STATS

The Magazine for Students of Statistics

Spring 2003 • Number 37

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

3 Careers in Biostatistics: High Demand and Rewarding Work
Dennis Dixon and Julie Legler

8 Employment Advice for Undergraduate Statistics Graduates
Mary Ann Ritter

12 Statistics Education Around the Globe: A Graduate Student's Experience
Katie Makar

Departments

2 **Editors„ Column**

15 **A Day in the Life**
Database Marketing
Scott McNitt

18 **Student Projects**
Unseen, Unfelt, and Understated: The Dangers Posed to Children by Arsenic-Treated Lumber in Playgrounds
Katherine D. Van Schaik

Getting Involved with Science Fairs
Joe Ward

24 **Data Sleuth**

25 **AP Statistics**
Understanding Mulitple Regression Output
Josh Tabor

A Look at the 2002 Exam
Roxy Peck

Editors' Column

Beth Chance Allan Rossman

In the springtime, students' thoughts can shift from focusing on studying and preparing for exams to graduation and launching careers. Accordingly, we focus on professional development and career advice for students in this issue.

Our lead article comes from Dennis Dixon of the National Institutes of Health and Julie Legler of St. Olaf College, who has recently returned to Minnesota after also working for NIH. Dennis and Julie provide compelling evidence that the U.S. faces a critical shortage of biostatisticians. They offer advice for students who would like to pursue careers in biostatistics, which involve valuable work and generous compensation.

In the next article Mary Ann Ritter, a retired statistician and manager from General Motors, shares her experiences by offering practical advice to help Bachelor's level statisticians to land jobs in industry. Her suggestions are especially valuable since they are based on responses to a survey of eight non-academic statisticians about what they look for in their hiring practices.

Katie Makar, a graduate student in mathematics education at the University of Texas, reflects on her experience attending a research conference as a student. She describes what she learned and how she participated in the conference, which was held in Cape Town, South Africa. See how valuable she found the conference for networking and for talking in person with leaders in her field of statistics education.

For a glimpse into statistical careers in marketing research, Scott McNitt offers a "Day in the Life" article describing his work for the Sigma Marketing Group. Scott depicts the variety of challenges and tasks that such a statistician faces on a daily basis and conveys the excitement and satisfaction that can result.

This issue's "Student Project" article comes from Katherine Van Schaik, a junior in the Discovery Math and Science Magnet Program at Spring Valley High School in Columbia, South Carolina. Katherine won several prizes at the International Science and Engineering Fair last spring, including the ASA Award for Best Use of Statistics. In this article she describes her study of the effects of using arsenic to treat lumber used in playgrounds. Joe Ward, a tireless devotee of statistics education from San Antonio, introduced Katherine to *STATS* magazine, and we include a piece by Joe that offers advice about science fair competitions for both students and statisticians.

Our AP Statistics columnist Josh Tabor picks up where he left off in the last issue, in which you may recall he was trying to predict a teacher's salary. Not content with the simple linear regression analysis that he provided last time, Josh proceeds to include more variables in his model. He explains the fundamental ideas of multiple regression, including the subtle notion of "adjusted R^2," while refining his ability to predict a teacher's salary.

Roxy Peck, Chief Reader of AP Statistics, supplies a report on the 2002 exam. As she plans for 60,000 test-takers this year, see what students have been asked to do and how they have been performing.

This issue's "Data Sleuth Mystery" was contributed by Michael Kahn of Wheaton College, who provides data appearing to suggest that smokers have greater lung capacity than non-smokers. Please see if you can solve this mystery, and consider sending us your own example of data that tell a mysterious tale.

Careers in Biostatistics:
High Demand and Rewarding Work

Dennis Dixon

Julie Legler

Biostatisticians responded quickly following the anthrax outbreak in the U.S. by modeling the potential impact of the prevention strategies employed. Their models produced estimates of the number of cases of inhalation anthrax that may have been prevented through the regimen of antimicrobial prophylaxis used after the initial attacks. Their results underscored the importance of rapid identification of exposed persons and disease surveillance (Brookmeyer and Blades, 2002).

Biostatisticians continue to work on determining optimal prevention strategies given constraints on resources such as vaccines. While the general public may not associate exciting, vital work with statisticians, biostatisticians are frequently involved in life-or-death issues of critical importance. Meanwhile, there is a serious shortage of qualified biostatisticians – demand is high and so are salaries, making it an excellent career choice for quantitatively adept students with interests in biology.

Who are Biostatisticians?

It is a little difficult to precisely define what is meant by "biostatistician." In its simplest terms, a biostatistician is a professional dedicated to addressing biostatistical problems. The term "biometric" hearkens back to the early 1900s, when it was used to refer to the field of development of statistical and mathematical methods applicable to data analysis problems in the biological sciences. In the last few decades, the methods employed in the field of biostatistics have become more well-defined and specialized, while at the same time the scope of problems addressed by biostatisticians has broadened considerably. One way to get an idea about what a biostatistician does is to take a look at their professional organizations.

The largest and oldest organization of statisticians in the United States is the American Statistical Association (ASA), which was founded in 1839 and now includes over 16,000 members in the U.S., Canada, and overseas. Interestingly, the sections devoted to biometrics and biopharmaceutical statistics constitute two of the largest groups within the ASA. One of the world's largest organizations specifically addressing biostatistical problems is the International Biometric Society (IBS). This group describes its mission as advancing "biological science through the development of quantitative theories and the application, development and dissemination of effective mathematical and statistical techniques." The IBS membership includes biologists, mathematicians, and statisticians, as well as biostatisticians. The Society's journal, *Biometrics*, reflects the wide range of interests among biostatisticians with articles on "statistical methods for the analysis of data from agricultural field experiments to compare the yields of different varieties of wheat, [to] the analysis of data from human clinical trials evaluating the relative effectiveness of competing therapies for disease, [to] the analysis of data from environmental studies on the effects of air or water pollution on the appearance of human disease in a region or country" (ENAR website).

Another international society for biostatisticians with a slightly narrower focus is the International Society for Clinical Biostatistics. This group's primary concern is the application of statistical methods in the realm of medicine and clinical research. While a large proportion of practicing biostatisticians do work in these fields, it is helpful to keep in mind that biostatis-

---

For the past 13 years Dennis Dixon (ddixon@ niaid.nih.gov) has worked as a mathematical statistician at the National Institutes of Health, primarily in support of clinical HIV/AIDS research. Before that he did cancer clinical trials work at the University of Texas MD Anderson Cancer Center. He takes office as chair of the ASA Biometrics Section in January 2004.

Following eight years in the D.C. area where she worked at the National Institutes of Health, Julie Legler (legler@stolaf.edu) is back home in Minnesota and currently the Director of the Statistics Program at St. Olaf College in Northfield. St. Olaf has a vibrant statistics program with 16–20 students concentrating in statistics at any given time. Julie's interests include undergraduate statistics education, biostatistical methods, and family activities with her 3 children and husband.

tics actually encompasses the much broader context of all biological sciences. With some minor editing on our part to accommodate this broader perspective, this Society's goals provide a useful description of what biostatisticians do. Namely, they work:

- To define and stimulate research on the biostatistical principles and methodology used in the biological sciences;
- To increase the relevance of statistical theory to all biological sciences including but not limited to clinical medicine;
- To promote high and harmonized standards of statistical practice;
- To collaborate with scientists in other closely related disciplines to advance the appropriate application of current biostatistical methods;
- To promote better understanding of the use and interpretation of biostatistics by the general public, and by national and international organizations and agencies within the public and commercial sectors with an interest in, and/or responsibilities for, public health.

What Biostatisticians Do

Just flipping through the pages of the journals of these professional organizations or of *STATS*, it is easy to discover that biostatisticians' work involves some of the most fascinating and at times controversial applications in all of statistics.

For example, in a recent article in *STATS*, Waller and Conlon (2000) describe how Geographical Information Systems (GIS) are being used by biostatisticians more and more frequently, particularly when studying environmental risks. They discuss how data from a GIS may be used to measure the relationship between race and exposure to toxic substances in an effort to determine equity or fairness in exposure to environmental risk. Such information may be used to investigate claims of environmental racism in locating landfills or sewage treatment facilities. Biostatisticians are also involved in environmental issues in a wide variety of ways, from studying ecological systems and how they are changing, to determining safe levels of exposures of certain substances for the Environmental Protection Agency and other organizations.

Virtually every area of clinical and population-based research involves biostatisticians. For example, Dr. Kathy Cronin of the National Cancer Institute (NCI) works on clarifying issues associated with the controversies surrounding mammography that have been the source of intense debate the past several years (Cronin, 1999). Her group has led a national effort to model the impact of mammography on breast cancer incidence and mortality by convening groups of preeminent biostatisticians who study breast cancer risk factors and trends. The results of their work appear in a variety of medical journals. In addition to modeling, other biostatisticians are assessing the quality and contributions of the results of past mammography studies (Berry, 1998). They are combining information from these studies (in the midst of heated discussions) so as to better understand the accumulated knowledge and evidence that these past studies provide concerning the value of mammography. Scientists, policymakers, and many others will be considering biostatisticians' work as the debate about the value of mammography continues.

Breakthroughs in treating diabetes, AIDS, cancer, and other devastating diseases involve biostatisticians working as key members of clinical research teams. One of the most interesting and important activities biostatisticians do is to help design and analyze data from clinical studies (often called clinical trials). Study design starts with identifying the main study objective and relating that objective to one or more observations to be made on the study participants. When the objective is to find out if an experimental treatment for patients with a particular disease is better than another treatment already available, the biostatistician will advise on how many volunteers should be asked to participate, how volunteers should be assigned to receive either experimental or current treatment, what specific measurements to take to rate the outcome, how to summarize the results for all volunteers assigned to each treatment, and how to reach a conclusion on the primary objective. Depending on the actual circumstances of the trial, the number of volunteers needed could be a few dozen, several thousand, or something in between.

You can be sure that a biostatistician was involved this past summer when a multi-million dollar study involving thousands of women was stopped prematurely. In this very complex experiment, one of the main objectives was to evaluate the ability of a specific form of hormone replacement therapy (HRT) in older women (already used by millions of women for other reasons) to protect against heart disease. When partway through the study the data showed the reverse, that risk of heart disease in women receiving HRT was an estimated 22% higher than in those not receiving it, this part of the study stopped (Nelson et al., 2002). This startling development affected the thinking of hundreds of thousands of women currently taking or considering HRT, and of course it had potentially serious financial implications for drug manufacturers. Examination of partial results from clinical trials is absolutely essential in terms of protecting volunteers and the public from unanticipated safety problems and reaching conclusions as soon as possible. Proper interpretation of partial results, however, is even trickier than interpreting final results and really cannot be done reliably without participation of a skilled biostatistician.

Where Biostatisticians Work

Biostatisticians find employment in a variety of settings, but primarily in academia, clinical research centers, government agencies, and private industry. A biostatistician at a School of Public Health or Medical School may teach one or two courses during the year and collaborate on a variety of research projects. There are a number of clinical research centers across the country that also employ statisticians. One of the largest is the Mayo Clinic in Rochester, Minnesota, where 22 Ph.D. statisticians, 44 Master's level statisticians, 62 data analysts, and 9 survey researchers work on over 1200 different studies at a time. The pharmaceutical and medical devices industries hire large numbers of biostatisticians at companies across the country and abroad. The federal government employs biostatisticians to analyze data concerning health and the environment. Many of the jobs are located at the Food and Drug Administration, the National Institutes of Health, and the Environmental Protection Agency in the Washington, DC area. In addition, biostatisticians involved in public health and epidemiology work at the Centers for Disease Control and Prevention in Atlanta.

