg (rhymes with "rogue") enjoyed a 51-year career at the University of Iowa, where he became a legend in the statistics community. Students know of Bob through his classic textbooks on mathematical statistics (do Hogg & Craig or Hogg & Tanis ring a bell?). In this issue Jackie Dietz presents an interview with Bob in which he recounts how he became interested in statistics, describes some of the professors and students with whom he has worked, and offers advice to current students of statistics. During the course of the interview Jackie and Bob even discovered that they are academic relatives of each other!

On the other end of the spectrum we present a "Day in the Life" segment from Paul Lupinacci, a recent Ph.D. graduate just beginning the third year of his career at Villanova University. Paul offers a glimpse of the range of activities that can occupy the daily life of an assistant professor. He also offers some advice for statisticians who reside in a Department of Mathematics and describes a mentoring program called Project NExT (New Experiences in Teaching) that he has found very valuable.

Robin Lock tackled football (sorry for the pun!) in his first installment of *The Statistical Sports Fan*, and he analyzed the four major North American team sports in his second feature. In this issue he turns his attention to the world's sport, soccer (or the real football, for our international readers). He analyzes times until the first goal scored in last summer's World Cup games and compares the fit of exponential and Weibull probability models to those data.

Chris Olsen addresses the perception problem that some statisticians feel about their chosen profession and offers advice for promoting our profession, useful for cocktail-party small-talk, in this issue's *μ-sings*. Gretchen Davis offers help for understanding the tricky concept of influential observations in this issue's AP Statistics column, illustrating the point with historical data from California missions.

This issue's *Data Sleuth* feature was submitted by Dan Teague, who asks readers to analyze some data on operating costs of aircraft and to explain an anomaly that arises. Students who took the Advanced Placement exam in 2002 may recognize the data as having supplied the context for one of the questions on that exam.

As always, we welcome your comments and suggestions for helping STATS to appeal to statistics students of all interests and ages.

Finding the Findings Behind the News

Christopher Paul

In this article I discuss the very real challenge of trying to assess research findings as presented in the popular press. The popular news media, be it television, print, or the web, abounds with interesting and titillating "findings" from studies done all over the world. When you read or pay close attention to the research as it is presented in the news, it is often hard to tell what the researchers actually did. This makes it hard for us, as consumers of statistical information, to evaluate their research design and decide for ourselves whether or not we accept their findings. Below I offer two things: first, a list of questions that, if you can satisfactorily answer, will allow you to evaluate almost any kind of research; and second, some suggestions on how to go about answering those questions starting with something as flimsy as a brief newspaper article or web headline.

Assessing Research Findings

In my efforts to be a good consumer of research (quantitative or otherwise), I always try to answer five questions whenever I seriously consider a piece of scholarship. These five questions are directly derivative of Maurice Zeitlin's (2000) "four questions," which I first encountered during my graduate study in the UCLA Department of Sociology. If you can answer these five questions, then chances are you have carefully read and understood an article or research presentation, have a good assessment of it, and are ready to talk or write about it.

1) *What is the research question?*
All research is trying to answer some question. Good presentations make clear what the question is within the first two paragraphs. Media reports often skip the question and go straight to the findings that the reporter finds interesting, often divorced from the researcher's original line of inquiry. Knowing what the actual central research question was is a big step toward understanding the research.

2) *What is the originating question?*
Where did this line of inquiry come from? Is there some social significance to the question? A problem to be solved? A policy to be evaluated or advocated? Is there a theoretical model being tested, or is it a question that is the logical next step in an ongoing strand of research? Understanding what motivates the research in the broadest possible way may help you make sense of things that might otherwise seem odd about the work or its findings, and may also help you maintain an appropriate skepticism if the research is clearly part of some kind of political or business "agenda."

Note that the *researcher's* originating question is the reason the *researcher* undertook the research, which may have absolutely nothing to do with why some *reporter* chose to write an article about that researcher's findings. While it is usually pretty easy to figure out why a reporter chose a topic (it is interesting to his audience and might help sell papers), it is often much more difficult (and more valuable) to figure out what a researcher's originating question was.

3) *What is their answer?*
In a good presentation, the answer follows the question. Equally frequently, the answer will be fairly clear, and the question can be inferred from it. This can be rephrased as "what did they actually find?"

4) *What evidence do they offer to support their answer?*
This is pretty self explanatory, and is where "the rubber meets the road," so to speak. Here is where all the methodological information you've learned gets tested out. What was their study design? Their

Christopher Paul (cpaul@rand.org) is an associate sociologist at RAND, presently working out of the Pittsburgh office. He considers himself an "amateur" statistician, having been formally trained as a sociologist but having a considerable amount of hands-on methodological and research experience. His teaching experience has all been at UCLA, in both the department of statistics and the department of sociology.

method of data collection? Their population? Their sample? Did they effectively control for confounding factors? Note that this general question (like all of these questions) is not constrained by disciplinary boundaries. This is an equally valid question to ask of the results of a clinical trial or an historical argument; "evidence" is defined differently in different disciplines, but still needs to be reflected on in all of them.

5) *Do you accept that evidence?*
This final question is an invitation to think, evaluate, make a judgment, and form an opinion. Based on the evidence you have just enumerated (and not based on a general sense that published research must be correct), do you accept their answer?

I find that these five questions give a great foundation from which to begin talking or thinking about a finding or set of findings. Unfortunately, virtually no news article contains enough information to satisfactorily answer the five questions. With the five questions as a "gold standard," what can be done?

Hunting Down the Answers

The answer is simple: the execution, less so. In order to find the answers to the five questions you need to track them down. The news article by itself likely just has tantalizing clues as to what the answers are, but does not provide sufficient detail to answer that fifth question, "Do you accept that evidence?" To answer those questions, you will need to go beyond the newspaper article. Tracking down and evaluating the source is a skill, and a skill worth having. Fortunately, in this, the information age, it is a considerably less daunting task than it once was. If you can identify the researchers and their institute, organization, or agency, you can find their web page, see their press release, and, more often than not, view the full text of their report. While research reports often run over 100 pages, after developing some skill at evaluating a table of contents, it is remarkable how few of those 100+ pages you actually have to look at in order to answer the five questions. Here I hope to demonstrate the process with a single example. This demonstration will not provide you with a recipe for tracking down five answers for every news presentation, but it should give you some idea of how to try to go about it, and an idea about how much effort is likely to be required.

An Example From My Own Reading

In this section, I describe the general process of tracking down research in the context of a specific example. While I use a specific (and reasonably interesting) article from my own reading as an example, this general approach will work on almost any news article reporting research findings.

Step 1: The original article

Articles discussing research results often present some core claim or finding without any of the information necessary to substantiate that claim or finding. Fortunately, they also usually present *whose* finding it is. This makes it possible to get more information, but before moving on, it is useful to spend some time puzzling over the newspaper article itself to see what can be learned. Trying to answer the five questions just from a summary of findings in a newspaper can be frustrating, but can be an interesting exercise. The five questions take for granted that you know who is asking the research question. However, in a newspaper article, there can be two lines of inquiry. The first is the one we want, the researchers who actually did the study. The second is the reporter (and the newspaper they work for). What about these findings is newsworthy? Does the newspaper article present the findings the researchers thought were most important, or the ones the reporter thought would be most interesting to his/her audience? Newspapers usually have different originating questions than the ones that actually drove the research, and may disagree with the researchers about what the central question was. This can make making sense of the answers, to the extent that they are given, potentially problematic.

The article I want to talk about is one that came across my Yahoo! news headlines on the Reuters wire on February 26, 2002: "Teens drink quarter of all alcohol consumed in US." I had been talking about binge drinking among college students in one of my classes, so I printed the article out and filed it for later examination, without really reading it. The next day (the 27th) I saw the same piece of research being reported both in the *Wall Street Journal* and in the *UCLA Daily Bruin*. I captured both of these, as I like to show students how different media write entirely different articles based on the same released findings. At this point I actually read the articles, and was intrigued to find a controversy right in the article! The reported study, conducted by a team of researchers at the National Center on Addiction and Substance Abuse at Columbia University, had reported that 25% of all alcoholic beverages were consumed by teens, but the Distilled Spirits Council disputed the findings and claimed that, due to a methodological flaw, the correct percentage was something like 11-12%. This controversy demanded further attention. Who was right? Any time you want to know more than the news article tells you, you are going to have to look beyond it. Fortunately, that isn't that hard.

Examining all three articles together, I noticed that the same facts were reported differently in each. I made a list of core claims and information about the study as presented in the news pieces so I would be able to proceed. I learned that the National Center on Addiction

and Substance Abuse at Columbia University was the source of the report, and that the most publicized claim was that 25% of all alcoholic beverages consumed in the U.S. were consumed by teens. I was unable to tell, however, if that claim was *the* answer to the central question of the report, or if it was just *an* answer to a question asked in the report, seized upon by the press and opposed by the liquor industry claim of 11 or 12%.

Step 2: Going to the source

Once as much information (and as many questions as possible) have been gathered from the newspaper article, it is time to dig: Take the name of the center, organization, or study group (in our example "National Center on Addiction and Substance Abuse at Columbia University") and type it into your favorite internet search engine. You will usually find a link to their homepage. The National Center on Addiction and Substance Abuse homepage is at *www.casacolumbia.org/*. On the homepage you will likely find many potentially useful items. Generally, an organization will have something like an "about us" link that provides the background for the organization, information that may help explain their motives and define their originating questions. Also on the front page there will usually be something like "what's new" or "recent releases" or something that will have a link to the information you want. It may, however, be buried in "press releases" or "publications." If you don't see what you want immediately, try one of those.

Entering the above url took me to the National Center on Addiction and Substance Abuse homepage. Right on their main page there was a link to CASA REPORT ON UNDERAGE DRINKING (which is, as of this writing, now under the "Newsroom" link). Since the center's name told me everything I really needed to know about their motives and reason for being, I followed that link right away. The link brought me to the press release.

Usually there will be a link to the relevant press release somewhere on an organization's main page. You can see pretty easily how everything that was mentioned in the news came straight from the press release. It is noteworthy when there is much more information in the press release than was in the newspaper, which shows that the news reporters were pretty selective about what they chose to report. Usually, based on the press report, you can answer questions 1, 2 and 3 of the five questions, deducing the originating questions, the central questions, and the answers. Maybe you can make a start on question 4, about evidence, but usually you only get vague methodological information that does not tell you enough about the sample design or the data collection to really assess the evidence.

On the CASA press release, I found all kinds of interesting information, including the somewhat quizzical information that their survey was a random sample of 900 adults (how do they know anything about teen drinking from adults?) *and* that they re-analyzed existing data from five other surveys (Ah-ha! some of those must be about teens). I found quite a few interesting claims about the drinking habits of children, a major section called "A CASA Checklist for Parents" and another called "Recommendations for Policy Makers, Educators and Prevention Experts." Just these headings, without actually looking at what's under them, gave me some clear insight into what the goals of this report are – to encourage policy that decreases potentially harmful alcohol use among teens. Interestingly enough, the highly contentious "25% of all drinks go to teens" claim, that headlines the newspapers, is hidden.