Many of these organizations and agencies offer students opportunities to learn about work in their sector through summer internships. A current listing of internships is accessible from the American Statistical Association's web site.

Growth in Market Demand

With opportunities to work on such a variety of interesting and important projects, surely there must be huge numbers of young people wanting to become biostatisticians. There are indeed many, but not enough to keep up with the needs. In the last 20 years, the demand for biostatisticians has grown dramatically. Areas of increased activity include epidemiological studies, clinical trials, health services assessments, basic laboratory research, biomedical imaging and most recently, genomics. These trends are expected to continue, particularly in the areas of genomics and bioinformatics.

Although no one knows exactly how many jobs there are for biostatisticians worldwide, we do know how many vacant positions are being announced in the *Amstat News* (the official newsletter of the American Statistical Association). Most academic and government research Ph.D.-level positions are advertised in this newsletter, as are some industry positions. Smaller proportions of positions requiring a Master's degree or Bachelor's degree, and senior or management positions in industry, are also advertised there. Thus, the number of positions listed in *Amstat News* is lower than the actual number of vacant positions.

DeMets et al. (1998) reviewed the number of positions advertised in *Amstat News* during the years 1990 to 1994. As shown in Figure 1, the number of advertised positions steadily increased during that period. The number of Ph.D.-level positions advertised rose from 60 per year to over 200 per year, most of which were in academia with 29% in biostatistics and biometrics. The number of advertised Ph.D. positions during that period was more than the number of graduates by perhaps a 2 to 1 ratio. Currently, the ratio is likely to be much larger. Although no update of the DeMets et al. (1998) report is available, the October 2002 issue of *Amstat News* contained advertisements for approximately 90 Ph.D. positions in statistics, of which 48 were for biostatisticians.

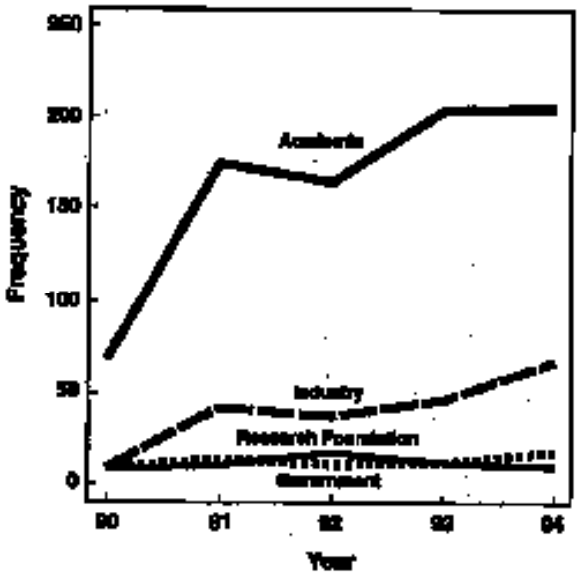


Figure 1: Frequency Plot of Source of Positions versus Year.

Most biostatistics/statistics departments or biomedical research facilities are having increasing difficulty hiring new faculty or staff. Positions may remain vacant for many months or even years. As a result, market forces have driven starting salaries up dramatically. Table 1 (abridged from Ashikaga (2002)) contains the results of salary surveys for the years 1996 through 2001. As can be seen, starting salaries have increased substantially during this period, partly due to a dramatic catch-up increment in 2000. Significant increases are likely to continue as long as the number of jobs increases faster than the number of new biostatisticians. Generally, salaries for those in industry and government tend to be a little higher than for those in academia.

Preparing for a Career in Biostatistics

Satisfying and rewarding careers in biostatistics are open to persons with Bachelor's, Master's, and Doctoral training. If you earned credit for an

Table 1: Mean Salary and Sample Sizes: Biostatistics and Other Biomedical Statistics Departments and Units (12 Months)

Rank/ Year in rank	Fall 1996	Fall 1997	Fall 1998	Fall 1999	Fall 2000	Fall 2001
Assistant						
Starting	$ 53,250 (14)	$ 57,000 (14)	$ 59,000 (12)	$ 60,905 (12)	$ 73,217 (15)	$ 67,504 (16)
1–3	$ 57,463 (14)	$ 55,000 (39)	$ 60,000 (43)	$ 61,876 (43)	$ 65,000 (58)	$ 70,000 (56)
4 or more	$ 57,463 (25)	$ 62,804 (25)	$ 63,860 (25)	$ 67,275 (20)	$ 67,500 (20)	$ 71,450 (24)
Associate						
1–2	$ 72,589 (24)	$ 69,000 (33)	$ 75,232 (24)	$ 80,625 (25)	$ 84,464 (29)	$ 85,608 (14)
3 or more	$ 68,234 (53)	$ 73,840 (63)	$ 76,200 (59)	$ 78,749 (48)	$ 86,875 (59)	$ 89,714 (48)
Full						
1–6	$ 93,067 (41)	$ 96,188 (44)	$ 98,094 (40)	$101,278 (41)	$106,947 (35)	$110,401 (32)
7 or more	$ 98,296 (75)	$106,438 (91)	$117,695 (79)	$116,827 (71)	$115,885 (70)	$125,692 (59)

Advanced Placement Statistics class in high school, that's a great start, but it's not adequate preparation for even an entry-level job as a biostatistician. To prepare for a career in biostatistics, you'll need to take undergraduate courses in statistical theory and as many courses in applied statistics as possible. Not many schools offer a course in biostatistics at the undergraduate level, but it would certainly be worthwhile if one is available at your institution. Other subjects to look for include regression analysis, nonparametric statistics, and experimental design.

Understanding and using biostatistical methods requires a knowledge of mathematics through calculus and linear algebra at a bare minimum. Even better would be to include courses in differential equations, advanced calculus, and complex variables. Naturally, computing skills are necessary, especially the ability to use some of the common database management programs and statistical analysis packages.

While there exist Bachelor's degree level positions in biostatistics, many more opportunities are available to those who earn Master's or Doctorate degrees. Those seeking employment with a Bachelor's degree will most likely be involved in managing data, statistical programming, and producing descriptive summaries of data. In many settings, recipients of Master's degrees can assume more responsibility. Project directors, researchers, and professors most often hold doctoral degrees. Those looking forward to graduate training would do well to shift the emphasis slightly towards theory of probability and statistics and mathematics, rather than towards the more applied courses. With a growing recognition that discussion of applications enriches even the most theoretical of courses, it may not really be necessary to make a stark choice in this respect. Courses in Bayesian statistics, categorical data analysis, multivariate analysis, stochastic processes, and survival analysis are helpful if offered for undergraduates.

One of the great benefits of being a biostatistician is having the opportunity to work with biomedical scientists, biologists, and others. On the job you'll need to communicate with these specialists, so it's a good idea to have an understanding of biology. Taking a college biology course as an undergraduate is highly recommended. In fact, the more you can pick up — not just freshman biology, but also other biology courses that may interest you such as physiology, microbiology, genetics, environmental science or zoology — the more you will be able to "hit the ground running" with collaborative projects.

Biostatistics Graduate Programs and Students

In contrast to the growth in demand, training of new biostatisticians seems to have reached a plateau. Since the 1970s, few new biostatistics training programs have been developed. Also, new quantitative science programs such as computer science and informatics are now competing for mathematically skilled graduate students (Shettle and Gaddy, 1996).

A survey by Hoffer et al. (2001) reports that in 2000 there were 92 Ph.D. graduates in biometrics and biostatistics. The corresponding number for 1990 was 47. Although the number of new graduates thus increased by 96% last decade, the current volume is still less than half the number of positions available in 1994. There are fellowships and grant money to fund biostatistics students that have gone unused in recent years due to a lack of students. Opportunities are out there and efforts by professional groups to increase the

supply of biostatisticians will only add to the existing prospects. An excellent place to begin your search for graduate programs is at the ASA's Schools Offering Degrees in Statistics page, where you can find a directory of schools offering degrees in Statistics in the United States and Canada. This listing includes schools offering degrees in statistics as well as biostatistics.

There are several routes one may pursue to a career in biostatistics. Some biostatisticians are trained only in statistics, at either a Master's or Ph.D. level and then learn about public health or biology or drug development on the job. This route typically involves training in theory and mathematics. Some statistics programs also allow concentrating or minoring in a subject area and students may have the choice of biostatistics while earning a degree in statistics. However, you will also find that there are programs in schools of public health and medical schools which train students specifically for careers in biostatistics. These programs supplement the statistical curriculum with a wide variety of applications in medicine and health. This can provide students with exposure to a number of different areas that they may never have otherwise been aware of. This can be a particularly good choice for students who have made a definite commitment to biostatistics.

Summary

The critical shortage of biostatisticians presents an opportunity for students who enjoy quantitative work and biology. Students taking mathematics, statistics and biology as an undergraduate will be well prepared to pursue advanced training in biostatistics. Increasing demand is anticipated for the next decade along with competitive salaries. Best of all, the field offers the professional a cornucopia of choices. Professionals can choose from a spectrum of work environments from academia to government to industry. And the potential applications are virtually unlimited from molecular biology to medical practice and pharmaceuticals to public health to entire ecosystems. The choices could be yours!

References

Ashikaga, T. (2002), "Salary Survey Results," *Amstat News*, 297, 10.

Berry, D.A. (1998), "Benefits and Risks of Screening Mammography for Women in their Forties: A Statistical Appraisal," *Journal of the National Cancer Institute*, Oct 7, 90(19), 1431–9.

Brookmeyer, R. and Blades, N. (2002), "Prevention of Inhalation Anthrax in the U.S. Outbreak," *Science*, 295, 1861.

Cronin, K. (1999), "A Day in the Life of a Statistician at the National Cancer Institute," *STATS: The Magazine for Students of Statistics*, 25, 12–13.

DeMets, D.L., Woolson, R., Brooks, C., and Qu, R. (1998), "Where the Jobs Are: A Study of *Amstat News* Job Advertisements," *The American Statistician*, 52(4), 303–307.

Hoffer, T.B., Dugoni, B.L., Sanderson, A.R., Sederstrom, S., Ghadialy, R., and Rocque, P. (2001), *Doctorate Recipients from United States Universities: Summary Report 2000*. National Opinion Research Centers at the University of Chicago.

Nelson, H.D., Humphrey, L.L., Nygren, P., Teutsch, S.M., and Allan, J.D. (2002), "Postmenopausal Hormone Replacement Therapy: Scientific Review," *Journal of the American Medical Association*, 288(7), 872-81.

Shettle, C.I., and Gaddy, C. (1996), "The Labor Market for Statisticians and Other Scientists," *National Science Foundation Report*, Arlington VA: National Science Foundation.

Waller, L.A. and Conlon, E.M. (2000), "Statistical Issues in Assessments of Environmental Justice," *STATS: The Magazine for Students of Statistics*, 28, 3-13.

Web Resources

American Statistical Association: www.amstat.org/

Listing of internships: www.amstat.org/education/internships.html

Schools offering degrees in Statistics: www.amstat.org/education/sods

Biometrics: www.stat.tamu.edu/Biometrics/

International Biometric Society: www.tibs.org/

ENAR: www.enar.org/

International Society for Clinical Biostatistics: www.iscb-homepage.org/

Employment Advice for Undergraduate Statistics Graduates

Mary Ann Ritter

Introduction

Traditionally, students pursued graduate degrees before embarking on careers in statistics. Despite this focus on advanced degrees, many people are now graduating from college with Bachelor's degrees in statistics. In the United States there were 504 Bachelor's degrees granted in statistics in 1996-97. People receiving Bachelor's degrees in statistics have choices when they graduate—they may go to graduate school, seek immediate employment, start their own business, or they may attend a professional school such as medical, law or business school. This article offers advice for students who will seek employment immediately after graduation. While the skills discussed below are the ones most likely sought by employers, rest assured that no employer expects a student to have all of them.

Where Does This Advice Come From?

Eight non-academic statisticians were asked six questions about employment for new statistics graduates. They answered the questions as well as offered their own ideas. The contributors came from a variety of companies and fields of application:

Baxter Healthcare Corporation (pharmaceutical)
Delphi Automotive (quality)
Ernst & Young (management consulting)
Intel Corporation (manufacturing)
Merck & Company (pharmaceutical basic research)
National Agricultural Statistics Service, US Department of Agriculture (government)
Southwest Technology Consultants (consulting)
Westat (consulting)

The six questions about positions for new statistics graduates were

- What jobs are out there? (Position titles)
- What do these statisticians do? ("Key job elements")
- What do students need to know? (Candidate qualifications)
- What do they REALLY need to know? (Detailed qualifications)
- What's most important? (Rating of qualifications)
- Advice or recommendations

Jobs For New Graduates

The rest of this article summarizes the answers to these questions about jobs for new statistics graduates. It is based on the opinions of a few knowledgeable people rather than a large survey. The surprising agreement among the contributors lends weight to their advice.

What Jobs Are Out There?

Very few positions were identified as exclusively for new statistics graduates. Instead, new graduates qualify for a number of positions for which statistics is just one of several appropriate backgrounds. The position title most commonly mentioned was (no surprise here!) "statistician" modified by an adjective describing the field of work. Some of these titles were:

Statistician
Biostatistcian
Agricultural statistician
Survey statistician

Mary Ann Ritter (ritterma@ix.netcom.com) is retired from General Motors where she held management positions in strategic planning, production, materials management and engineering. Before working for General Motors she was employed by MIT and SRI International. She holds a B.S. degree in mathematics and an M.S. degree in statistics from Stanford University, and an S.M. degree in management from MIT. Mary Ann is currently assistant principal violist with the Livingston Symphony Orchestra and violist with the Ypsilanti Symphony Orchestra.

This article was adapted from "Advice From Prospective Employers on Training BS Statisticians," The American Statistician*, February 2001, Vol. 55, No. 1, 14–18, by M. A. Ritter, R. R. Starbuck and R. V. Hogg.*

Mathematical statistician
Automated data processing statistician

Other common positions were:
Staff (in a consulting organization)
Programmer
Analyst

What Do These Statisticians Do?

Employers often describe what an employee is expected to do by listing "key job elements" that are the major responsibilities assigned to the person in the position. These are often used by employers to describe positions at job fairs, in on-campus recruiting notices or newspaper advertisements, or in formal personnel records.

Many companies use terms in a special way inside their company or industry. To find the best employment opportunity, students should understand the specific terms as different employers use them. For example, if a job posting reads "responsible for programming PC's," does this mean programming personal computers in an office environment or programmable controllers on the shop floor? Or does it mean responding to problem communications from suppliers or customers? The way to find out is to ask someone knowledgeable in the industry. Company interviewers, faculty advisors and recent graduates in the same industry are good sources for this information.

The contributors identified key job elements in three categories:

Statistical (specific statistics theory or methods)

- Apply statistical methods (the specific method varied with the industry)
- Apply statistical theory
- Collect, analyze, interpret data
- Perform general statistical consulting
- Review and diagram processes
- Prepare sampling frames
- Draw samples

Technical (mathematical, engineering, or computer-related activities)

- Write SAS computer programs
- Use databases
- Conduct web-based searches

Non-statistical (activities outside statistics methods or theory)

- Write reports
- Make presentations
- Participate in teams

What Do Students Need to Know?

Key job elements describe a position. Qualifications describe a candidate or what a candidate needs to know. Usually the list of qualifications describes the "ideal" candidate. Employers often do not find a person who is an exact match, but they attempt to select someone who comes as close as possible on the most important or required areas.

The formal qualifications mentioned by the contributors were:

- Bachelor's degree with two to four years' experience
- Major in math, statistics or operations research
- Master's degree strongly preferred (MS with no experience was seen as equivalent to bachelor's with two to four years experience)
- Minor in the field of application (examples: a science, social science or engineering)
- Specific statistics course work (more about this in the next section)
- Communication skills (written and oral)
- Computer programming skills (SAS was mentioned most often)

There was considerable agreement on these qualifications among the contributors despite their widely different experience. Almost all of them mentioned that a statistics degree was only one of several qualifying degrees, that an advanced degree was highly desirable, and that knowledge of the subject area was very important. Students can demonstrate subject matter knowledge by taking additional courses (more than just the introductory course—perhaps even a concentration or a minor) in the area or by summer or co-op employment in the field while still in school. The best solution would be to have both course work and related employment.

This list begs the question about whether a statistics Bachelor's degree really is enough to find a first job after graduation. The answer is a definite yes based on the contributors' and author's experience. It also immediately raises the big question for all people looking for their first job: how do I get the experience required for the first job until I've had my first job? This question is not unique to statistics degree holders and the advice section at the end of this article offers some answers.

What Do They REALLY Need to Know?

The formal qualifications listed above are very general but were mentioned by almost everyone. The contributors also offered more detailed descriptions of qualifications, but these tended to be specific to a firm or industry.

The detailed qualifications below (in no particular order) were mentioned by more than one contributor, but not by all. They suggest specific courses, class proj-

ects or work study experiences that students might have during college.

Statistical

- Analysis of variance/general linear models
- Simple analysis methods
- Reliability statistics
- Survival statistics
- Variance component analysis
- Variance propagation
- Acceptance sampling
- Exponentially weighted moving average
- Design of experiments
- Nonstandard experimental design
- Graphical analysis (box and whiskers, etc.)
- Statistical process control
- Sampling
- Principles of statistics and variation
- Survey methods and techniques
- Research methods and techniques
- Data collection/handling
- Limitations of methods
- Statistical experience/hands-on work

Technical

- Tolerancing
- Measurement capability analysis
- Calibration
- Statistical package (especially SAS, although S-plus and Minitab were also mentioned)
- Database programming/structure/large database experience
- Mathematics including advanced calculus, linear algebra
- Subject matter knowledge

Non-statistical

- Written communication
- Oral communication
- Work organization
- Consulting (practical experience preferred)
- Meeting participation (agendas, minutes, etc.)
- Team membership/collaboration
- Interpretation of statistics to non-statisticians

That is quite a list! No single student can or will graduate knowing about all of the topics. How can students know what are the most important items for them?

What's Most Important?

To help answer this question, the contributors were asked to rate the qualifications in importance on a scale from 1 to 5, with 5 being the most important. They consistently gave two qualifications the highest importance:

- The statistical methods most often used in their field of application. This was a different method in different fields.
- Communication skills. Written and oral communications were equally important.

Advice

The last question for the contributors was open-ended. It asked them to give any comments or advice they felt would be useful to graduating statistics students. Their answers covered a lot of issues both statistical and non-statistical and were offered with great energy and conviction. Their advice can be summarized in eight points.

Experience

Get as much experience with applied statistics as possible while still in school.This may be in the form of class projects, co-op experiences, work-study, or internships. Work in a campus consulting center if at all possible.

Employers often use summer hiring to check out prospective full-time employees. Students can use summer employment in a similar manner by looking for summer work in a field or at a firm that holds long-term interest for them. Doing this also helps overcome the "experience required" hurdle that people face seeking their first full-time job.

Field of application

Learn about statistics as used in a specific field. Take a minor or introductory course sequence in a field of application such as biology or marketing.

Find out what fields the statistics faculty members have worked in. Ask them which methods are most used in that field. Ask statistics instructors what fields use the methods they are teaching. When visitors come to campus to give talks, attend the talks and notice what methods they use. Ask them what methods are used in their fields.

Acquire experience in the specific field. It is most useful if the experience involves statistics, but it is also useful to acquire any experience in an industry or firm that is of long-term interest. For example, it is valuable experience to have worked in the tooling or production department of a manufacturing plant in a summer or co-op job if a student wishes to seek a position as a quality statistician after graduation.

Teams

Get experience working in project teams. Learn to fill the different roles in teams (member, leader, facilitator, etc.) Take a class in organizational behavior that explores team dynamics. Sign up for classes with term projects assigned to teams. Learn to become comfortable in roles that are not familiar. "Take charge" people should practice facilitating rather than leading; "followers" should volunteer for leadership roles.

Communication

Develop excellent written and oral communication skills. Take a writing class, then take another! Learn to write long, detailed reports and one-page summaries. Prepare presentations with bullet points. Become comfortable speaking in front of groups. Offer to make the presentation for team projects. Learn to convey statistical information to non-statisticians.

Learning

Plan to continue learning. Take personal responsibility for continuing to learn after graduation. Any technical field such as statistics is continually growing. To stay useful, students must plan to grow with their field. This can be done after graduation by joining professional societies and attending meetings, taking classes offered at work by their employers, taking evening classes either through distance learning or local colleges, reading professional journals. Find out whether an employer values specific degrees or certification and work towards one. Ask for new assignments on the job that require expanded skills.

Programming

Learn to use a statistics application package and a high level programming language. The most commonly mentioned package was SAS. Each employer usually has a statistics package that is used at the company. Learn to use this package as quickly as possible. Learn to use as much of the capability of application packages as possible. Develop a careful, error-free approach to programming.

Data

Develop good data collection and management skills, including the use of a database management program. Work with the largest data sets available. While still in school, work on projects that require the design of an analysis plan, the collection of the data to support it and the construction of the database from the collected data. Learn data quality assurance and documentation methods. Learn to move data between application packages. If possible, take a class in data base management to supplement hands-on experience constructing and maintaining databases.

Graphical methods

Learn to think about and explain statistical analyses using graphical methods. Learn to use the graphing routines of statistics application packages. Learn to present an analysis using only graphical methods. Learn as many graphical forms as possible, their appropriate use and their limitations. Incorporate graphical summaries in any analysis prepared for class, projects or work-study, even if this is not a specific requirement.

This advice is not significantly different from advice offered in previous articles. What may be different is the increased emphasis on computing and database skills and the very heavy emphasis on the need for real-world experience and the non-statistical skills of communication and team participation. Without these skills and experiences in their toolkits, new statistics graduates will not be competitive for employment, regardless of the field in which they seek employment or the statistical methods they have learned.

Summary

The undergraduate statistics degree has received relatively little emphasis at most institutions compared to graduate statistics degrees. However, it is a degree with great potential and one that offers students much flexibility in their career choices. The advice offered in this article should make the job-seeking process a little easier for students graduating with undergraduate degrees in statistics.

Statistics Education around the Globe
A Graduate Student□s Experience

Katie Makar

Background

I am a doctoral student in mathematics education at the University of Texas – Austin, concentrating on the statistical reasoning of secondary teachers. Although there are no faculty members at my university concentrating on research in statistics education, I have been fortunate enough to work on a grant with my advisor Jere Confrey, who provides a lot of opportunity for me to focus on such research. Specifically, we work with secondary teachers (inservice and preservice) helping them interpret the data from their students' state test. Dr. Confrey has tailored the project to allow me to work with the teachers to teach them statistics using a piece of learning software called *Fathom* (Finzer, 2001). I had presented some work on the project in 2001 at the Second International Research Forum on Statistical Reasoning, Thinking, and Literacy (SRTL2) in Australia from a paper I worked on with Dr. Confrey. I'm currently collecting data for my dissertation and plan to graduate in December 2003.