Normally after examining the press release I would be ready to go to the report itself to answer the last two of the five questions. There is usually a link to the *full report* or something similar. In this case, however, I had the "25% of drinks consumed by teens" controversy in the forefront of my mind. I saw at the bottom of the press release a link to the full report and another link to CASA's statement on the release of the report. Since I knew the report itself would be large and potentially difficult to pull good information out of, I decided to give the latter a try first. In this separate statement they admit that the 25% figure is that 25% of the *drinks consumed by their sample* were consumed by teens, but the study those numbers are based on has a teen oversample, and if you weight the sample (as you must) you end up with a figure closer to 11 or 12%. So, the liquor industry was right! The 25% claim is totally bogus. Did that keep it off the Reuters wire? No. Did I know that it was a false claim until I went to the trouble of looking it up? No.

Step 3: Examining the report

Now, in my example, we've already reached the climax. Since they debunked their own claim (and it wasn't that important of a claim to them anyway), we may not feel compelled to examine their evidence. However, in more conventional explorations of findings, you would still be wondering: what evidence have they got? Is this a well executed and designed study, or not? Now is when we take the plunge and click on *the report*.

When I did this, I saw that the report is over 100 pages (152 printed pages in the CASA report) and it is in hypertext .pdf format (some .pdf documents are fully searchable and often indexed as well; others are just images and can only be searched visually). Fortunately, I only really needed about five of those 152 pages, and the report writers made it convenient for me to find them. The first page is the title page, followed by some acknowledgements, neither of which is useful to us. Then, finally, the first treasure: the table of contents. The table of contents will tell us where to look for what

we want to know. Now, even though this is the roadmap to what we want, it is still rather daunting. The table of contents usually fills two or three pages. Most of it is stuff you don't care about, or, at least, don't care about at this phase. Generally there are two things I look for: the abstract or executive summary (useful if you want a good quick summary of what they think is important in their findings), and an appendix about methods. In the CASA report, sure enough, there is Appendix C – Survey Methodology, beginning on page 109. Quickly scrolling down to that, I found that it details, in a wholly adequate way, how they conducted their survey of the 900 adults. At that point I saw and accepted their evidence for claims based on adult attitudes. What about the 25% cum 12% claims? How can I evaluate those data? Well, looking back at the table of contents, I also saw Appendix A – Survey Descriptions. Scrolling to that, I found what I was looking for: a detailed description of the other studies CASA pooled for their report, and the organizations that conducted them. If I were still curious, I had at that point identified other organizations whose homepages I could visit, and repeat the process.

Step 4: Conclusions and take-away

It took me less than 10 minutes to track down everything I needed to answer the five questions. I solved my riddle and got adequate answers to the five questions, all with a few mouse clicks. I believe that you could do this too. I suspect your internet skills are better than mine, based solely on generational advantage. I've just described how these things are generally organized in terms of the agency or organization web page, the press release, and the actual report. All that is left is for you to get out there and start asking, and answering, those five questions. The next time you hear or see a news agency report research findings that are amazing, interesting, or incredible, don't just take their word for it or remain uncertain. Follow it up and find the findings for yourself!

References

Daily Bruin Wire Service (2002), "Underage Drinking an American Epidemic," *Daily Bruin*, 2/27/02.

Soares, C. (2002), "Teens drink quarter of all alcohol consumed in US," Washington: Reuters, 2/26/02

Wall Street Journal Staff Reporter (2002), "Underage Drinking Study Has Liquor Industry Riled," *Wall Street Journal,* 2/27/02.

Zeitlin, M. (2000), "The Four Questions Elaborated," Los Angeles: University of California.

Web Resources

Here are some of the websites visited in my search. Unfortunately, the articles on the news websites are no longer available.

National Center on Addition and Substance Abuse: *http://www.casacolumbia.org*

CASA Report on Underage Drinking: *http://www.casacolumbia.org/newsletter1457/newsletter_show.htm?doc_id=103334*

CASA Statement on the release of the report:

http://www.casacolumbia.org/newsletter1457/newsletter_show.htm?doc_id=103428

Teen Tipplers: America's Underage Drinking Epidemic (full report): *http://www.casacolumbia.org/user_doc/Underage1.pdf*

"Stats Camp" Research Experiences for Undergraduates

Marilisa Gibellato and Kristin Duncan

From the Editors

Research Experiences for Undergraduates (REU's) are programs sponsored by the National Science Foundation to encourage undergraduate students to join research projects during the summer (*www.nsf.gov/home/crssprgm/reu/start.htm*). These are held at many sites around the country. Each site consists of a group of ten or so undergraduates, who work in the research programs of faculty at the host institution. The following article was written by two graduate students who provided support to the REU program in applied statistics and biostatistics at Ohio State University this summer. To learn about other REU opportunities in statistics and other mathematical sciences, you can consult the American Mathematical Society site at: *www.ams.org/employment/reu.html*.

Marilisa Gibellato (mgg@stat.ohio-state.edu) is in her fifth year of studies as an MD-PhD student at The Ohio State College of Medicine and Department of Statistics. She has completed the first two years of medical school and is now beginning her fourth year in the stats department. She holds a BA in mathematics from the US Naval Academy (1996), an MA in Natural Sciences–Pathology from Cambridge University (1998), and an MS in Statistics (2002). Her main career objective is to become a research physician for the Navy with the ability to act as a statistical consultant and/or primary investigator.

Kristin Duncan (blenk.2@osu.edu) is currently a fourth year graduate student in the Statistics Department at Ohio State where she serves as a Graduate Teaching Associate. She earned her Bachelor's degree in Mathematics from the University of Dayton in 1999 and her Master's degree in Statistics from Ohio State in 2001. As an undergraduate she participated in the Carleton/St. Olaf Colleges Summer Mathematics Program for Women and in the REU in Industrial Mathematics and Statistics at Worcester Polytechnic Institute.

Introduction

Eight undergraduate students from six different universities stand next to the restaurant all wearing red shirts that read "2002 REU in Statistics and Biostatistics, Ohio State University." One turns around to check that the sign for the "Stats Café" will be centered in the picture the photographer is about to take. As she turns, the back of her shirt is visible. It sports a graph and equation for the gamma distribution and reads "We're far from normal." It's hard to say what the people passing by on the street think of this scarlet posse… but in the Ohio State Department of Statistics these students are thought of as valuable researchers, potential statistics graduate students, and the most entertaining infusion of creativity and enthusiasm that we get to experience all year.

Research Experiences for Undergraduates at OSU

The Research Experience for Undergraduates in applied statistics and biostatistics has been running now for two summers at Ohio State University under the direction of Professors Doug Wolfe, Haikady Nagaraja, and Stan Lemeshow, with the assistance of two graduate students — us! This program attracts talented undergraduate students between their junior and senior years who are considering graduate school in statistics or a related field. These students attend classes and complete an eight-week data collection and analysis project under the joint tutelage of a statistics mentor and a scientific advisor. The experience is not only rewarding in an intellectual capacity, but also in many other ways.

Room and board are provided to the students who live in their own campus house for the duration of the program. The house has a communal kitchen and living areas, internet access, phones, and laundry facilities. It acts as the home base and the social epicenter. Each

student receives an OSU debit identification card that can be used at dining halls and various restaurants around campus. So students don't even have to use the kitchen if they would rather eat soggy broccoli or deep fried cuisine all summer. On top of all of this, each student receives a generous stipend.

The program presents a busy summer that is also filled with camaraderie and social activities. We graduate student assistants act as tour guides/ social coordinators/ Cedar Point trip organizers (Cedar Point is the most amazing roller coaster park on earth) and general Columbus, Ohio experts. We try to make sure that everyone has fun and that not all of the students' time is spent in the laboratories or staring at computer screens or trying to find the third moment of the Cauchy distribution (don't bother — there is none!). The students also plan events themselves and have been known to attend local concerts, play on the departmental softball team, organize ultimate Frisbee games against the physics REU students on campus, and generally enjoy the numerous activities in the young and vibrant city of Columbus, Ohio.

Of course the entire summer is not merely fun and games. The official academic program runs from mid-June to the beginning of August. The eight weeks is comprised of statistics courses, learning the science behind proposed projects, meetings with project mentors, weekly group meetings, and the completion and presentation of the projects.

The students choose to attend one or both of the following: an early-start mathematical statistics class offered to incoming statistics graduate students to prepare them for the Ph.D. level course offered in the fall (referred to by some as "easy start"), or an applied biostatistics course that focuses on statistical reasoning, the use of statistical software, data management, and the performance of standard statistical analyses. Students select the courses according to their backgrounds and abilities. No grades are given in the courses, so the students are able to learn without the pressure of the normal school year (although the motivated REU students usually choose to turn in homework assignments with pleasure). This also allows them to focus on their primary objective of producing and presenting their final projects.

The projects originate from faculty research in many different departments. Upon the students' arrival, their scientific advisors welcome them into their laboratories, clinics, and offices as members of a research team, while the statistics mentors guide them through the analyses and technicalities of the statistical tools required. As the graduate student support, we provide statistical assistance as well as scientific references and help. We serve as the first respondents to participants' queries.

The oil that keeps the cogs of the REU program turning smoothly is the weekly meeting. All REU students, the graduate student assistants, and the directors attend these meetings. They provide a healthy forum for discussing all topics pertinent to the students' experiences as well as providing the opportunity to describe progress, present interesting questions, identify subjects on which short courses are needed, and plan future social events. Some short courses given in response to requests from the students have been on logistic and Cox regression techniques. Also, lectures in UNIX and the S+ computing language were subsequently offered to aid students in the analysis portions of their projects.

The culmination of a summer's worth of hard work is the all-day mini-conference at which the students give oral presentations and display scientific posters. Most of the mentors also attend to support their students. The presentations are very impressive and reflect the time, energy, and enthusiasm that the students have for their projects. In fact, the quality of the talks is usually so high that we encourage the REU students to present their projects at their home institutions and at professional scientific meetings. There also have been opportunities for some students to become authors on related scientific publications.

Conclusion

We hope that this REU program demonstrates to undergraduates that there are many interesting opportunities for research in statistics and biostatistics and encourages them to continue their study at the graduate level. However, we are confident that they will all be successful in whatever paths and careers they choose. The nine graduating REU students from the first summer (2001) of our program all began graduate school in the fall of 2002. Eight have enrolled in statistics or biostatistics programs and one began a mathematics graduate program — all of them at top schools.

So here is the part where we try to recruit our next group of wonderful REU students… If you are presently a junior at an undergraduate institution studying math, statistics, biology, or some other technical major and are interested in trying out a summer of statistics (and getting paid a healthy stipend in the process) — then give it a shot! Meet this year's REU students and read about their projects in Box 1, and see what they have to say about the experience in Box 2. The application deadline is in February and the details are on the OSU statistics website: *www.stat.ohio-state.edu.* This is a great opportunity for the academically minded, and the experience will look impressive on any graduate school application.

Box 1: REU Student Biographies and Project Summaries

My name is Melissa Ludack and I'm from Ashland, Wisconsin. I'm currently finishing up my last year at the University of Wisconsin-River Falls where I am studying math and biochemistry. Upon graduation I'm planning on working toward my Ph.D. in statistics. My project this summer involves understanding how Drosophilae (fruit flies) behave when they experience different taste sensations. I'm working with several scientists in the entomology department who are trying to become the first to develop a taste assay for the fruit flies. I'm checking that the experimental design is appropriate with the information they are trying to obtain.

I'm Katie Pichotta from Waukegan, IL. I am a math major with a statistics concentration at St. Olaf College in Northfield, MN. My plans for attending graduate school have been undeterred by the fact that I do not yet have an idea of what I would like to study. My summer project involves analyzing the results of an interaction study between varying concentrations of a non-Cox-II inhibitor and a soy isoflavone, both of which have previously been shown to reduce the number of living mouse bladder cancer cells.