Before beginning my doctoral studies, I taught high school mathematics for 15 years, first in the U.S. and then at international schools in Nepal and Malaysia. My academic background (Bachelor's and Master's degrees) is in pure mathematics, but through encountering increasing applications of data analysis in the school curriculum, the International Baccalaureate program, and finally AP Statistics in my last year of teaching, I began to learn statistics and found it fascinating. In particular, I found it was a real lever for me to change my teaching practices towards a more student-oriented approach. In addition, I found it helped me create an entirely different mindset about mathematics and teaching mathematics. Statistics has all of the elements of reform-based instruction that I had struggled to bring into my classroom: deep contextual connections, engaging opportunities for student inquiry, strong integration with technology, the struggle with ambiguity and messy authentic problems, and the prospect for students to debate and defend their thinking with evidence. I felt that if learning statistics could help me as a teacher to change my thinking about teaching mathematics, perhaps it could have a similar impact on other teachers. To do research in this area, I first needed a better grounding in statistical thinking and reasoning, and this is where my research is currently taking place. My dissertation work deals with teaching secondary preservice math and science teachers the statistics they need to interpret assessment data. The context of assessment data is a critical one for teachers in the U.S. who are faced with increasing pressures to improve their students' test scores. My doctoral thesis examines how these teachers develop statistical reasoning through opportunities to conduct inquiries with student assessment data as well as how that reasoning impacts their understanding of equity.

Katie Makar (kmakar@mail.utexas.edu) is a doctoral candidate in Math Education at the University of Texas – Austin. She taught high school math in California and Oregon, and then went overseas for 9 years, teaching in international schools in Kathmandu and Kuala Lumpur, where she was department chair. She loves to travel, play Indian rummy, and do puzzles with her 5-year-old daughter.

ICOTS-6

Last summer, I attended the Sixth International Conference on Teaching Statistics (ICOTS6) in Cape Town, South Africa. ICOTS is held once every four years and is one of the main conferences for members of the International Association of Statistics Education (IASE). The focus of the conference is split between statistics education in the schools and statistics education for professionals and the workforce. As a second year doctoral student in mathematics education, the chance to attend this conference was a great opportunity. The highlights of the experience were reconnecting with researchers in statistics

education that I've met at other conferences, meeting many new people, having the opportunity to present at the conference (even though I was SO nervous!), learning about new research in the field (which really helped when it came to writing my qualifying examination!), and of course, getting a chance to visit the beautiful city of Cape Town.

First, I had to find money to attend ICOTS. It was an expensive endeavor to be sure (the airfare alone was over $2000), but I'm a strong believer in my own professional development. I have also found that the contacts I had made at other conferences have really paid off professionally, so I considered the expense a good investment. In the end, I was able to find unexpected departmental funds at the last minute to cover almost half of my expenses, which was a great relief to my budget. Although it is usually harder to get funding for international conferences, many departments have some funds available for graduate students to attend local conferences, particularly if they are presenting.

I arrived in Cape Town a few days ahead of the conference to spend a little time exploring and to have some time to get over my jet lag. During that time, I rented a car and toured some of the local vineyards, did some local sight seeing, and took some scenic drives around Cape Town. By the time the conference finally started, I was ready to see some friendly faces after three days of solitude. One of the things I most looked forward to was regenerating friendships I had made at previous conferences and hopefully making some new ones. I was able to meet up with people that I knew the first night of the conference and went out to dinner with a group, allowing me to meet a few new people. That made the rest of the conference much easier, because I could usually find someone to talk with during the breaks or meals, and it also gave me more opportunities to meet other researchers and (a few) graduate students.

The setting at ICOTS was fairly informal, and with only a few hundred attending the conference, it was relatively intimate. There were so many sessions though! I had a really difficult time choosing. Each session was organized around a central theme, with three to five presenters and a session chair, who generally led the discussions. Presenters spoke for 15–20 minutes followed by five or ten minutes of discussion. I tried to pick a theme for myself and balance areas that I'm already interested in, such as sessions on teaching variation and distribution, with sessions on areas of research that I wanted to learn more about, like teaching multiple regression. Most of the sessions, lasting between one and a half to two hours, were wonderful, and the quality of the presentations was much higher than most of the education conferences I've attended in the U.S. I did get a bit sleepy in the afternoons from jet lag and cold medicine, and I missed one or two sessions for a nap. Fortunately, the conference proceedings were on CD and handed out the first day of the conference, so you could read abstracts or full papers of sessions you didn't attend or wanted to get references from, or to help you decide which sessions to choose from.

My favorite session was "Research-based Design and Use of Software for Teaching Statistical Concepts." This was organized as a "double session" (two consecutive ninety-minute sessions). The first speaker was Arthur Bakker, a doctoral student (one of the very few at ICOTS) from the Freudenthal Institute in the Netherlands. I had met Arthur in Australia the summer before at SRTL2, so it was a pleasure to catch up with him again. His presentation focused on the categorization of statistical software into what he termed "route-based software," where the trajectory a student experiences is fairly limited," and "landscape-based software" that is more open and multipurpose, and he discussed the difficulties and opportunities each type provides for teachers. The next speaker was Tim Erickson from Epistemological Engineering in California. He talked about interesting investigations with *Fathom* that link mathematics and science. Bill Finzer from KCP Technologies then discussed some of the design challenges in creating and expanding *Fathom* in a commercial environment, while still maintaining a strong connection to the development and enhancement of statistical thinking and reasoning. I had met Bill before, but this was a new setting – discussing *Fathom* as a research tool with other researchers. Most *Fathom* users are secondary teachers (as I was when I first encountered it), so it was a thrill to be able to watch and participate in a discussion led by the software developer himself. Similarly, I had used *Tinkerplots* (Konold and Miller, 2001) in a research project that examined how to teach exploratory data analysis to a group of urban upper elementary students, but in the second session, I had my first opportunity to hear Cliff Konold, from the University of Massachusetts and the creator of *Tinkerplots*, talk in person about its purpose and how it was used. He focused on how *Tinkerplots* could be used with middle school students to understand covariation without the use of scatterplots. Then Andee Rubin from TERC, an education research and development organization in Massachusetts, talked about her research with teachers using *Fathom*. Lastly, Dave Pratt, from the University of Warwick in England, led a wonderful discussion after the presentations about the process of linking software design with researchers in the field.

The room was packed and the discussion very animated. Most of the other sessions were more formal, so this one was really exciting to participate in. It was incredible for me to have all these developers in the room at the same time and to be able to witness and even participate in their exchanges. We discussed where software in the field is headed, or where we wish it could be headed, what opportunities these types of software pro-

vide for doing research in statistics education, and how their developers can work in partnerships with researchers.

The social events were another one of the major highlights of the conference—when I was able to find other people to talk to, that is. When I couldn't find anyone, it was pretty lonely, especially since there were not a lot of other graduate students there (no surprise, given the price of the airfare) and I'm pretty shy when it comes to meeting people for the first time, especially experienced researchers. The conference tours were also a wonderful opportunity to meet people. During one of the tours, I sat with Iddo Gal, whose work I've admired since I first started doing research in statistics education. He was so friendly and gave me a lot of advice about looking for jobs. He gave me names of people to contact, even several months later through email. Meeting him was certainly one of the highlights of the trip.

Conclusion

It is always such an interesting experience to meet the people behind the papers you've been reading in graduate school. At first I always feel a little starstruck, but I'm always amazed at how nice and how down-to-earth almost all of the people are that I've met, no matter how well known they are. Getting to know the community in statistics education has been one of the greatest appeals for me being in this research area. As a relatively young field of research, there is a great amount of encouragement for graduate students. And because I lived overseas for many years, and plan to again, I've really enjoyed how international the field is (although it makes the conferences awfully expensive!). Getting to know the other researchers in the field has also really helped me, looking back, as I now begin the process of applying for jobs.

It might sound as if going to conferences is all about meeting people and making connections; that's partially true. I think one of the main reasons I do attend conferences like ICOTS is to meet other researchers in my field. The people that you get to know will be your future colleagues after you graduate, particularly if you decide to continue conducting research. Getting to know people early on helps you better understand their work, be in touch with new findings in the field, ground your research in areas that are interesting, and gain important feedback and direction for your own work. I highly recommend trying to connect with other researchers and graduate students close to your own area of research as early as you can in your program, especially if you are in an area of research that doesn't match with other professors at your university. Meeting other researchers in statistics education has given me additional literature resources (before they're even published!), opened doors to opportunities to publish and present at conferences, and helped me better understand the field. I have kept in contact with some of the researchers I have met at ICOTS and SRTL and look forward to reconnecting with them at later conferences. A few of them have even become good friends.

I look forward to ICOTS7 in Brazil! That gives me until 2006 to save up for the airfare…

References

Finzer, W. (2001), *Fathom! (Version 1.16)* [Computer Software]. Emeryville, CA: KCP Technologies. www.keypress.com/fathom

Konold, C., and Miller, C. (2001), *Tinkerplots (v. 0.45)* [Computer Software]. Amherst, MA: SRRI, University of Massachusetts. www.umass.edu/srri/serg/projects/tp/tpmain.html

Web Resources

ICOTS6: www.swin.edu.au/maths/iase/ Post_ICOTS6.html

Presentations from Session 7F, "Research-Based Design and Use of Software for Teaching Statistical Concepts": ictos6.haifa.ac.il/scientific_program/sessions/session7f.htm

Epistemological Engineering: www.eeps.com

Fathom: www.keypress.com/fathom

KCP Technologies: www.kcptech.com/

SRTL-3: tc.unl.edu/srtl/

Systemic Research Collaborative for Education in Mathematics, Science, and Technology: www.syrce.org

Tinkerplots: www.umass.edu/srri/serg/projects/ tp/tpmain.html

A Day in the Life of a Database Marketing Statistician

Scott McNitt

Background

Market research is a dynamic field with many challenges as well as opportunities to apply statistical analysis techniques. It is rewarding, as it can directly benefit both consumers and businesses. Consumers benefit as they realize a richer range of choices in the marketplace for products ranging from medicines to automobiles. Businesses benefit as market research can facilitate a better understanding of the needs and desires of consumers. It can also help businesses measure what drives positive results from their own processes, as well as the challenges that competitors bring to their market segment. What is most refreshing for me is that the statistical analyses I am involved in are almost always turned into actionable marketing plans as opposed to studies that become "shelf-ware."

With the arrival of networked, computer server technology, and the focus on measuring every aspect of doing business, there is more data available for market research activities than ever before. Thin profit margins and a highly competitive marketplace drive the need for what we term "business intelligence," achieved through the right statistical analyses. There are many aspects to market research, some more data-intensive than others. A person in this field can be involved in developing the brand recognition for a product, in new product launch activity, in convening focus groups, in new customer acquisition, or in using econometric techniques to develop the most effective marketing strategies (to name just a few).

Some approaches to market research include the utilization of "nonpublic personal information." Purchasing additional data can enhance information that a company has about its customer base, whether that is individuals, households or businesses. These external "overlay" data contain demographic, purchase behavior or other characteristics such as lifestyle interests. Some recent events that have impacted the field of market research include the passage in 1999 of the Gramm-Leach-Bliley Act amendments that require all financial institutions to provide consumers with information regarding the use and sale of collected personal information, with a mechanism for opting-out of the institutions' shared information practices. Further, the tragedies of 9/11 and the anthrax-contaminated letters have somewhat chilled direct marketing efforts in general.

The company I work for, Sigma Marketing Group, focuses on database marketing services, acting in a partnership role with our clients. We are a consulting group that works with a variety of businesses. These companies are involved in financial services, (e.g., Bank One), consumer packaged goods (e.g., Proctor and Gamble) and the automotive industry (e.g., General Motors). Some are focused on Business to Consumer transactions (B2C), while others sell their products and services to other Businesses (B2B).

We use a wide range of data sources for analytical purposes. Some are the company's own, such as internal management systems, while others are data developed for them by external vendors. If called for, Sigma also develops databases appropriate for analysis using data warehousing techniques. This is sometimes necessary, as a company's internal data sources are frequently developed to run the company, not to serve statistical consultants. While snapshots of data captured over short periods of time may be sufficient for enterprise management, analytic goals may dictate the need for a substantial longitudinal view of the company's activities. A repository of data that has been optimized for analysis is frequently called a "warehouse." For some projects we develop and execute surveys or convene focus groups. In contrast to the role I have as an external vendor, many market research

Scott McNitt (smcnitt@rochester.rr.com) has been a statistical modeler at Sigma Marketing Group in Rochester, NY for the last two years. Prior to this he was director of research for the Texas Health and Human Services Commission. He received his B.S. in Applied Statistics from the Rochester Institute of Technology and his M.S. in Statistics from the University of Connecticut.

professionals work inside of companies.

My primary role is that of a statistical modeler, responsible for developing predictive regression models based on the goals of our clients. These models can be focused on predicting customers who are most likely to cancel, to buy other products offered from the same company, or to trade up to a higher level of product. I'm also involved in strategy surrounding the development of analytic products, in consultation with other teams within the company and, more recently, in efforts to grow our business by discussing potential analytic opportunities with current clients. Other statistical activities include survey design, survey data analysis, correlation studies, marketing campaign back-end analysis, customer (or market) segmentation, and statistical forecasting. My current work demands substantial SAS database programming as well as statistical analysis.

Sigma uses a team approach for each client's work, with different essential skill sets represented on each team. Besides statisticians and econometricians who are responsible for the statistical modeling, teams include information analysts, typically SAS or SQL programmers who develop analytical datasets for the use by the modelers. Other team members include account specialists of different flavors, some with specific industry expertise, who do strategic and tactical planning as well as meeting and task coordination.

A ‡Typical· Day

Although the following description is not of an actual day, it incorporates all real events that I typically experience.

8:30 AM

I arrive at work, boot up my computer and warm up my brain with some tea. The first course of business is to check new e-mail and review my schedule. Some mornings there are only a few new messages, but if there is an impending deadline for a client project generating a lot of activity, then there could be a flood waiting for me. While e-mail needs to be attended to, it can eat up a substantial part of the day. E-mail in the work place (as well as voice mail) has facilitated the ability for quick responses, but it has also raised the expectations of those waiting to hear back. Thus, there's a balance that I'm constantly working to find, prioritizing what needs to be responded to and how quickly versus my ongoing work of the day.

I'm currently involved in the development of a logistic regression model that predicts whether or not a customer will cancel a specific type of contract. The client for this project sells hardware and services to other businesses. These contracts are complex and involve multiple types of equipment and many employees, generating hundreds of millions of dollars in revenue for the client. Our client may write one or multiple contracts with each of their customers. Distilling a complex business process down to a 0/1 response variable (the occurrence or not of an attrition event), has been challenging. However, through numerous consultations with our client these details have been hammered out. Other work on this project entails the development of potential covariates that are associated with the characteristics of the contract. A programmer has been preparing data on possible predictor variables related to equipment performance and usage. Before the final model development begins, we will join our two datasets.

9:10 AM

After responding to the most important e-mails, and printing out a few others (yes, even though it's 2003, I sometimes use paper copies of e-mails as documentation), it's time to code. I use a server version of SAS to access MS SQL Server database tables, join them using the appropriate keys and finally construct potential covariates for the logistic regression model. I test associations using correlation analysis and chi-square tests. Later in the model development process I'll use CHAID (Chi-square Automatic Interaction Detection) to explore potential interactions. CHAID diagrams, which the user can build interactively using statistical software, allow a graphical look at what different subgroups may be significantly different as measured by the response variable. We use lift charts formed by looking at stacked decile rankings to measure performance of the regression model.

For this project, I'm interested in determining to what extent different measures of machine usage may be related to both the response variable (attrition) and to each other. A high correlation (as determined by examining a matrix of Pearson correlation coefficients) is found between some of the machine usage metrics, which is undesirable because correlation among covariates contributes to multi-collinearity in a regression model. I'll use caution in deciding which of these usage variables to include in the model.

I run my programs in a networked server environment where both the data and software reside. With this structure, the power of each individual PC is not very relevant as it's the server that is the workhorse, storage unit and software source. Resources can be focused on the servers to maximize performance. It reminds me of working as a SAS programmer for IBM in the mid-80's as a Rochester Institute of Technology intern (or "co-op"), where we used dumb terminals connected to an all-powerful mainframe. "Back to the future" as it were.

10:30 AM

I'm off to a meeting regarding a potential new project my team has been approached about. The client is interested in anchoring an internally derived customer

satisfaction measure to some concrete outcomes such as lower attrition and increased revenue. Included in the meeting are two econometricians from other teams who have done similar work with both Sigma and past employers.

On Fridays this time is spent at a meeting of senior staff for all teams managed under a principal owner of the company. I represent the analytical (or as we call it, "knowledge mining") team, and there are also representatives from information technology and the major account managers. We discuss revenue projections, issues for each of the team members, and higher-level strategy.

11:30 AM

I check in with the SAS programmer who is primarily responsible for supporting my statistical modeling work. She had left a voice mail and needed clarification on some business rules for combining datasets with a series of complex overlapping dates. Fortunately, she is versed in SQL and SAS macro programming, and I appreciate our productive collaboration on this project.

Because much of the data used for this project was not originally meant for purely analytical purposes, we need to spend resources on appropriately preparing the data for modeling. Unlike my educational experience, I have always spent a substantial part of all my professional placements working with data to clean and develop a "cut" appropriate for statistical analysis. Therefore, having the ability to use at least one programming language is a great tool for a working statistician.

Lunch

Because we are a relatively modest sized company in our own stand-alone building, there are no cafeteria facilities onsite. Luckily, there are numerous options just a short ways away. I have lunch with a colleague who is in the statistical modeler role on another financial services team. We opt for a diner with a little something for everyone. Sometimes lunch turns out to be the best time to meet with a client as everyone's schedules are busy.

1:00 PM

I do some more programming as well as some 2x2 table chi-square testing.

2:10 PM

I have a quick conference call with our client regarding how to derive a metric focused on billing quality for use in the regression model on contract cancellation.

There is great value in looking at how "tight" a support process is from the consumer's point of view, and not only the performance of the equipment or the service a customer may receive. For example, incorrect invoicing might very well negatively impact a customer's ongoing loyalty with our client's products. Of course, the modeling process will bear this out.

2:25 PM

From there it's back to an informal meeting with some programming staff on deriving a control group for one of the waves of our client's scored customer records which go out to their sales force. All currently active products are scored using a previously derived regression cancellation model. Those in the top two deciles (the 20% most likely to cancel a service contract) are flagged, and the relevant sales staff will be notified electronically. We are holding back a small, statistically derived random sample of this data to serve as the "control group." Their rates of cancellation will be compared to the risk group that our model identified.

4:00 PM

Once a week at this time the Knowledge Mining team meets, giving a chance for all the statistical modelers and analysts to get together with each other and the director of modeling and market research. We discuss issues for each team as well as those affecting the entire company. Every few meetings a team member will present a synopsis of recently completed work. The discussions will focus on the more technical aspects of the project. There are plans to have staff present relevant journal articles in our quest to stay abreast with recent market research and statistical developments.

5:00 PM

I do some diagnostics on a table of updated revenue data that I need to use in the regression model. I compare monthly averages between the new dataset and what was previously used and find the new table has only approximately half as much data as the old. I develop a quick graphical presentation for a technical member of the client's team showing the discrepancy. I e-mail the team leader requesting that we meet as soon as possible to get this issue resolved. Hopefully, this issue will be cleared up tomorrow.

Although I usually finish my work during regular work hours, there are periods when I need to be "flexible" with my time in order to complete big projects. In other words, overtime. Fortunately, our IT group has recently developed a VPN (virtual private network) allowing me access to all the relevant data and programs from home.

6:00 PM

It's time for the short trip home, where I will undoubtedly be mugged by my two young children the moment I walk in the door. After a day in front of the computer, it's time to fire up a different part of my

Unseen, Unfelt, and Understated

The Dangers Posed to Children By the Use of Arsenic-Treated Lumber in

Katherine D. Van Schaik

Background

In Tallahassee, Florida, the Environmental Working Group and the Healthy Building Network recently demanded that the Florida government immediately ban the use of lumber that is pressure-treated with a preservative called chromated copper arsenate. On May 23, 2001, these two groups, in conjunction with many other environmental groups, released a study that enumerated the risks posed to children due to exposure to lumber that is treated with preservatives containing arsenic ("Coalition: Ban treated wood," 2001). Chromated copper arsenate, or CCA, is commonly used to treat lumber to prevent decay due to insects and fungi. The uses of CCA-treated lumber are many; it's used for everything from bridges to playgrounds to picnic tables.

The preservative CCA is a mixture of chromium trioxide (CrO_3), copper oxide (CuO), and arsenic pentoxide (As_2O_5). The arsenic is a pesticide, the copper is a fungicide, and the chromium fixes the arsenic and the copper to the wood.

Lumber is treated with CCA according to its use. The more a piece of wood is exposed to the ground and to the elements, the more preservative is impregnated into the wood. For example, lumber that has no contact with the ground contains 0.25 pounds of preservative per cubic foot, lumber that contains 0.40 pounds of preservative per cubic foot is used to build playgrounds and picnic tables, and lumber that is immersed in salt water contains 2.50 pounds of preservative per cubic foot (Florida Hazardous Waste and Waste Management Department, 2000).

D.E. Stilwell and K.D. Gorny of the Connecticut Agricultural Experiment Station determined that the arsenic in CCA-treated wood leaches out of the lumber into the surrounding environment (1997). They collected 85 soil samples from below seven decks, aged four months to 15 years. Concentrations of arsenic in the soil ranged from 3 mg/kg to 350 mg/kg. The mean arsenic concentration was 76 mg/kg.

The chemical profile of arsenic from the U.S. Environmental Protection Agency indicates that chronic exposure to small amounts of arsenic may cause decreased blood cell production and nerve damage. There is sufficient evidence that inorganic arsenic compounds are skin and lung carcinogens in humans (USEPA, 1987). Direct skin contact with inorganic arsenic compounds can cause swelling, redness, and irritation (USDHHS Toxicological Profile for Arsenic, 2000).

Minimal Risk Levels (MRLs) are developed by the Agency for Toxic Substances and Disease Registry (ATSDR) to establish "an estimate of the daily human exposure to a hazardous substance that is likely to be without appreciable risk of adverse noncancer health effects over a specified duration of exposure" (ATSDR, Minimal Risk Levels, 2001). MRLs exist for acute (1–14 days), intermediate (>14–365 days), and chronic (>365 days) exposure through inhalation and oral exposure routes. As of December 2001, a MRL for the dermal route of exposure had not been identified because the ATSDR was unable to find a suitable method for deriving a dermal MRL. The MRL for acute oral exposure to arsenic is 0.005 mg/kg/day, and the MRL for chronic oral exposure to arsenic is 0.0003 mg/kg/day.