My name is Shannon Fraker. I attend school at Virginia Polytechnic Institute and State University (Virginia Tech). I have been a math major since my freshman year and recently added a second major in statistics. I am doing a unique combined program for my Master's in statistics. My project is on the influence of nutrition in the decreased lung function of patients with HIV. The nutrition factors we are looking at are the Percent Predicted Body Mass Index, Total Protein, and Albumin levels. We are also examining specific white blood cell counts in the body and in the lungs themselves. We are performing the analysis using scatterplots, correlations, multiple regressions, and linear models.

Hi, I'm David Kadonsky. I will be finishing my undergraduate work at the University of Wisconsin-Madison in the spring of 2003 with majors in Economics and Mathematical Statistics. I plan to pursue graduate studies in either Mathematical Statistics or Economics. I am working with Dr. Elizabeth Stasny and Dr. Paul Robbins on a study from Dr. Robbins' recent survey involving lawn care chemical use and environmental views. I have been learning and using categorical data analysis techniques for analyzing this survey data.

Folks □round here call me Spongebob, but usually I'm known as Peter Sprangers. I'll be finishing up my undergraduate at St. Olaf College this year and leaving with a B.A. in Mathematics, Latin, and a concentration in Statistics. Next year, I plan to go to graduate school to earn my Ph.D. in Statistics. For my research project I analyzed a data set containing nutrient counts, interleukin (an immune response cell) counts, and Perceived Stress Scores, in the hopes of finding correlations between these things. I've had a great time.

I'm Jeremy Strief. There is very strong statistical evidence ($p<.0001$) that I was born in Des Moines, Iowa. After a typical childhood of school, sports, and t-tests, my studies led me north to St. Olaf College in Northfield, MN, where I am currently pursuing a bachelor's in mathematics and religion. I'm planning on going to grad school in biostatistics for either an M.S. or a Ph.D. My summer project examines the effects of psychological education upon children with mood disorders. The hope is that proper psychological education will make both child and parent more savvy consumers of mental health services.

My name is Susan Hunter and I'm from Raleigh, North Carolina. I attend North Carolina State University in Raleigh where I am a Statistics major with minors in Spanish, Physics, and Math. I have been accepted into a 5-year BS/MS (Master of Statistics) program at NCSU and eventually plan to enroll in a Statistics PhD program. For the summer, I am working under John Orban, Research Leader at Batelle Memorial Institute, on a project to minimize truck engine emissions to meet new EPA standards. The aim of my project is to provide an optimal map of EGR settings that minimizes harmful emissions at specific engine speeds and torques.

I'm Emily Johnson. I am a Math and Social Science major from Dartmouth College in Hanover, NH, and this program is my first exposure to statistics. Previously I have had only hazy graduate school plans, but I have become interested in pursuing a degree in biostatistics or an MD/PhD, thanks to Dr. Wolfe's tireless recruiting efforts! This REU has piqued my interest in the broad applications of statistics, especially to psychology and medicine. I have been working with Dr. Mario Peruggia and Dr. Trisha Van Zandt on reaction times to cognitive stimuli in the form of number recognition. We are trying to find an alternative explanation for so-called "fractal" noise in the brain.

BOX 2: REU Student Reactions

How has your experience at the REU changed your impressions of academic or clinical research?

I guess I hadn't realized before how tied together the academic and clinical sides can be. Going to a small liberal arts school, I don't see the same research going on around me like I have here at OSU. Actually, I can now see myself happy in a career like that.

Until I came here I had but a diaphanous understanding of the operational aspects of clinical/academic research.

Working in the Cancer Genetics Department has shown me that the research can be stimulating and applicable to medical research, while before I thought that research was all very theoretical.

I was excited to be able to participate in the complete research experience — from collection to analysis of the data. And having this experience just recommits me to a life of research!

It made research seem much less scary and much more fun. I loved my project and I will stay in touch with my advisor as he continues the research, just because I am interested in the results.

I've been very impressed by the academic research in biostatistics. I feel like I will be able to go home and show my parents and grandparents (none of whom are in the field of academia) the work I have done and have them understand at least the basics of what my research is about. That is totally refreshing to me... Biostatistics ended up being exactly what I am looking for.

How has your clinical experience at the REU influenced your thoughts about graduate school?

I have realized that there's more out there than just [my school], and I have so many options of very interesting [programs].

I've been convinced to pursue biostatistics — not sure which degree, but I am leaning towards a PhD.

It made me realize that there are lots of people pursuing higher degrees. Often I feel pressure not to or that it's a waste of time. So it was good to see how many people are still studying.

It has helped me understand what to look for in a graduate program such as research areas, talking to students, faculty, etc.

[It] encouraged me to continue to pursue research opportunities available in the academic environment.

I didn't know for sure that I wanted to go to graduate school for statistics before attending the REU. Now I do.

Interview with Bob Hogg

Bob Hogg

Jackie Dietz

Robert V. Hogg is professor emeritus of statistics and actuarial science at the University of Iowa, where he taught for 51 years. Dr. Hogg received his bachelor's degree from the University of Illinois in 1947 and his doctorate from the University of Iowa in 1950. He is the author of several widely used textbooks, including *Introduction to Mathematical Statistics* (Hogg and Craig, 1995) and *Probability and Statistical Inference* (Hogg and Tanis, 2001). At the 2001 Joint Statistical Meetings in Atlanta, Dr. Hogg received the second annual Noether Award, given by the American Statistical Association to a distinguished researcher in nonparametric statistics. A few days before the award presentation, I spoke to Dr. Hogg about his education and career.

EJD: How did you first become interested in probability and statistics?

RVH: I was really very interested in playing games. This is when I was a kid. I liked games with dice… Monopoly, Parcheesi, things like that. It was very clear to me, even before I knew anything about probabilities, that the 6, 7, and 8 came up a lot more frequently than snake eyes (two 1s) or boxcars (two 6s), so I knew a little something about the distribution of the sum. I'd sort of figured that out. You see, the best properties to get in Monopoly – because you go to Jail – are the orange properties, because they get hit more frequently – because they're 6 and 8 and 9 spaces away from the "Get Out of Jail." So I'd always try to get the orange or the red properties. This was just something I had observed from playing a lot of Monopoly games. Since then, I've seen the game simulated, and now I've seen that the orange properties are the ones that get the highest frequencies! Now, everyone always says you want to get Boardwalk and Park Place. Those aren't good properties at all. The orange ones are the best and then the red ones, and I think maybe even the yellow ones. Although sometimes I used to go get those blue ones… I liked those… But people go to Jail, and they've got to get out! That boosts up the frequencies of the orange properties and the red properties.

But, anyhow, I just liked games in general. When I got into high school, I started to play bridge and poker and other card games. I was always interested in probability at that time – kind of "seat of the pants" probability. I did take a probability course at the University of Illinois from a man named Evans Monroe, who was a Ph.D. student of Joe Doob's while he was at Illinois.

EJD: Is that when you were an undergraduate?

RVH: Yes, I was an undergraduate then. I actually had been in service, and when I got out of service, I finished up at Illinois. I took probability out of the book by Uspensky. I don't know if you've ever heard of Uspensky. But it was kind of a classic book. I used to kid that Evans Monroe was the one who started me on my statistical career because he would make a statement, and then he'd say something like "Huh?" at the end. And the students used to count the number of huh's. He was a Ph.D. student, and when he was trying to prove something, he'd be looking at his notes, and at first you could tell that he wasn't too sure. Then he'd get all excited, and you could tell that he saw how the proof was going to go. I think that Monroe went out to New Hampshire and taught up there.

He was a bachelor at that time. And of course I was single too. I remember that he told me the story of the n drunks. Now this is like stuffing the letters in the envelopes. What's the probability that you get at least one letter in the right envelope? He told about the n drunks. They lived in this housing development where all the houses looked alike. But they were going out on a big bash that night, and so they hired a driver for the bus. They went out, and they all got plastered. When they came back, the bus driver didn't know where they lived. He'd just drop off one at this house and one at that house. What's the probability that at least one, in his terminology, sleeps with his own wife? And of course it turns out asymptotically to be $(e - 1)/e$. It comes out to be about 62 or 63% or something like that…I forget exactly. But I always used to laugh about that, because I said that the probability of any of those guys sleeping with anybody's wife was zero!

EJD: Did you major in math at Illinois?

RVH: I majored in math as an undergraduate. Then I went to Iowa. I went there because I wanted to be an

actuary. There were three of us – Ralph Goebel and Will Kragel and I. We all graduated at the same time. We'd all been in service. There was this poster there: Do you want to be an actuary? None of us knew what we were going to do. We thought, "Oh, that would be kind of fun!" So I started out in actuarial science at Iowa. That was in the math department. But I discovered that I liked statistics better than the actuarial work. Kragel and Goebel went on to be actuaries, but I just stayed there at Iowa.

As a graduate student, all I really had was eight hours of statistics. Allen Craig taught a course he called Theory of Statistics. He also taught a summer course one time out of Wald's book, *Sequential Analysis.* This was when it first came out. I thought that sequential analysis was a great thing. Well, it's still a neat idea to try to make those decisions sequentially. Maybe the Bayesians have the best way of doing that. In any case, I didn't have very much statistics.

Later, when I joined the faculty, I was lucky to get that job. You're too young to appreciate this. There was the veterans' boom right after World War II. Then we went into the Korean War. Right around 1950, when I got my degree – my Ph.D. degree – enrollments were very low. Iowa had boomed up to about 12,000 with the veterans, but then we were down to about 7,000, and this happened to every school. Jobs were not plentiful. I probably could have gotten a job at someplace like Bradley. But Craig and I had hit it off so well. He was the only statistician on the math faculty, and so they asked me to stay on. We don't do that anymore; that is, we don't hire Iowa Ph.D.s.

It was a wonderful experience for me. Oh, right at first, I was kind of formal with Craig, and then we got to be great friends. I'm much more gregarious than he was. He was a formal, Southern gentleman. Great teacher. Oh, he could write out things on the board – just perfect sentences. And when I first started to teach, I tried to be like Craig. I tried to do that same thing. And then I said, oh, this is not me. We just had completely different styles.

It was a great experience working with Craig. He liked independence. He did the quadratic form theorem, $\mathbf{AB} = \mathbf{0}$. Are you familiar with that? It's about the independence of quadratic forms. If you have two quadratic forms $\mathbf{X}'\mathbf{AX}$ and $\mathbf{X}'\mathbf{BX}$, where $\mathbf{X}$ has elements that are independent and identically normally distributed, then they'll be independent if and only if $\mathbf{AB} = \mathbf{0}$.

My thesis was on the stochastic independence of a ratio and its denominator. Maybe the best thing I proved was that if you have the sum of a_iX_i divided by the sum of X_i, and if that ratio is independent of the sum of X_i, then you must be sampling from a gamma distribution or a negative gamma distribution (Hogg, 1951). It was a characterization of the gamma distribution. I've really been interested in independence practically all my professional life. I used to think I was sort of the fastest gun in the west. I used to go through the *Annals* and see if I could spot independent statistics! I was good! Probably I could still do that pretty well.