MRLs are especially applicable to sensitive individuals, such as children (ATSDR, Minimal Risk Levels, 2001). Children are at a greater risk for arsenic poisoning than adults because they are more likely to ingest soils that contain arsenic. Information also suggests that

Katherine D. Van Schaik (KDVSViking@aol.com) is a junior in the Discovery Math and Science Magnet Program at Spring Valley High School in Columbia, South Carolina. Last school year, she presented her science research project at the South Carolina Junior Academy of Science Spring Meeting, the National Junior Science and Humanities Symposium in San Diego, and the Intel ISEF in Louisville, KY. At the Intel ISEF, she received awards for first place, Best Use of Statistics, first place in the category Environmental Science from the U.S. Air Force, and second place overall in the category Environmental Science. She also enjoys tennis, fishing, volunteer work, and church activities.

children are less efficient than adults at internally converting inorganic arsenic into less harmful organic arsenic (Toxicological Profile for Arsenic, 2000).

The purpose of this research was to determine the effects of exposure time and concentration of arsenic on the wood on the amount of the preservative absorbed into skin. It is possible that children who are playing on playground structures and picnic tables constructed with CCA-treated lumber are absorbing arsenic into their skin. Children exposed to CCA-treated wood for long periods of time could absorb enough arsenic through their skin to be potentially hazardous to their health. Chicken skin was used to simulate human skin because the structural arrangements of its epidermis, dermis, and collagen fibers are similar to that of a young human child (Dr. Glenda George, personal communication, September 3, 2001). I hypothesized that the higher the concentration of arsenic and the longer the time of exposure, the greater the amount of arsenic that would be absorbed by the chicken skin.

Method

I obtained fresh chicken skins from Amick Farms in Batesburg, South Carolina, and I cut the skins into 66 squares approximately 2 cm by 2 cm. Next, I obtained sheets of glass and ten petri dishes and rinsed them with distilled water and 10% nitric acid. I purchased the pieces of lumber with preservative concentrations of .25, .40, and 2.50 pounds per cubic foot, and I sawed them into pieces approximately 2.5 cm by 3.5 cm. Finally, I collected bricks and broke them so that each piece weighed approximately 3.0 pounds, or 1.35 kg. This weight was chosen to simulate the weight of a small child, age 2–6, sitting on a playground structure or picnic table: 3 pounds of brick pressing on a 2.5 cm x 3.5 cm wood surface with chicken skin underneath the wood is proportional to a 40 pound child with 18 in^2 of bare skin, the approximate area of a small child's thigh, exposed to the wood.

Each trial consisted of an exposure duration (2, 8, or 12 hours) and a preservative concentration (.25, .40, or 2.50 pounds per cubic foot). Seven repetitions were conducted for each of the 9 (3 x 3) treatments. For each time duration, I used 22 squares of chicken skin. I placed one square (the control) into a petri dish, and I put the other twenty-one squares on top of the glass. I put seven pieces each of the 0.25 treated wood, the 0.40 treated wood, and the 2.50 treated wood on top of the twenty-one squares of chicken skin. I placed the bricks on top of each piece of wood (see Figure 1). The glass upon which the chicken skins were placed was a smooth, clean, unreactive surface.

I removed the bricks, chicken skins, and wood after the indicated amount of time elapsed. I then placed the chicken skins into three separate petri dishes according to the concentration to which they were exposed and labeled the dishes.

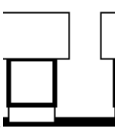


Figure 1:

I acid-digested the chicken skin samples with a medium that was 25% nitric acid, 25% sulfuric acid, and 50% distilled water. I further digested the samples with 30% hydrogen peroxide. The digested samples were then analyzed for arsenic with a Perkin-Elmer Atomic Absorption Spectrophotometer, to determine micrograms of As per cm^2 of skin surface.

Results

The experimental data partially supported my hypothesis. While the amount of arsenic absorbed by the chicken skin did increase as time of exposure increased, increases in preservative concentration did not significantly increase arsenic absorption. The scatterplots in Figures 2, 3, and 4 display the arsenic absorption amounts vs. duration times for .25 wood (Figure 2), .40 wood (Figure 3), and 2.50 wood (Figure 4). The mean arsenic absorption values for the exposure times and the amounts of preservative are illustrated in Table 2; the standard deviations are reported in parentheses.

I used two-way analysis of variance (ANOVA), conducted with the Minitab software package, to analyze the data. The results appear in Table 3. The null hypothesis that the chicken skin would not absorb more arsenic over increased periods of time was rejected. The F-statistic was 23.74, and the p-value was less than .001. The effect of the level of preservative concentration was not significant ($F = 2.17$, $p = .124$), and the effect of an interaction term was also not significant ($F = 0.84$, $p = .505$).

Figure 5 displays an interaction plot. All of the lines are approximately parallel and the differences between the levels of the dependent variable, amount of arsenic absorbed, are roughly the same distance for each value of the significant independent variable, time of exposure. In the graph we see that increasing the time of exposure substantially increases the amount of arsenic absorbed.

Because time of exposure was statistically significant, I used a Pearson product moment correlation coefficient (r) to determine the strength of the relationship between the time of exposure and the amount of arsenic absorbed by the chicken skin. The correlation for the 0.25 treated wood was $r = 0.643$, with a two-sided p-value of 0.002. For the 0.40 treated wood, $r = 0.740$ with $p < 0.001$. For the 2.50 treated wood, $r = 0.659$ with $p = 0.001$.

Extension

Stilwell (1998) used a wipe test with a polyester cloth to determine the amount of arsenic picked up by the cloth after exposure to Type C 0.40 treated wood, the same type of wood used in this experiment. He placed a polyester cloth under a cement block, and pulled this structure across the surface of the CCA-treated

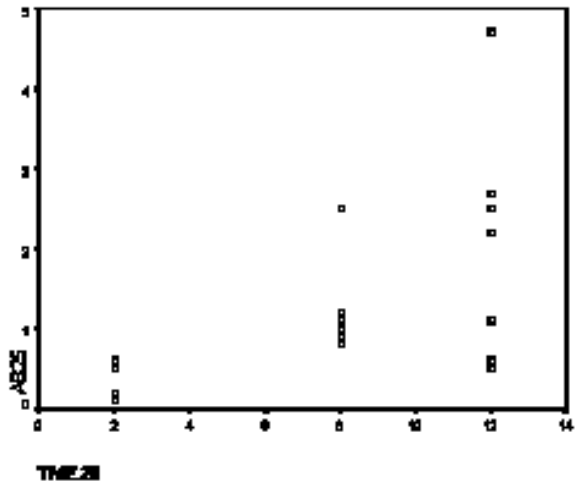


Figure 2: Scatterplot of Absorption vs. Duration for 0.25 Wood

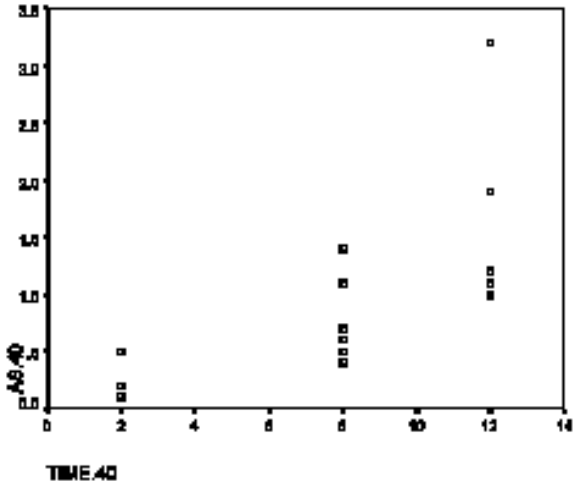


Figure 3: Scatterplot of Absorption vs. Duration for 0.40 Wood

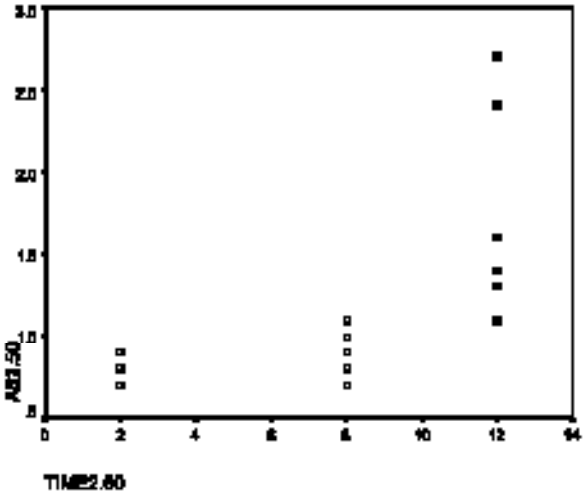


Figure 4: Scatterplot of Absorption vs. Duration for 2.50 Wood

Table 2: Mean (and standard deviation) of the amount of arsenic absorbed by skin (ug/cm^2)

Time of Exposure	Preservative Concentration (n_i = 7) 0.25	0.40	2.50
2 hours	0.27 (.20)	0.19 (.15)	0.81 (.09)
8 hours	1.2 (.59)	0.74 (.37)	1.53 (.93)
12 hours	2.04 (1.48)	1.53 (.80)	1.84 (.64)

wood five times. He repeated this four times, using 3–4 boards in each repetition. The ranges of the amounts of arsenic dislodged onto the polyester cloth are shown in Table 4. I converted his values, as shown on the right, since my values are given in µg/cm^2, and his values are given in µg/100 cm^2.

These results are fairly consistent with my mean values of 0.19, 0.74, and 1.53µg/cm^2 for each of the exposure levels for the 0.40 treated wood.

Table 5 shows the values after they were converted from µg/cm^2 into mg/kg of the child's body weight. This allowed me to compare my results to the Minimal Risk Levels for both acute and chronic oral exposure to arsenic.

With the exception of two, all of the values shown are greater than or equal to the minimal risk level for acute oral exposure to arsenic (0.005 mg/kg/day). All of the values are greater than the minimal risk level for chronic oral exposure to arsenic (0.0003 mg/kg/day). This indicates that the skin absorbed arsenic in quantities that exceeded the threshold at which the effects of arsenic exposure begin to occur in sensitive individuals.

Discussion

The purpose of this research was to determine the effects of time of exposure and concentration of the CCA preservative on the amount of arsenic absorbed into chicken skin. I found that as time of exposure increased, the amount of arsenic absorbed by the skin increased. For the 0.25 wood, for example, the mean amount of arsenic absorbed by the skin increased from 0.27 µg/cm^2 (2 hours), to 1.2 µg/cm^2 (8 hours), to 2.04 µg/cm^2 (12 hours). The data supported the above hypothesis.

The data did not support the hypothesis that the

Table 3: Two-Way ANOVA Summary

Source	DF	SS	MD	F	p
Exposure time	2	20.370	10.185	23.74	< 0.001
Preservative Conc	2	1.864	0.932	2.17	0.124
Interaction	4	1.444	0.361	0.84	0.505
Error	54	23.171	0.429		

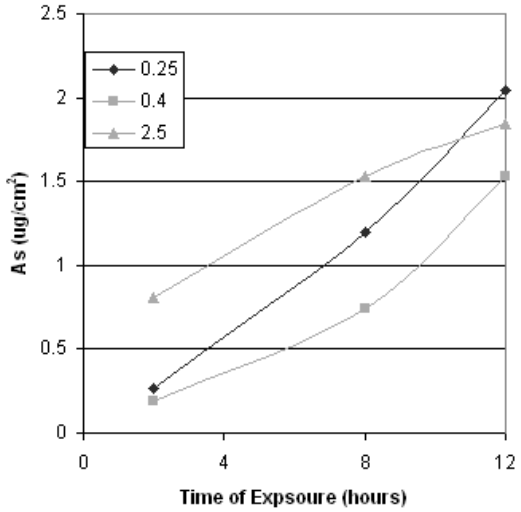


Figure 5: Interaction Plot.

greater the amount of preservative, the greater the arsenic absorption, but this could be due to the small sample sizes (7) failing to detect an effect that may actually be present. Also, several uncontrollable variables could account for this. Since I purchased the wood from two different states, the treatment process could have been different. (I could not buy the 2.50 and 0.40 wood in my hometown of Columbia, SC, so I bought them from a treatment facility in Florida.) Also, some of the wood could have been older than other pieces, or could have been exposed to more weathering.