Basically I used – are you familiar with Basu's theorem? Well, it's an easy theorem, and actually I had proved a version of it earlier. When you have a vector that is a complete sufficient statistic for some parameters, and you have another statistic whose distribution does not depend on those parameters, then the two statistics will have to be independent. Well, it's an interesting thing. Neyman had done something like this back in the 1930's in a paper that I had found. In the winter of 1951–52, I submitted a paper to a journal. I had a single parameter and a single sufficient statistic, and I showed that if you have another statistic whose distribution is free of that parameter, then they'll be independent. I had assumed the regular exponential form. By the spring of 1952, the referee of my paper had come back with what was essentially Basu's theorem. We were teaching "Basu's theorem" in Iowa three years before Basu's paper (Basu, 1955) came out! I always had a hard time calling this Basu's theorem, because it was really the referee's theorem! Lehmann and Scheffé had done a nice paper on complete sufficient statistics in *Sankhya* about 1950 (Lehmann and Scheffé, 1950), as I recall, and I figured maybe one of those fellows refereed the paper and just saw, hey, the thing I sent in can be generalized. I always felt like I had a little ownership on that theorem, but it was the referee who generalized it. I didn't know exactly what to do. I probably should have asked the editor if I could correspond with the referee, and maybe we could have written a joint paper or something like that. But at that time, I really didn't know what to do.

Craig had a big influence on me. I always liked probability. Craig didn't like to teach it, so I got to teach probability from Feller. Feller is a classic. I don't know what book you used for probability, but Feller is a wonderful book. When Scheffé's book came out on analysis of variance, Craig didn't want to teach that either. And so he taught one semester of this course we called Theory of Statistics, and the second semester we did Topics. One year I'd teach Scheffé and, the next year, something else. I was just learning – I hadn't taken any of those courses.

EJD: I was going to ask you about that! You said that you'd only taken a couple of statistics courses.

RVH: I didn't know anything about this analysis of variance or multivariate methods; I just learned along with the students. Scheffé's book, as I remember, was in the middle 50's. Anderson came out with his multivariate book in maybe the late 50's. And I said, "Oh, I'll just teach this!" Craig would do his thing on quadratic forms; he'd build a good basis for me. Then I'd just take off on topics in the second semester.

Anderson's book was good. I did a lot of that stuff. When I joined the faculty, Craig recognized that I hadn't had very much statistics. In all honesty, Craig probably never had very much statistics either!

EJD: So was your Ph.D. also in math?

RVH: Oh, yes – math! I took topology and abstract algebra – all those things. Don't quiz me on it now! It was a degree in math because I really only had that six hours plus the two-hour summer course in statistics.

EJD: So you taught yourself those statistical topics while you were teaching them?

RVH: Oh, yes! I'd just grab those books and teach them. Another thing – which was a great idea – was our seminars. We had an actuary – a man named Byron Cosby – who sometimes joined us in our seminars. The first year that I was out, we just decided to read books. The first book we read was big Cramér – *Mathematical Methods of Statistics*. It was a Princeton publication, about 1946. It's a real thick, well-written book. We just read through that book. We had about three faculty members, including the actuary, and three graduate students. The six of us would just sit around, and we'd ask each other questions. We'd say, "We'll read from here to here for next week." And we'd study. Nobody would get up and lecture, and we didn't ever go to the board. We'd just sit around and talk about it. "This is an interesting idea. Can we get an example of that?" It was the greatest experience for me – going through Cramér.

One year we read Fraser's book on nonparametrics. A book that had a big influence on me was Lehmann's book, *Testing Statistical Hypotheses*. That was an important one. In the early 60's, we had a student named Jim Hickman, who actually went on to become the dean of the business college at Wisconsin. He was a Bayesian, so we read Raiffa and Schlaifer the one year Jim was in the seminar. The seminar was a great experience.

EJD: How did you get interested in nonparametric methods?

RVH: Well, I had actually published a couple of things on nonparametrics from our seminars in the late 50's. I mentioned Lehmann's book. Lehmann's book doesn't have nonparametrics in the title, but he did a lot of nonparametrics in his book. Fraser had also come out with a nonparametrics book before Lehmann. Now, I'm going to tell you about this – if $F(x)$ is a continuous distribution, then the order statistics are complete sufficient statistics for F. So any time you get a distribution-free test statistic, by Basu's theorem, it's always going to be independent of those order statistics. Thus I was able to use Basu's theorem in nonparametrics. I used it a lot of times in some interesting places.

Ron Randles joined our faculty in 1969. He had been a student of Myles Hollander in Florida State. At that time I said, "Well, we don't really have a nonparametrician on the faculty. I do some things in that area – let's try to work together." As I look back, I was 45 at that time. Ron must have been 25 or 27 – a young guy. Probably from his point of view, having a senior faculty member take interest in him was a good idea. But of course I learned a lot more from Ron than he learned from me! But we worked from 1969 to the early 80's when he went to Florida. I bet we wrote maybe 10, 11, 12 papers together. Most of these were nonparametric papers.

Incidentally, Ron is very well organized – I imagine he was a great chairman of the department down there in Florida. Dick Scheaffer had chaired that department, and then Ron Randles, and now George Casella. By the way, Casella and Berger beats out Hogg and Craig now for first year math stat! As a matter of fact, I'm in the process of revising Hogg and Craig. Joe McKean at Western Michigan is going to help me with this.

EJD: Will this be the sixth edition?

RVH: It will be the sixth edition. The publisher wants us to boost up the mathematical level just a little bit, so it will compete more with Casella and Berger. That's Joe's job! Joe is actually an academic grandson of mine. That is, Tom Hettmansperger was one of my Ph.D.s, and Joe was a Ph.D. of Tom's. They've done a lot of good things on nonparametrics.

EJD: I just realized when you said that that you and I are related too. Tom Hettmansperger was the advisor of Tim Killeen, who was my advisor at UConn.

RVH: I didn't realize Tim was your advisor! In 1974, I had a sabbatical, but our kids were at an age when it wasn't convenient for me to go away for a whole year. They were involved in lots of school stuff. I was married then to my first wife Carolyn, who died in 1990. But I made a fairly extensive trip through the east. At that time, I had gotten pretty well acquainted with Gottfried Noether, who was at UConn. So I made Storrs, Connecticut a stop on my trip. Most of my conversations with Gottfried Noether were about statistics education. Of course, he believed very much that statisticians were making a mistake by not taking the opportunity to teach statistical ideas with nonparametrics.

EJD: I was a teaching assistant at UConn, and we used his book *Introduction to Statistics: A Fresh Approach*, which introduces inference using nonparametric methods.

RVH: Did you know he was the nephew of a very famous abstract algebraist?

EJD: Emmy Noether.

RVH: The mathematicians know Emmy Noether. She was a super mathematician. I'd heard of her when I was back in the math department. I wasn't great at abstract algebra – I did well enough to get A's in courses – but I had heard of Emmy Noether. I knew who Gottfried Noether was, and that he had kind of fixed up Pitman efficiency. Also I knew he was the nephew of this famous mathematician. You won't believe this, but I was kind of shy. When I went to meetings through the 50's and 60's – I think maybe it was a lack of confidence – maybe I thought, I can't do things with the big boys – I never went up to Gottfried and said, "Hey, I'm Bob Hogg, University of Iowa." But then we got tossed together, because we were on the ASA-NCTM Committee together for about six years. And we got to be good friends. Now he was very conservative and reserved, like Craig. I used to dash off little notes – of course, we didn't have e-mail at the time – we had these little blue note pads. They were more or less within-the-department notes, but I'd always write to Noether on the little blue notes. I'd think of something that I wanted to ask him or a comment that I wanted to make on one of our discussions. I'd write it out long hand and put it in the mail. Gottfried used to kid me about my blue notes.

But I had put Storrs on my tour. While I was there, I talked to Tim Killeen about the distribution-free multivariate two-sample problem. As a matter of fact, I don't think there's ever been a very good solution to that problem even now. But after I went home, I talked to Jim Broffitt about our ideas, and then Ron Randles got in on the discussion. Jim knew something about discrimination. He said, "Hey, maybe that will work with discrimination." And we got a real nice distribution-free technique in discrimination. I won't explain here; it's too hard. But I'd been out there to visit Gottfried, and then I got talking to Tim. And so you're a student of Tim's! And so that makes you a great granddaughter of mine.

EJD: I never realized that because I didn't know you were Hettmansperger's adviser. When you said that McKean had been Hettmansperger's student, I remembered that Tim Killeen was also.

RVH: Tom Hettmansperger had a lot of students. As a matter of fact, Tom probably produced most of my grandkids and great grandkids.

EJD: To conclude our conversation, is there any advice you would give to your great great grandkids or other new statisticians just starting out?

RVH: It's important for statisticians to support each other and cooperate for the good of the profession. Senior faculty should be mentors for junior faculty and graduate students, and advanced graduate students should help beginning students. I think it's important to have colleagues that you can interact with and bounce ideas off of. I've been fortunate to have met lots of interesting people and made lots of good friends through statistics. Statistics has been a super career for me for over 50 years, and I wouldn't trade being a professor of statistics for any other position. I hope that young people today find statistics as rewarding a career as I have. They should also make sure to have some fun along the way! Young people need to find something they enjoy doing, because they'll be doing it for a long time.

Books and articles mentioned in the interview:

Anderson, T. W. (1958), *An Introduction to Multivariate Statistical Analysis*, New York: Wiley [2nd ed., 1984].

Basu, D. (1955), "On Statistics Independent of a Complete Sufficient Statistic," *Sankhya* 15, 377-380.

Casella, G., and Berger, R. L. (1990), *Statistical Inference*, Belmont, CA: Duxbury Press.

Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton, NJ: Princeton Univ. Press.

Feller, W. (1950), *Probability Theory and its Application*, New York: Wiley.

Fraser, D.A.S. (1957), *Nonparametric Methods in Statistics*, New York: Wiley.

Hogg, R. V. (1951), "On Ratios of Certain Algebraic Forms," *Annals of Mathematical Statistics*, 22, 567-572.

Hogg, R. V., and Craig, A. T. (1995), *Introduction to Mathematical Statistics* (5th ed.), Upper Saddle River NJ: Prentice Hall.

Hogg, R. V., and Tanis, E. A. (2001), *Probability and Statistical Inference* (6th ed.), Upper Saddle River NJ: Prentice Hall.

Lehmann, E. L. (1959), *Testing Statistical Hypotheses*, New York: John Wiley [2nd ed., 1986].

Lehmann, E. L., and Scheffé, H. (1950), "Completeness, Similar Regions, and Unbiased Estimation," *Sankhya* 10, 305-340.

Noether, G. E. (1971), *Introduction to Statistics: A Fresh Approach*, Boston: Houghton Mifflin.

Raiffa, H., and Schlaifer, R. (1961), *Applied Statistical Decision Theory*, Studies in Managerial Economics, Boston: Harvard University [Wiley Classics Library ed., 2000].

Scheffé, H. (1959), *The Analysis of Variance*, New York: John Wiley.

Uspensky, J. V. (1937), *Introduction to Mathematical Probability*, New York: McGraw-Hill.

Wald, A. (1947), *Sequential Analysis*, New York: John Wiley.

A Day in the Life of an Academic Statistician in a Mathematics Department

Paul Lupinacci

Congratulations!!! You just earned your Ph.D. in statistics and you are contemplating a career in academia. However, you are not sure what the job entails. What are the research requirements? What is the teaching load? What do they mean by service? All professors have asked these questions while trying to kick-start their careers. You are certainly not alone! To help budding academic statisticians, I set out to write an article on the topic, "A Day in the Life of an Assistant Professor of Statistics." However, this is just not possible. The day-to-day activities change, well, day to day. Therefore, singling out one day's worth of activities will not accurately describe the ever-changing nature of the job. In my opinion, the fact that no two days are ever the same is one of the greatest aspects of a career in academia. Therefore, I will proceed to talk about the various activities that fill up a week or a month for me rather than single out one particular day.

Background

Before I begin, I must describe the type of university and the type of department in which I am located. The research and teaching requirements will differ from department to department. However, there are universal similarities that all new academics can relate to. This is the start of my third year as an Assistant Professor in the Department of Mathematical Sciences at Villanova University. Villanova is a mid-sized university that is located approximately ten miles west of Philadelphia. The Department of Mathematical Sciences is one of the largest departments in the College of Liberal Arts and Sciences. We have 32 professors, associate professors, assistant professors, and instructors. At the current time, there are approximately 80 students in the math major. The department functions mainly as a service department, teaching calculus and introductory statistics to the students majoring in one of the many subjects housed in the College.

My department has three Ph.D. statisticians. Although this number is not large, we are not isolated. There are plenty of opportunities to talk shop with the other statisticians. Having the statisticians housed in a Mathematical Sciences department has its advantages and disadvantages. One advantage is that you get to be known as the "stats guy." You are looked upon to make decisions regarding the statistics curriculum and to shape the future of the statistics program. In some departments, for one reason or another, there may be tension between the mathematicians and the statisticians. However, at Villanova, I am very lucky. There is no animosity between the two groups. With only three Ph.D. statisticians in the department and over 15 statistics classes to cover in any one semester, there is little chance that I will teach something other than statistics.

As an Assistant Professor, I am cognizant of the ticking of the proverbial tenure clock. In preparing to apply for tenure in a couple of years, I am constantly juggling many professional activities. These activities include the usual "Big 3" on which I will be evaluated at tenure time: teaching, research, and service. In addition to those activities, I also perform some statistical consulting on the side. What follows is a discussion on how I am attempting to fulfill each of the job requirements.

Teaching

My current teaching load is three classes per semester. This is a reduced load that will be in effect until my tenure application is decided. Since the department offers a Master's Degree in Applied Statistics, one of my three classes is at the graduate level. For the fall

Paul Lupinacci (paul.lupinacci@villanova.edu) is an Assistant Professor in the Department of Mathematical Sciences at Villanova University. Paul received his B.S. in mathematics from Villanova in 1995 and his M.S. in statistics from Temple University in 1997. He received his Ph.D. in statistics from Temple in January of 2001. His thesis focused on the area of Experimental Design. Paul is a Philadelphia sports enthusiast and an avid golfer.

semester of 2002, I am teaching two sections of Introductory Statistics for Liberal Arts students, as well as the Master's level course in Statistical Methods. The undergraduate Introductory Statistics sequence is taken predominantly by undergraduates to fulfill a math requirement. As such, these students are not overly excited to find themselves in a statistics class at 8:30 in the morning every Monday, Wednesday, and Friday. (Yes, I picked that time!) Nevertheless, my biggest thrill comes from working with these students. On the first day of class, I inquire into my students' perceptions of statistics. Some of their responses have been "boring," "drudgery," and "watching paint dry would be more exciting!" I love the challenge of trying to change these perceptions and motivating the students to the point where they are excited about (well, ok, interested in) statistics.

The graduate students present a different dynamic. They are eager to learn the material, and they keep me on my toes by asking some very thought provoking questions. The Statistical Methods course takes a very long time to prepare. I probably spend twice as much time preparing for the graduate level course as I do the undergraduate course. The most time consuming aspect of the preparation is finding appropriate data sets that both illustrate the statistical method as well as demonstrate the application to a "real world" problem. In my opinion, an ideal graduate level data set 1) demonstrates the need for the methodology under study; 2) is taken from a context to which the student can relate – a student is much more eager to learn the material when the necessity of the knowledge is clearly made; 3) is a little bit messy, i.e., some massaging is necessary before the data set can be analyzed; and 4) can be extended to a more advanced method. For instance, suppose we are studying inference for one mean. If I have a data set that contains information on cholesterol level of patients, I would also like that data set to include a treatment variable and a post-treatment measurement. I can then illustrate inference for one mean, inference for two means, and inference for paired samples. This allows the student to easily compare the data structure that is required for each method. The internet is such a wonderful resource for obtaining great data sets. There are a number of sites that I have bookmarked from performing searches in the past. However, a great way to find data sets is to simply search for a topic using your favorite search engine, such as Google.

Research

The second branch of the "Big 3" is research. Regarding this area, I recently received some great advice from a former professor of mine at Temple University, Burt Holland. He said that "you must always have multiple irons in the fire. You must always be working on more than one research project at a time. If one goes cold, you can put it aside for a while and make progress on another." I think that this is good advice for any new faculty member. When I was first hired, my research suffered because I was adjusting to a new school and new colleagues. I spent most of my time preparing for classes because I wanted them to be perfect! You will find that the sooner that you can settle in and get your research started, the better off you will be. However, don't panic if it takes you a while to settle in. It happens to all of us. Currently, I am working on two research projects with a third to begin very soon.

One of these projects is in collaboration with a friend and former colleague of mine, Terry Hyslop. Terry is a biostatistician at Thomas Jefferson University Hospital in Philadelphia. We are finishing a paper titled "A Nonparametric Fitted Test for the Behrens-Fisher Problem." Terry and I are also working on a second research project that develops a new mixed effects model for a data structure that is poorly dealt with by existing methods. The third project, which is currently on the back burner, concerns one of my favorite topics, fractional factorial designs. I am starting to do the background research for this project in collaboration with a colleague of mine from Villanova, Joseph Pigeon. None of these projects is related to my dissertation work. The collaboration with Terry stemmed from some of the consulting work that I do on the side. This is one of the reasons why I believe statistical consulting is so important for someone in academia. The consulting gives the academic statistician the ability to keep abreast of the new statistical methods that are used in the real world. The mixed effects model research came from an actual study that was performed at TJU. We realized that the methodologies available did not adequately address the question of interest, so Terry and I set off to develop a new method. This is a great example of the real world driving research.

Service

The third branch of the tenure tree is service. Service can be broken down into smaller branches: service to the department, service to the university, and service to the profession. I fulfill service to the department in a number of ways. First, I am the faculty advisor to the math club. In this capacity, I oversee the activities of the club which range from sponsoring a career night to running the student/faculty softball game. Second, I am the advisor to a few of our math majors. I help the students create their semester rosters while ensuring that the schedule they have created will keep them on track to graduate in four years. Finally, I am also the course coordinator of the introductory statistics sequence that I previously mentioned. I am in charge of selecting the textbook for the sequence and I oversee the development of its curriculum. I am the go-to guy if any of the faculty members have

difficulties with any aspect of the course.

There are many ways to fulfill service to the university. Some faculty members serve on committees while some get involved with student organizations. Let me talk about a unique way that I am fulfilling this obligation. I was the faculty evaluator for the Special Olympics of Pennsylvania Fall Festival that is held annually at Villanova. Every fall, hundreds of Special Olympics athletes come to Villanova from all over the state to compete in the Fall Festival. This is a major event on Villanova's campus. It is almost entirely student-run. My job was to wander from event to event, take notes, interview athletes and their families, and then prepare a document that reports on how smoothly the event ran, where things could improve, etc.

In terms of service to the profession, I have become more involved with the preparation of Advanced Placement statistics teachers. During the summer, one of my colleagues, Tom Short, ran a two-day "Beyond AP Statistics" seminar at Villanova. The goal of the two-day event was to introduce the teachers to material that goes a step or two beyond that which is needed for the AP course. Hence, the teachers gain a better understanding of where the AP material falls in the grand universe of statistics. At the seminar, I was given the opportunity to present material on Design of Experiments. I have also become an active participant in a group called P.A.S.T.A. (Philadelphia Area Statistics Teachers Association). The goal of the group is to get area high school statistics teachers together with their university counterparts to discuss new teaching strategies and enhance the understanding of statistics by teachers across the region.

Consulting

How did I get attracted to statistics in the first place? It was the ability to work with data to solve problems and answer questions that drew me to the profession. To this end, I try to fit in as much statistical consulting as possible. I am currently analyzing data for the nurses at the "Living Well After Cancer" program at the University of Pennsylvania. I am also the consulting statistician to a group of nurses and doctors at Pennsylvania Hospital in Philadelphia. I really enjoy analyzing medical data. You see, I am a medical idiot! I cringe when I see an athlete break a bone or twist a knee. However, by analyzing data that are associated with new advances in medicine, I feel like I am on the front line of medical breakthroughs.

One must be careful not to get too caught up in the consulting side. At Villanova, this is not an activity that will be recognized come tenure time. However, I have already mentioned how my consulting led to ideas for future statistical research. Also, I have been lucky enough to have been included as a co-author on some of the non-statistical papers that have resulted from my consultations. At Villanova, this type of paper is not counted as much as a paper developing new statistical theory, but it does count towards tenure. Therefore, I believe that consulting has been very beneficial for me in terms of growth as a statistician, as well as preparing me for tenure.

Project NExT

Let me leave you with one piece of advice. One of the best decisions that I have made since coming to Villanova is getting involved with a program called Project NExT. Designed for new and recent Ph.D.s in the mathematical sciences, Project NExT (New Experiences in Teaching) is a yearlong professional development program that addresses faculty responsibilities in teaching, research, and service. Project NExT is a program that was developed by the Mathematical Association of America. Although statistics is a particular branch of mathematics, there are enough differences that I was hesitant to apply. To learn more about the program, I was directed to John Holcomb at Cleveland State University, who is a statistician and a graduate of the Project NExT program. John made it very clear that this program is very beneficial for statisticians. I want to echo those remarks.

As a Project NExT Fellow, I have attended three meetings in the last year. One of the meetings was held last summer in Madison, Wisconsin. A second was held this past January in San Diego, and the third and final meeting was held in July in Burlington, Vermont. I attended sessions ranging from "Getting Your Research off to a Good Start" and "Obtaining Grants from the NSF" to "New Methods of Assessing Student Learning" and everything in between. Each session was informative and beneficial. You pick up something here and you learn something there. In the end, you leave the program with a wealth of information that would have taken years to accumulate.

However, as good as the professional development sessions were, the best part of Project NExT was the ability to form a network of friends that are at the same point in their careers. An email distribution list was set up so that we can bounce questions off of the other Project NExT Fellows who may have dealt with these same issues. For example, someone posted a question about student cheating to the list and within an hour there were at least ten responses. This list is also monitored by selected mentors. The mentors are accomplished teachers and researchers in the various areas of mathematics. Their input in the discussions has been an invaluable resource for the Project NExT Fellows. There were six statisticians among the 2001-2002 Fellows, and we have formed our own email distribution list. This program is one that every new statistician entering academia, particularly in a Mathematics Department, should inquire about. After going through the program, I now feel prepared and confident that I

will succeed in this profession.