Discrepancies in findings also could have resulted from the uneven surface of the chicken skin and the fat content in the chicken skin. It is possible that an increased fat content could affect absorption (Gensie Waldrop, personal communication, November 3, 2001).

In the future, I could dip the skins in water to remove surface arsenic before they are digested. Also, I could analyze sand samples from local playgrounds for arsenic content.

Acknowledgements

I would like to thank Ms. Gensie Waldrop, Mrs. Marilyn Senneway, Dr. Glenda George, Dr. David Stilwell, Ms. Debbie Easler, Mr. and Mrs. M.K. Weingarth, Amick Farms, and Mr. and Mrs. Douglas L. Van Schaik.

References

Agency for Toxic Substances and Disease Registry (ATSDR). (2001), "Miminal Risk Levels (MRLs) for Hazardous Substances," on the web at www.atsdr.cdc.gov/mrls.html.

American Wood Preservers' Association (AWPA) (2001), on the web at www.preservedwood.com.

Atomic Absorption Spectrophotometry (2001), on the

Science Fair Excitement

When my pre-Research teacher, Dr. Glenda George, told me and the rest of my classmates in the Discovery Math and Science Magnet Program that we would need to finalize our research project ideas before the end of the summer, I was nervous but excited at the thought of such a challenge. So, in the summer between my freshman and sophomore year of high school, I spoke with Dr. George and found a topic that both worried me and intrigued me: arsenic-treated wood was being used to build playgrounds.

Over the next six months, I wrote my paper, developed my experimental design, and carried out the experimentation and analysis. At this point, I came to what my Research teacher, Mrs. Marilyn Senneway, called the "meat" of the entire research paper: the statistics. After several attempts at finding a way to input the data so as to yield results that would appropriately reflect the data, I finally found a way, and, in doing so, learned the "significance" of statistics in research. Mrs. Senneway had continually told us that, without statistics, we would have no way of knowing what our results really were. After looking at the two-way ANOVA table, the interaction graph, and the Pearson product moment correlation table, I knew she was absolutely right. My resulting graph and tables told me more about what had actually happened between the variables than I had ever thought possible.

In April 2002, I competed at our local science fair and won the top award, and with it, the chance to compete at the Intel International Science and Engineering Fair in Louisville, Kentucky in mid-May 2002. To say that the Intel ISEF was the highlight of my sophomore year would be a gross understatement – it was the experience of a lifetime. Over 2000 students and teachers from 39 nations attended, as well as Nobel Prize recipients who spoke with students and answered questions from the audience.

The first of three award ceremonies was the day after the judging, and it was at this award ceremony that the American Statistical Association presented the award for the Best Use of Statistics. When the announcement was made, my initial reactions were shock and overwhelming excitement. The director of my magnet program, Mrs. Jennifer Richter, immediately used her cell phone to call Spring Valley High School and my principal, Dr. Greg Owings. As I shook Chapter President Bill Wunderlin's hand, I was trembling with excitement. All the hours of work and, at times, frustration, were worth it. I knew I was hooked on science, math, and exploring areas of the environment where I felt I could have an impact. I felt I had found a career path that was challenging, exciting, and very worthwhile.

Table 4: Results of Stilwell□s Wipe Test

Set Number	Amount of Arsenic Dislodged (µg/100 cm²)	Amount of Arsenic Dislodged (µg/cm²)
1	15–31	0.15–0.31
2	6–33	0.06–0.33
3	56–122	0.56–1.22
4	15–26	0.15–0.26

web at campus.murraystate.edu/academic/faculty/

Table 5: Comparison to ATSDR Minimal Risk Levels (mg/kg of the average child□s body weight)

Time of Exposure	Preservative Concentration 0.25	0.40	2.50
2 hours	0.002	0.001	0.005
8 hours	0.008	0.005	0.010
12 hours	0.013	0.010	0.012

judy.ratliff/graphite.htm.

"Coalition: Ban treated wood." (2001), *St. Petersburg Press*, May 23, A1.

Florida Hazardous Waste and Waste Management Department (2000), "What is Treated Wood," on the web at www.ccaresearch.org.

George, G. (2001), personal communication, September 3.

Gradient Corporation (2001), "Focused Evaluation of Human Health Risks Associated with Exposure to Arsenic from CCA-Treated Wood," on the web at www.preservedwood.com/safety/ccafocus.pdf.

Stilwell, D.E. (1998), "Arsenic from CCA-treated wood can be reduced by coating," *Frontiers of Plant Science*, 51, 6–8.

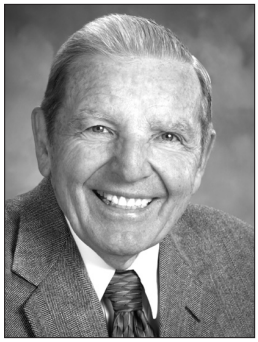
Stilwell, D.E. and Gorny, K.D. (1997), "Contamination of soil with copper, chromium, and arsenic under decks built from pressure-treated wood," *Bulletin of Environmental Contamination and Toxicology*, 58, 22–29.

U.S. Department of Health and Human Services (2000), "Toxicological Profile for Arsenic," Syracuse Research Corporation under contract no. 205-1999-00024.

U.S. Environmental Protection Agency (USEPA) (1987), on the web at www.epa.gov/swercepp/ehs/profile/1303282p.txt.

Waldrop, G. (2001), personal communication, November 3.

Getting Involved With Science Fairs



Joe Ward

Science Fairs and other science competitions are held across the United States each year with the goals of encouraging students to share ideas, motivating innovation, and showcasing cutting-edge science projects. These are very important goals not only for individual students but also for the future scientific progress of our nation. Students compete for awards and scholarships in these competitions, and the statistical analysis incorporated in their projects can play an important role in the final quality of their work. Judges at these contests frequently include statisticians.

For several years I have attended the International Science & Engineering Fair (ISEF). I always visit the ASA Special Awards winners to congratulate them. Katherine Van Schaik won not only the ASA First Special Award of $500 and a plaque, but also a U.S. Air Force First Award of $3,000, and an Environmental Protection Agency Second Award of $1,500.

I also give a presentation titled "Combining the Power of Statistics and Computers to Enhance Science Fair Projects" to Fair Directors, teachers, parents and ISEF Finalists. I present some ideas about applying computer-based statistical analysis techniques to improve research projects.

Also, of most interest is a discussion of ways to obtain statistical support for student research. I encourage participants to go to www.amstat.org to identify ASA Chapters and Chapter officers to locate statisticians in their area who might assist with statistical design and data analysis techniques. Some students have not only contacted nearby assistance but have also received valuable long-distance guidance by phone and email.

I also believe that statisticians can do more to help high school science fair participants. I suggest that statisticians consider:

- assisting high school students with statistical design and data analysis techniques that will enhance the quality of their research projects;
- providing special statistics awards and judges to select the award winners at local science-related fairs;
- offering students information about the association between winning statistics special awards and winning specific science-category awards;
- encouraging special statistics award-winners to enter the ASA Poster & Project Competitions (see www.amstat.org/education/index.html#K12) and to submit their research papers for publication consideration in journals and magazines such as *STATS*.

These activities can contribute much toward "selling" science fair students on the value of using quality statistics in their research projects.

To learn about science-related competitions, consult sources such as Science Service (to locate the Science Fairs in nearby locations), your state's Junior Academy of Science (to locate Junior Academy competitions) and the Junior Science and Humanities Symposium. Since many school districts have web pages, it is easy to locate nearby schools and teachers who are already engaged in science and engineering research projects.

Web Resources

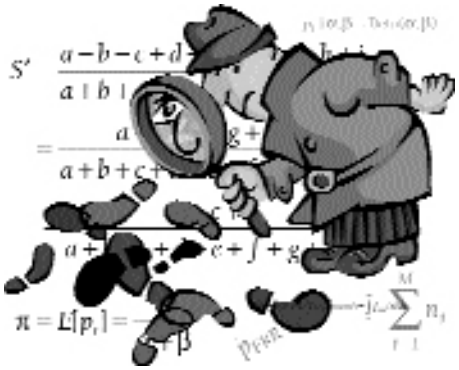
Sceience Service: www.sciserv.org

Junior Science and Humanities Symposium: www.jshs.org

Joe Ward (jwardjr@satx.rr.com) was employed as a civilian by the U.S. Air Force for over thirty years to bring the combined power of statistics and computers to problems involving personnel and training research. Starting in 1958, before computers were introduced into the school curriculum, he volunteered to present free basic computer courses for high school students and teachers who were interested in learning about the new technology. He continues to volunteer today as a member of the Board of the Alamo Regional Academy of Science and Engineering to encourage high school science research students to use the resources of statistics and computers to enhance their projects.

Data Sleuth

An Exhalent Problem



Contributed by Michael Kahn, Wheaton College (MA)

A dataset from Rosner's *Fundamentals of Biostatistics* concerns the relationship between forced expiratory volume (FEV, a measure of respiratory function) and smoking, along with several other variables. The data include information from 654 children and young adults, ranging from 3 to 19 years of age. The variables considered here are FEV (in liters), self-reported smoking status, and age (in years).

The boxplots in Figure 1 compare the distributions of the smokers' FEV with the nonsmokers' FEV.

Question #1: Using the boxplots, do nonsmokers appear to have, on average, higher FEV scores than smokers?

Question #2: Is it sensible to use these data to discuss, in isolation, the effects of smoking on FEV? In particular, would you conclude that smoking causes young people to strengthen their respiratory function? If not, can you suggest an alternative explanation for the differences in the boxplots?

Figure 2 compares the nonsmokers' and smokers' relationships between FEV and age. The "curves" are computed using *lowess* (Cleveland, 1979); they provide estimates of the (conditional) average FEV for a given age.

Question #3: Using the scatterplot and lowess curves, do the nonsmoking 16-year-olds appear to have, on average, stronger respiratory function than those 16-year-olds who smoke? 19-year-olds? 10-year-olds? Suggest some possible explanations for the inconsistencies in your answers to these questions.

The solutions appear on page 28.

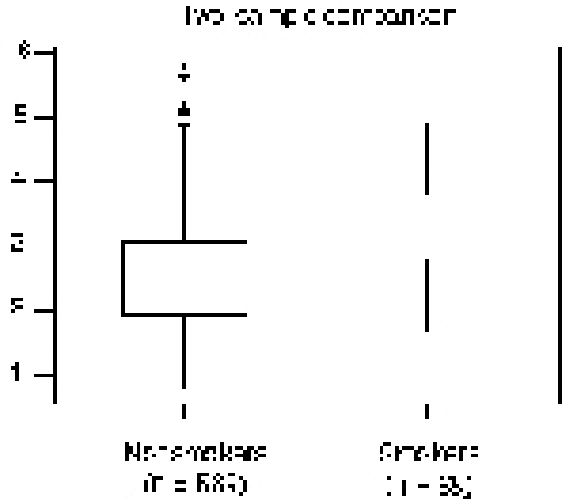


Figure 1: Forced Expiratory Volume by Smoking Status

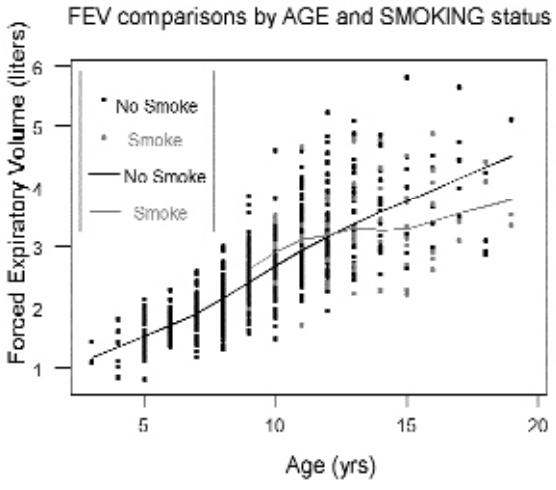


Figure 2: Forced Expiratory Volume by Age and Smoking Status

References

Cleveland, W.S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829–836.