Conclusion

I have tried to give you a broad understanding of the activities that find their way into my day. However, you must be asking yourself how I can find the time to fit all of these activities into my schedule. This was the hardest adjustment that I had to make coming out of graduate school. My research stalled during my first year because I didn't know how to juggle all of the activities. The research was the activity that didn't have an immediate deadline and so it kept getting pushed off. I realized that this pattern wasn't going to earn me tenure and I needed to structure my day differently. My new structure has been set up around my teaching preferences. During the school year, I teach early in the morning. Therefore, the mornings of each day are dedicated to teaching. I prepare for my classes, teach, grade, etc. in the mornings. Once noontime comes, I stop whatever I am doing and switch gears to work on research for the rest of the day. My research time includes statistical research, as well as my consulting. However, I generally try to keep my consulting to the weekends. My day is never this black and white; however, over the last year I have tried very hard to keep this schedule intact. My research is now up and running, and I believe that it is progressing nicely. This schedule may or may not work for you, but the important idea is to structure your day. In academia, the workday has no structure other than the few hours that we are in class. This can be very dangerous for a new academic. You will definitely spend your first year finding out what schedule works best for you. Find it and stick to it. You will soon see progress on all fronts.

In conclusion, let me emphasize that before you focus on the tremendous workload, realize that the rewards of the profession are greater. As a professor, you are able to get involved with many activities both inside and outside of the field of statistics. Everyone that enters this profession will have to do some research, be a very good educator and perform some service. I have given you some ideas on how I have tried to accomplish these tasks. However, everyone is different. You must find your own niche in your own department. Find something in your department that you can grab ahold of and make it yours. Do it and do it well, and you will have created a niche for yourself. This is what I am trying to do, and I am having a ton of fun doing it! I couldn't imagine doing anything else!

Data Sleuth

Mystery 1: Aircraft Operating Costs

Contributed by Dan Teague, North Carolina School of Science and Mathematics

Every year, the Air Transport Association (ATA) issues an Annual Report on the US Airline Industry. These reports provide Aircraft Operating Statistics for the major aircraft flying in the United States. The reports (since 1994) can be found online at *www.airlines.org/public/publications/display1.asp?nid=916*. Among the statistics presented are the average number of passenger seats within each model of plane, the average speed in flight measured in miles per hour, the average length of flight in miles, and the average cost per hour in dollars. Table 1 below presents these values for aircraft in 1996 whose average passenger capacity is greater than 200. These data were presented in one of the free-response questions for the 2002 AP Statistics Exam.

The following is the least squares regression equation for predicting the average cost per hour from the average number of passenger seats:

predicted average cost = 1136 + 14.7 *average num seats*

Question 1: Provide an interpretation of the slope coefficient in context.

Question 2: Explain why the sign of the slope coefficient makes sense.

The least squares regression line can also be calculated for just those aircraft whose capacities are between 250 and 350 passenger seats:

predicted average cost = 13605 - 29.97 *average num seats*

Question 3: Explain how it can happen that the slope coefficient is positive for the entire dataset, but negative for this subset of the data.

From the data in Table 1, it is clear that the different aircraft fly at different speeds. Perhaps cost per *mile* would be a more valid measure for comparison than cost per *hour*.

Question 4: Create the new variable Avg Cost per Mile by dividing Avg Cost per Hour by Avg Speed. Does the relationship between Avg Cost per Mile and Avg Passenger Seats reveal the same anomaly as the relationship between Avg Cost per Hour and Avg Passenger Seats?

(continued on page 28)

Table 1: Averages for Commonly Used Aircraft Models

Model	Average Number of Passenger Seats	Average Speed in Flight (miles/hour)	Average Length of Flight (miles)	Average Cost per Hour (dollars)
B747-100	410	518	2882	6567
B747-400	400	539	5063	7075
B747-200/300	369	529	3231	7790
L-1011-100/200	305	498	1363	5081
B-777	291	513	2451	4194
DC-10-10	286	498	1493	5092
DC-10-40	284	504	1963	4684
DC-10-30	272	516	2379	5859
A300-600	266	467	1126	5123
MD-11	260	524	3253	6335
L-1011-500	222	523	2995	4764
B767-300ER	216	495	2331	3616

The Statistical Sports Fan

Survival at the 2002 Soccer World Cup

Robin Lock

Teams representing thirty-two countries gathered in South Korea and Japan in June of 2002 to compete for the quadrennial FIFA World Cup. After sixty-four matches, Brazil emerged as the champion with a 2-0 victory over Germany in the final game. A total of 161 goals were scored during the tournament, yielding an average of just over 2.5 goals per game. Most of those goals were critically important, as 43 of the 64 games (67%) ended in a tie (16 games) or were decided by the margin of a single goal (27 games). In the non-tie games, the team scoring first went on to win in 39 out of 48 cases (81%). Given the scarcity of goals at this elite level of soccer, teams often adjust their style of play after the first goal of the game is scored, either becoming more conservative to protect their lead or attacking more aggressively to try to score an equalizing goal. The time (in minutes) until the first goal of a game can be studied using ideas from survival analysis. Such techniques have been applied in medicine (survival of a patient after treatment), actuarial sciences (lifetimes), and industry (reliability of mechanical or electrical components before a breakdown). We will define "survival" in soccer as the time from the start of a game until either team's defense "breaks down" and allows the first goal. Data on survival times until the first goal for 2002 World Cup matches can be downloaded from *www.amstat.org/publications/stats/data.html* .

Finding a Model to Describe Soccer Survival Times

Figure 1 shows a histogram of the survival times for all but three of the 2002 World Cup matches. The three omitted games ended as 0-0 ties with no goals being scored. The mean for the times is 33.5 minutes with a standard deviation of 24.2 minutes. What might be a reasonable distribution to model these times?

One of the basic probability models used in survival analysis is the exponential distribution with density function given by

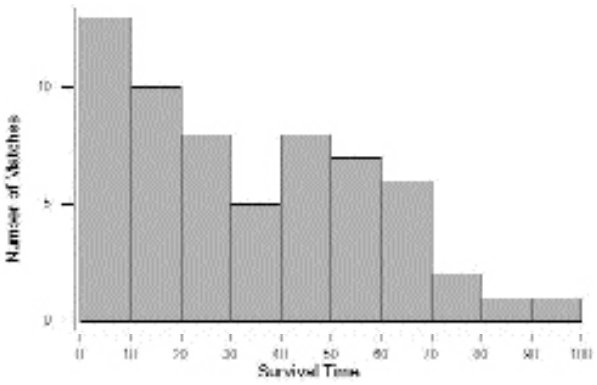

Figure 1. Time (in minutes) until the first goal in 2002 World Cup Games.

$$f(x) = \frac{e^{-x/\lambda}}{\lambda} \text{ for } x > 0.$$

The single parameter, λ, is the mean of the distribution, so it can be easily estimated using the mean of the sample, which is also the maximum likelihood estimate. Does an exponential distribution with parameter $\hat{\lambda} = 33.5$ provide a good model for the World Cup survival times? One way to examine the fit is to use a probability plot (left graph in Figure 2) where the vertical axis of percentages is scaled so that a perfect exponential sample would fall precisely on a straight line. In our case, the plotted times curve below the line towards the end of the first half (45 minutes), indicating that the times in the lower half of the distribution are longer than we would expect from the exponential model (i.e., the actual median is larger than the model's median). The points in the probability plot then begin to rise more steeply than the line for times in the second half, showing that the top 10% of the distribution has shorter times than we would expect. This can also be seen by comparing our histogram to this exponential model (right graph in Figure 2). There are fewer matches with quick goals than we would expect, more matches with the first goal occurring in

the 40-70 minute range, and fewer matches with no goals until after the 70th minute.

A More Sophisticated Model

One of the features of an exponential distribution is sometimes called the "memoryless" property, i.e., no matter how long the game has gone on without a score, the chance of a score in the next interval (say ten minutes) is exactly the same as the chance of a goal in any other interval of the same size (for example, the first ten minutes). A more sophisticated model would allow the probability to change as the game progresses to account for players adjusting to their opponent's strategy or becoming fatigued as play proceeds. One such probability model is the Weibull distribution with density given by

$$f(x) = \frac{\alpha}{\beta^{\alpha}} x^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{\alpha}} \quad \text{for } x > 0.$$

One can see that the exponential is a special case of the Weibull distribution (when $\alpha = 1$). The estimation process is a bit more complicated with two parameters, so we rely on statistical software (in this case Minitab) to obtain maximum likelihood estimates for the parameters that give the best Weibull fit to our soccer survival times. The estimates are $\hat{\alpha} \approx 1.31$ and $\hat{\beta} \approx 36.2$, and the probability plot for this Weibull distribution (Figure 3) shows a distinct improvement over the exponential.

Censored Data

But what about the 0-0 games? Two of these (France vs. Uruguay and Nigeria vs. England) occurred during the initial rounds of play and ended after 90 minutes with a tie result. The third 0-0 score (Spain vs. South Korea) happened during the elimination quarterfinals where a winner must be determined so the teams played an additional 30 minutes of scoreless overtime before settling the outcome with a series of penalty kicks (South Korea advancing with a 5-3 advantage). Another quarterfinal game (Senegal vs. Turkey) also was 0-0 after regulation time, but Turkey scored in the fourth minute of overtime to produce a survival time of 94 minutes. Our analysis so far has ignored the 0-0 final scores by treating them as missing values, even though they represent three of the four longest waits to see a goal. In survival analysis, such cases are called *censored* data because we couldn't follow them long enough to find the time until the stopping event (the first goal) occurred. Such circumstances arise regularly in medical studies when patients (fortunately) survive beyond the length of the study. Our parameter estimations should take these cases into account when trying to model the length of survival times. Fortunately, the maximum likelihood estimates can be recomputed to include the likelihoods of the censored cases (see, e.g., Lee, 1992). Table 1 shows the parameter estimates for both the exponential and Weibull models with and without the censored observations. The increases in the estimated values for λ and β in reflect a higher mean time until the first goal that helps account for the long time with no goals in the 0-0 games. In the Weibull case, the mean also increases as α decreases, so the mean of the fitted Weibull distribution based on censored data (37.9 minutes) is longer than the mean without the censored data (33.4 minutes).

Figure 2.

A Nonparametric Alternative

Although the Weibull model provides a reasonably good fit to the censored survival times, perhaps there is something unique about soccer scoring (for example, the halftime rest break at 45 minutes) that is not captured by a simple two-parameter model. Another way to display such models is with a *survival plot* that shows the proportion of cases that should still be surviving as a function of the time *t*. A typical survival plot starts at one (at time zero) and decreases to zero as time increases; it may be interpreted as one minus the distribution function for the survival times. The smooth curve in Figure 4 shows the survival plot for the Weibull distribution that we fit to the censored survival times.

A survival curve may also be estimated directly from the data as a step function with decreases at the observed data values. A Kaplan-Meier survival curve adjusts the step function to account for right-censored data (when all observations beyond a certain time value are censored) while a Turnbull plot allows for arbitrary censoring. More details on these nonparametric approaches can be found in Lee (1992). Since our dataset has a valid observation at 94 minutes, with censored values at both 90 and 120 minutes, the Turnbull estimates were used (via Minitab) to produce the step function shown in Figure 4. Note that, although the parametric (smooth) curve tracks the nonparametric (step) function fairly well, the "halftime" effect is noticeable as the Turnbull approximation shows a slightly faster scoring rate at the beginning of each half of the game with distinct sections that flatten more than the parametric curve near the end of each half.

Conclusion

Techniques from survival analysis can be used effectively to model the distribution of times until the first goal in World Cup soccer matches. The presence of 0-0 ties at both 90 and 120 minutes provide interesting challenges to handle as censored data. For additional work, one might examine survival times for games from other levels of soccer, such as professional leagues, colleges or schools. Does the halftime effect observed at World Cup games appear in other circumstances? Would it be better to model the first goal in each half separately? Other sports such as

Figure 3.

lacrosse, ice hockey and field hockey regularly report goal times. What would their survival models look like? Finally, what about second or later goals? Does the style of play really change after a first goal or would the times until the next goal (which would exhibit a much more elaborate censoring scheme for all the soccer games that don't have a second goal) follow a similar survival pattern?

References

Lee, E. T. (1992), *Statistical Methods for Survival Analysis*, New York: John Wiley & Sons.

Turnbull, B. W. (1976), "The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data," *Journal of the Royal Statistical Society,* Vol. 38, pp. 290-295.

Web Resources

2002 World Cup game summaries can be found at ESPN's website *http://worldcup.espnsoccernet.com/results*

Table 1. Parameter estimates for Exponential and Weibull models

	Exponential	Weibull	
	$\hat{\lambda}$	$\hat{\alpha}$	$\hat{\beta}$
Without censored cases	33.5	1.31	36.2
With censored cases	38.4	1.19	40.2

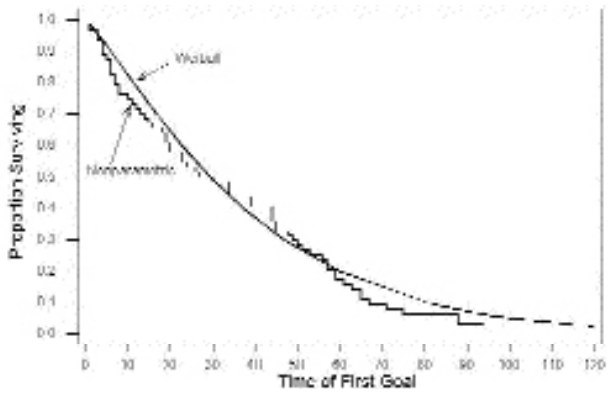

Figure 4. Survival plots for Weibull and Nonparametric models.

AP Statistics

Some Thoughts about Influential Points

Gretchen Davis

Many concepts in statistics are not entirely mathematical in nature. Part of the richness of the discipline is the combination of algebra and "art," where interpretation and subjective judgment are key. As a consequence some definitions, even when well written, can leave a concept difficult to understand. For example, in the Peck, Olsen, Devore (2001) text, we read, "When a single observation plays a big role in determining the slope of the least squares regression line, it is therefore called an influential point." The Yates, Moore, and McCabe (1999) book reads, "An observation is influential if removing it would markedly change the position of the regression line. Points that are outliers in the *x* direction are often influential." We may still find ourselves wondering what "a big role" is and how "marked" the change needs to be. The difference between an influential point and an outlier can be especially confusing.

Illustrating the concept with a concrete example may be the breakthrough that we need to increase our understanding. Looking at the 1832 census data from the 21 California missions (Table 1) helped me and my students understand the distinction between an outlier and an influential point.

Looking at Indian and Livestock Populations

Let us first focus on the relationship between the Indian population and livestock population at the missions during that time. We notice that the data in Figure 1 have a positive relationship. Missions with larger than average Indian populations tend to have larger than average livestock populations as well (r = .873). The least squares regression line is:

predicted livestock = 1921 + 15.8 *Indians*.

Table 1: California Mission Data from the 1832 Census

Mission	Date	Indians	Livestock	Crops
San Diego de Alcala	1769	1455	18200	158675
San Carlos Borromeo	1770	185	5818	103847
San Antonio de Padua	1771	640	17491	84933
San Gabriel Arcangel	1771	1320	26342	233695
San Luis Obispo	1772	231	8822	128751
San Francisco de Asis	1776	204	9518	67117
San Juan de Capistrano	1776	900	16270	83923
Santa Clara de Asis	1777	1125	20320	98356
San Buenaventura	1782	668	7616	135303
Santa Barbara	1786	628	5707	151143
La Purisima Concepcion	1787	372	13985	191014
Santa Cruz	1791	284	9236	58072
Nuestra Senora de la Soledad	1791	339	12508	68408
San Jose	1797	1800	24180	222809
San Juan Batista	1797	916	12333	69577
San Miguel Arcangel	1797	658	12970	105049
San Fernando Rey	1797	782	9060	95172
San Luis Rey	1798	2788	57330	92656
Santa Ines	1804	360	9860	179925
San Rafael Arcangel	1817	300	5492	74609
San Francisco de Solano	1823	996	5063	10991

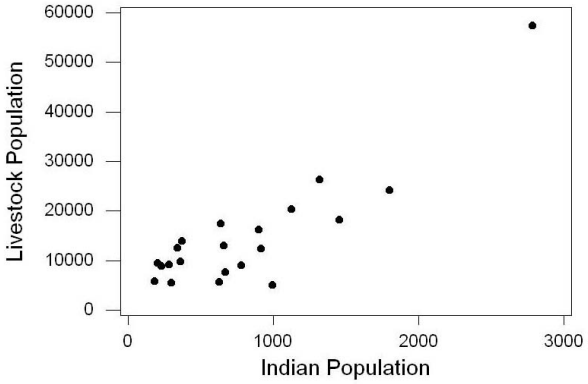

Figure 1: Scatterplot of the Indian Population and Livestock Population at the 21 California Missions in 1832

There is one mission that seems to be of special interest. San Luis Rey (located north of San Diego in Oceanside, CA) has both the highest population of Indians and the highest livestock population and could be considered an outlier in both the *x* variable and in the *y* variable. This point has the largest residual, making it an outlier in the regression model as well. We also note that if we remove the point, the slope of the regression line changes considerably, dropping to 9.82 (see Figure 2).

Looking at Crop Production and Livestock Population

The San Luis Rey data point is also of special interest in Figure 3, which displays livestock population vs. crop production at the missions. With San Luis Rey, the cloud of points is very wide, and the slope is a rather flat .04. However, when we remove the San

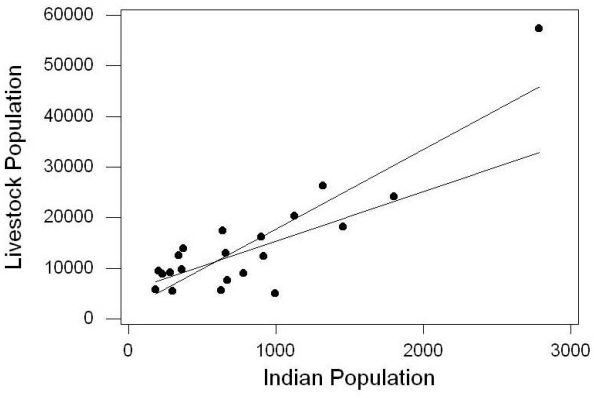

Figure 2: Scatterplot with Regression Lines with and without Data from San Luis Rey

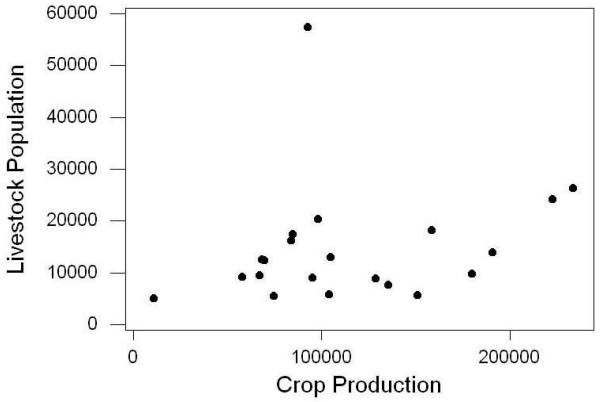

Figure 3: Scatterplot for Livestock vs. Crops at the California Missions

Luis Rey point this time, the slope changes to 0.059, which is not a very dramatic change, as we see in Figure 4. While the point is clearly again an outlier in the regression model (with a much larger residual than the rest), it is not an influential observation in this case.

What's different in this situation? While San Luis Rey was an outlier in both variables in Figure 2, it is only an outlier in the *y* variable (livestock population) here. However, what if we reverse the axes and consider crop production as the response to livestock population? This does not change the correlation coefficient, but now San Luis Rey is an outlier in the *x* variable. Notice that if we remove the point there is now a dramatic effect on the least squares line (Figure 6)! It appears that whether or not a point is influential is related to whether or not it is extreme in the *x* direction.

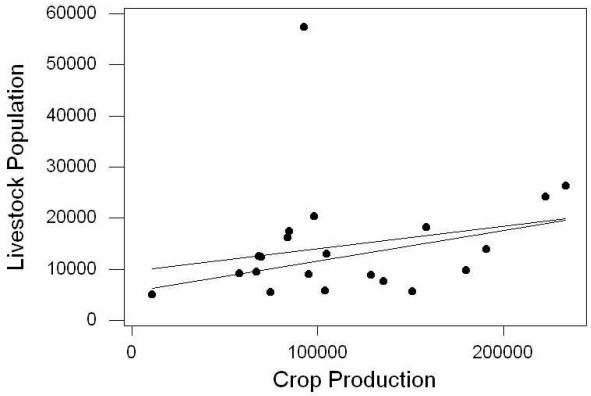

Figure 4: Scatterplot with Regression Lines for Livestock vs. Crops

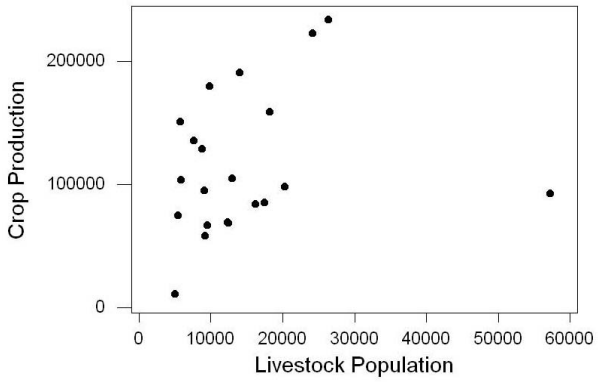

Figure 5: Scatterplot for Crops vs. Livestock at the 21 California Missions in 1832

Looking at the Algebra

To see why this happens, let's consider the algebra of the calculations of the least squares slope coefficient. The first thing to notice is that the slope depends on both the variability in the y direction (as measured by s_y) and the variability in the x direction (s_x). However, once we substitute in the equation for the correlation coefficient, the s_y terms cancel and the estimate for the variance of x (s_x^2) dominates the denominator of the slope. A large deviation in the x direction thus has a much greater influence on the slope of the regression line than a large deviation in the y direction.

When a statistical software package "flags" an observation as being potentially influential, it typically does so by looking for points that have x values far from $\bar{x}$. These are sometimes called "leverage points." These points would need to be individually removed from the data set and the change in the regression slope measured to determine whether a point's potential for

$$
\begin{aligned}
slope &= r\left(\frac{s_y}{s_x}\right) \\
&= \left(\frac{1}{(n-1)s_x s_y}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right)\left(\frac{s_y}{s_x}\right) \\
&= \left(\frac{1}{(n-1)s_x}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right)\left(\frac{1}{s_x}\right) \\
&= \left(\frac{1}{(n-1)s_x^2}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right).
\end{aligned}
$$

influential is realized. One such measure is called "Cook's distance." This formal measure goes beyond the scope of the AP course, but the concept of influence is important to understand for interpreting scatterplots and least squares line thoughtfully.