Rosner, B. (2000), *Fundamentals of Biostatistics* (5th ed.), Pacific Grove, CA: Duxbury Press.

AP Statistics

Understanding Multiple Regression

Josh Tabor

In the last issue, we tried to estimate the salary of a prospective spouse using simple linear regression. Using the model we created (*predicted salary = 40.61 + 1.686 years*), we estimated that a teacher with 8 years of experience makes about $54,098. The standard deviation of the residuals was 4.611, which means that our predictions using this model will be off by about $4,611, on average. This was a big improvement from making a prediction based solely on the average teacher's salary ($50,882) which had a typical error of $6,490. Additionally, 54.6% of the variability in salary was explained by the relationship between salary and years of experience.

However, in addition to years of experience, there are other factors which affect teachers' salaries. These include the number of post-graduate units earned and extra duties, such as coaching or chairing a department. To include these factors in our analysis, we must use statistical software with multiple regression capabilities. There are many packages available and, fortunately, the output from the various packages is fairly standard. In this article, we will again use JMP-Intro to do our analysis. Table 1 displays salaries and years of experience for our 11 teachers and also includes three more explanatory variables, the number of post-graduate units earned, whether or not they coach and their gender.

Our first step will be to add post-graduate units as an additional predictor variable. The results of the multiple regression are shown in Figure 1.

Examining the "Summary of Fit" table shows that adding "units" greatly improves the model. The typical error (root mean square error) decreases from $4,611 to $2,673 and the percentage of the variability in salary that is explained by the model (Rsquare) increases from 54.6% to 86.4%. Of course, whenever we add an additional explanatory variable, the value of R^2 will never decrease, because adding an additional variable cannot decrease the amount of variability in the response variable explained by the model.

Thus, if we want to have a high R^2 value, we just need to add more and more explanatory variables. However, after a while, adding additional variables will cease to have a big effect on the value of R^2. Meanwhile, the model will become unnecessarily complicated. To address the tug-of-war between high R^2 values and simple models, statisticians developed a measure called *adjusted* R^2.

$$adjR^2 = 1 - \left[\frac{n-1}{n-(k+1)}\right]\frac{SSError}{SSTotal}$$

$$R^2 = 1 - \frac{SSError}{SSTotal}$$

Note: In the formula above, n = the total number of observations and k = the number of explanatory variables. Verifying the value from Figure 1,

Table 1: Data on Teachers' Salaries

Salary	39.9	47.6	49.3	51.6	47.0	46.2	48.5	51.7	58.1	56.1	63.7
Years	4	8	5	9	1	4	4	6	8	7	11
Units	8	18	39	32	35	48	63	48	68	72	58
Coach?	No	No	Yes	No	Yes	No	No	No	No	No	Yes
Gender?	M	F	M	F	F	M	F	M	M	F	M

Figure 1: Regression of Salary on Years of Experience

Summary of Fit

Rsquare	0.864295
RSquare Adj	0.830369
Root Mean Square Error	2.673482
Mean of Response	50.88182
Observations (or Sum Wgts)	11

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	364.17631	182.088	25.4758
Error	8	57.18005	7.148	**Prob > F**
C. Total	10	421.35636		0.0003

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	34.270682	2.463627	13.91	<.0001
years	1.3818053	0.305404	4.52	0.0019
units	0.1843385	0.042547	4.33	0.0025

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
years	1	1	146.31805	20.4712	0.0019
units	1	1	134.16606	18.7710	0.0025

$$adj\,R^2 = 1-\left[\frac{11-1}{11-(2+1)}\right]\frac{57.18}{421.36} = 1-(1.25)(.1357) = .830.$$

Notice that the value in the brackets will always be greater than 1. Thus, the value of *adj* R^2 will always be smaller than the value of regular R^2. Also, notice that if the number of explanatory variables increases without a corresponding decrease in SSError, the value of *adj* R^2 be smaller in the new model. Therefore, *adj* R^2 penalizes a model that adds extra explanatory variables that don't greatly increase the precision of the predictions.

In the first line of the "Analysis of Variance" table, we see that the degrees of freedom for the model have increased from 1 to 2 with the addition of the second explanatory variable. Also, the variability (measured by sums of squares) in salary that is explained by the model has increased from 230 to 364 (out of a total possible 421). The F-ratio for the model is 25.48

$$\left(F = \frac{SSModel/dfModel}{SSError/dfError} = \frac{364.18/2}{57.18/8} = 25.48\right)$$

which indicates that the model is useful for making predictions (p-value = .0003).

In the "Parameter Estimates" table, we find that our model is: *predicted salary* = 34.271 + 1.382 *years* + 0.184 *units*. Thus, we can estimate that a new teacher (with no experience and no post-graduate units) will make $34,271, on average. Also, for each additional year of experience, the model predicts that the salary will go up by $1,382 and for each additional graduate unit, the salary will go up by $184.

In the "Effects Tests" table, we find an analysis for each predictor variable. The Sum of Squares listed for each variable is the difference of SSModel when the variable is included in the model and the SSModel when the variable is excluded. For example, in Figure 1, SSModel equals 364.18 and before SSModel was 230.01. The difference of these 2 numbers, 134.17, is the additional variability that is explained by including "units" in addition to "years." Likewise, if you were to create a model using only "units" to predict salaries, adding "years" would increase the Model Sum of Squares by 146.32.

Note: These 2 values do not add up to SSModel. This is because "units" and "years" are not independent. Since there is a positive relationship between "years" and "units," some of the variability that is being explained by years could also be explained by units. The Venn diagram below shows the allocation of the Total Sum of Squares:

Figure 2: Regression Model with Coaching

Summary of Fit

Rsquare	0.960576
Rsquare Adj	0.94368
Root Mean Square Error	1.54048
Mean of Response	50.88182
Observations (or Sum Wgts)	11

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	404.74481	134.915	56.8523
Error	7	16.61155	2.373	**Prob > F**
C. Total	10	421.35636		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	32.701946	1.469389	22.26	<.0001
years	1.4512859	0.176777	8.21	<.0001
units	0.1835298	0.024517	7.49	0.0001
coaching	4.3321229	1.047761	4.13	0.0044

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
years	1	1	159.94395	67.3994	<.0001
units	1	1	132.98295	56.0382	0.0001
coaching	1	1	40.56850	17.0953	0.0044

Figure 3: Regression Model with Coaching and Gender

Summary of Fit

Rsquare	0.960796
Rsquare Adj	0.934661
Root Mean Square Error	1.659253
Mean of Response	50.88182
Observations (or Sum Wgts)	11

Analysis of Variance

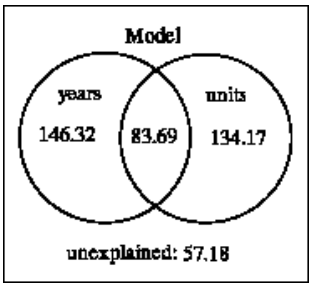
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	404.83765	101.209	36.7617
Error	6	16.51872	2.753	**Prob > F**
C. Total	10	421.35636		0.0002

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	32.771869	1.627845	20.13	<.0001
years	1.4552108	0.191602	7.59	0.0003
units	0.1835174	0.026407	6.95	0.0004
coaching	4.3657054	1.143267	3.82	0.0088
male	-0.187804	1.02275	-0.18	0.8604

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
years	1	1	158.80899	57.6833	0.0003
units	1	1	132.96417	48.2958	0.0004
coaching	1	1	40.14563	14.5819	0.0088
male	1	1	0.09283	0.0337	0.8604



The F-ratios and p-values in the "Effect Tests" table tell us if including the variable in the model significantly improves our ability to make predictions. For example, the F-ratio for "units" is

$$F = \frac{SSUnits / dfUnits}{SSError / dfError} = \frac{134.17/1}{57.18/8} = 18.77$$

which gives a p-value of .0025. Thus, including "units" gives us additional predictive power beyond what can be attributed to random chance.

Note: the p-values in the "Parameter Estimates" table are the same as in the "Effect Tests" table. Also, the t-ratios are simply the (signed) square roots of the F-ratios.

Figures 2 and 3 show the multiple regression output if we add two categorical variables: coaching and gender. To include binary variables in the model we must convert them to numerical values by coding them. For example, teachers who coach are coded 1 and teachers who do not are coded 0. Likewise, male teachers are coded 1 and female teachers are coded 0. (Categorical variables with more than two category possibilities require additional variables in the model.)

As you can see in Figure 2, including "coaching" significantly improves our model (F-ratio = 17.10, p-value = .0044). The *adj* R^2 value went from 83.0% to 94.4% and the typical error went from \$2,673 to \$1,540. The coefficient estimate for "coaching" (4.332) also tells us that coaches earn an extra \$4332, on average (the predicted increase in salary as the coaching variable changes from 0 to 1).

However, acording to Figure 3, including the variable "gender" doesn't seem to help our model (F-ratio = 0.034, p-value = .8604). Although including gender slightly increases the R^2 value from .9606 to .9608, the *adjusted* R^2 value actually decreases from .9437 to .9347. The large p-value and decrease in the *adjusted* R^2 value are indications that "gender" is not particularly helpful for predicting salary if we already know the first three variables.

Thus, our final model is:

predicted salary = 32.702 + 1.451 *years* + 0.184 *units* + 4.332 *coaching*

If our prospective spouse has been teaching 8 years, has earned 40 post-graduate units, and coaches, her predicted salary would be approximately \$56,000.

So, although we have answered the original question ("How much does my prospective spouse make?") with more confidence, the answer to the more important question ("Will you marry me?") remains uncertain…

Solutions to Data Sleuth Mystery

Question #1: From the boxplots it appears that the smokers have, in general, higher FEV scores. The median FEV for the smokers is over 3 liters while the median FEV for the nonsmokers is less than 2.5 liters.

Question #2: In any observational study in which we are looking to assess the relationship between a self-selected category (smoking), and some other measure (FEV), we should always consider other confounding reasons that might help clarify and/or explain the apparent relationship. In this case, since these data include so many very young children who do not smoke and whose lung capacity are unlikely to be as large as the older children, we should consider accounting for the age of each subject. In the actual dataset from Rosner, the subjects' height and gender are also included and make for an even more complete and interesting story.

Question #3: For 16-year-old nonsmokers the estimated average FEV is just under four liters, say 3.8 liters, while smokers have an average FEV of just over three liters, say 3.2 liters. Similarly, for 19-year-old nonsmokers the estimated average FEV is over four liters, say 4.5 liters, while smokers have an estimated average FEV of about 3.5 liters. Finally, for 10-year-old nonsmokers the estimated average FEV is under three liters, say 2.5 liters, while smokers have an estimated average FEV of about 2.8 liters. Though the standard errors of these estimates have yet to be expressed, these data do indicate that the older subjects who do not smoke have, on average, larger FEV than the smokers. The inconsistency arises when considering the youngest smokers since the estimates suggest that for 10-year-olds smoking is associated with larger FEV. This has a variety of possible explanations, some of which include the very small number of young smokers, the short time and relatively small amount that these children are likely to have smoked, and errors in self-reporting.