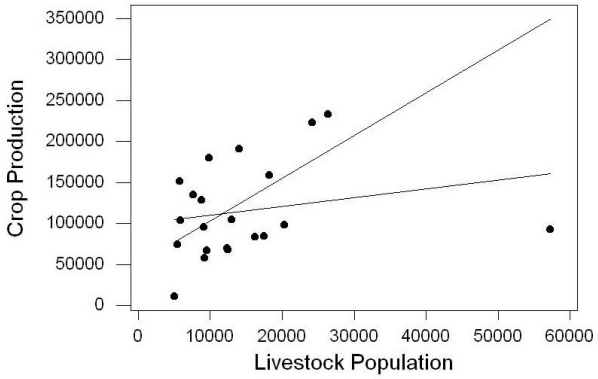

Figure 6: Scatterplots with Regression Lines for Crops vs. Livestock

Conclusion

Looking first at the scatterplots and then at the summary statistics for regression and finally the algebra behind the formulas may help us see when and why an outlier becomes an influential point. Your homework assignment is to determine why the San Luis Rey mission was such an unusual observation!

Acknowledgements

I am grateful to Linda Hermosillo, a fourth grade teacher at 7th Street School in LAUSD who shared the California Mission data during a *Dealing with Data* course, sponsored by the UCLA Math Content Program.

References

Peck, R., Olsen, C., and Devore, J. (2001), *Introduction to Statistics and Data Analysis*, Pacific Grove: Duxbury.

Yates, D., Moore, D., and McCabe, G., (1999), *The Practice of Statistics: TI-83/89 Graphing Calculator Enhanced,* New York: Freeman.

Web Resources

California Missions Website: *missions.bgmm.com*
Mission San Luis Rey Website: *www.sanluisrey.org*

Java applets illustrating influence:
statweb.calpoly.edu/chance/applets/LRApplet.html
statweb.calpoly.edu/reins/StatDemo/All.html
www.stat.sc.edu/~west/javahtml/Regression.html

ø-sings

We're Number 2!

Chris Olsen

My well-worn list of words spelled differently than I think they should be spelled — i.e., my dictionary — informs me that the word "pantheon" refers to the realm of the heroes or persons venerated by any group: e.g., a place in the pantheon of American literature. I really don't have any conception of a pantheon of American literature, except possibly that collection of folks who know how to spell, but I do read the poet Robert Frost now and then. It was Frost, as you may recall, who penned those immortal words, "Two roads diverged in the woods, and I took the one less traveled by, and that has made all the difference." The professional road I chose is teaching mathematics and statistics, and over the years I have constructed a pantheon of sorts of professional roads-less-traveled-by. This is a personal list of professions that serve to elevate my spirits after a proverbial hard day at the office. For example: "I messed up 3 trig identities, 2 normal curve table look-ups, and 4 differentiations of polynomials. But things could be worse; I could be a(n) _____."

I presume that each of us has a similar list of such occupations, and for many people it looks like this:

Standard Could-Be-Worse List

1. Accountant
2. Statistician
3. [Others…]

I'm sure the reason for this ordering is the popular view that accountants and statisticians spend their whole working lives consulting encyclopedic sets of books of collections of pages of rows and columns of numbers. Now, I must say that I'm a bit concerned about this. No, no, I'm not concerned about statisticians being number 2; I'm concerned they may soon move into the number 1 spot! Recent events, as reported in the press, have unearthed the heretofore very well hidden fact that accountants are not tied to their office chairs reading sets of books of collections… etc. I'm gleaning from the press that there is a branch of this profession known as "Creative" Accounting, championed by the well-known firm of Arthur Andersen. Creative Accounting. Starts with C, and that rhymes with T, and that stands for Trouble in River City. Fellow students of statistics, we must fight back. Should the general public get the idea that accountants lead creative and exciting lives, this will surely result in:

Standard Could-Be-Worse List (Revised)

1. Statistician
2. [Others…]

Notice, there isn't even a credible replacement for number 2. Obviously we statisticians are all going to have to do our part to avoid being number 1. If we fail, we will be doomed to a life of two strategies when meeting people at parties or elsewhere. One, watch people edge away from us when they find out we do statistics, and Two, be disingenuous, e.g., claim to be in counterespionage — or possibly, an accountant. How can we convince others that statistics is a vibrant, exciting profession, populated by captivating individuals solving interesting problems? Well, I myself am doing my part by pointing out what you can do for your part.

First of all you can start in your immediate circle by mentioning to everyone (e.g., department heads) what exciting and captivating and interesting people your statistics professors are. This will surely get back to them, and your grades may even improve.

Beyond your immediate circle — such as when you are small-talking at parties — better evidence of statisticians as exciting people will be needed! And by some quirk of fate I have just the cyberspace place to go for information on these men and women. At a click of your mouse, you can have descriptions of the life and work of statisticians, pictures and portraits of them, and in some cases access to images of their original writings! I know you are breathless with anticipation already, so here is the URL:

http://www.york.ac.uk/depts/maths/histstat

No mere thumbnail sketches of statisticians here!

There is extensive discussion, complete with the sources of the information presented, just right for those ice-breaking conversation starters at parties. For example, you can download pictures of an astragulus (four views!) and/or a quincunx and just happen to have them for easy showing on your host's coffee table. (Suggestion: practice saying these words before you unveil the images at a party.) For those into astrology and signs — (no, I am NOT misspelling astragulus) — there is a section on the first use of probability and statistics symbols. As is well known, probably even by accountants, there are two common systems for indicating parameters and estimates: Greek letters and hats. But is it common knowledge that both systems have their beginning in the writings of R. A. Fisher? Well, I guess not! Armed with such knowledge, you will be the statistical raconteur of every party.

Nor will you be at a loss if someone tries to grab the limelight by bringing up religion and politics. You can download John Arbuthnot's discussion, "An Argument for Divine Providence, taken from the constant Regularity observ'd in the Births of both Sexes" in Postscript, pdf, or LaTeX format. Or, you can expound sagely on the political careers of William Farr (Chief Statistician of the UK) and Florence Nightingale, the Passionate Statistician, and her efforts to reform army health practices after the Crimean War.

And what about those people who insist on pulling out pictures of their children — or worse, grandchildren?!? After a visit to this site, you will not be at a conversational disadvantage simply because you don't have any children to crow about. You will be able to bring forth images of the statisticians of old, slyly pointing out that they appear to be just as handsome/beautiful as the statisticians of today. (One should not use these images indiscriminately — I would counsel against the picture of Sir John Sinclair, 1754 – 1835, who appears to be either surrendering to Washington at Yorktown, or picking up his date for the prom, or both.)

Students of statistics, your duty is clear. Visit this site, stock up on information and pictures, and sally forth in the fight to liberate statisticians from their stick-in-the-mud image the next time you go out socially! Take up thy cyber-cudgel, and remember the words of Don Quixote after his most famous battle: "We're number 2!"

Data Sleuth, from page 19

Mystery 2: Aircraft Operating Speeds

Contributed by Dan Teague, North Carolina School of Science and Mathematics

The dataset above gives rise to some more oddities.

Question 5: Examine the relationship between Avg Cost per Hour and Avg Cost per Mile. Is it strongly linear? Explain why this makes sense.

The least squares line for predicting Avg Cost per Hour from Avg Cost per Mile turns out to be:

predicted Avg Cost per Hour =
-546.48 + 562.75 *Avg Cost per Mile*

Data Sleuth Solutions

Mystery 1: Aircraft Operating Costs

Question 1: The slope coefficient of 14.7 suggests that if one model of aircraft has one more seat on average than another model, then the predicted average operating cost for the first model would be almost $15 more per hour than for the second model.

Question 2: It seems reasonable that aircraft models with more seats cost more to operate.

Question 3: The three models with the most seats cost the most to operate, and the model with the fewest seats costs the least to operate. Those extreme cases lead the overall association to be positive, even though the association is slightly negative for the remaining cases. A scatterplot (Figure 1) reveals this well.

Question 4: The relationship is very similar whether ones uses Avg Cost per Hour or Avg Cost per Mile, as shown by the scatterplot in Figure 2.

Mystery 2: Aircraft Operating Speeds

Question 5: The relationship between Avg Cost per Hour and Avg Cost per Mile is quite linear, as shown in the Figure 3 (r = .988). This makes sense because the aircraft operate at very similar speeds, so the cost per mile is nearly a linear transformation of the cost per hour.

Question 6: The units on this slope coefficient are dollars per hour divided by dollars per mile. The dollars cancel, leaving the units as miles per hour.

Questions 7,8: The intercept term is one reason that the slope does not estimate the average speed well. Fitting the regression model without the intercept term produces the least squares equation: *predicted Avg Cost per Hour* = 513.81 *Avg Cost per Mile*. The slope estimate here is much closer to the mean of the models' average speed measurements (510.33). If the average speed had been constant across models, then you can show algebraically that the slope would be equal to that value. The slope won't quite equal the average of the speeds due to properties of sums.

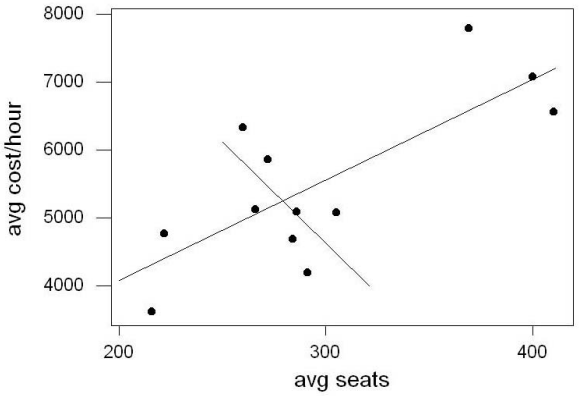

Figure 1: Avg Cost per Hour vs. Avg Seats, with Two Regression Lines

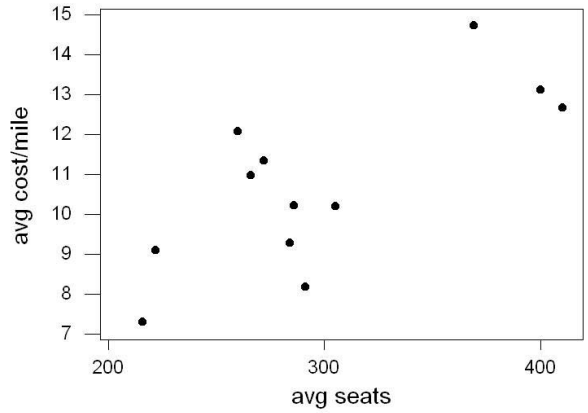

Figure 2: Avg Cost per Mile vs. Avg Seats

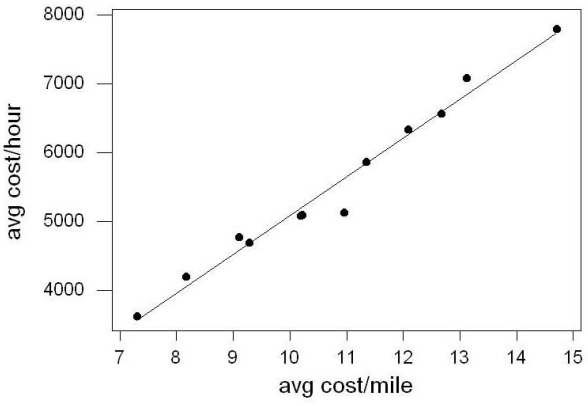

Figure 3: Avg Cost per Hour vs. Avg Cost per Mile