STATS

The Magazine for Students of Statistics

Fall 2003 • Number 38

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 Statistical Adventures in Litigation
Ulric Lund

10 A Rainy Day in Trinidad
Winston A. Richards and Jeffrey R. Wilson

14 Preparing for Life After Graduate School: A Recap of Our Personal Experience
Sumithra J. Mandrekar and Jayawant N. Mandrekar

Departments

2 **Editors' Column**

18 **A Day in the Life**
AP Statistics Teacher
Brian Kotz

23 **Data Sleuth**
Where Have All the Boys Gone?

24 **The Statistical Sports Fan**
A Streak Like No Other
Robin Lock

27 **ø-sings**
The Greatest Experiment Ever Performed on Women
Chris Olsen

Editors' Column

Beth Chance **Allan Rossman**

Lawyers have received much more frequent and favorable depictions on television shows than statisticians have. Perry Mason and *L.A. Law* have no doubt inspired many to enter the legal profession, but popular culture offers no such role models for statisticians. As you prepare for a career in statistics, though, you probably realize that statisticians have long played an important role in many legal proceedings. Moreover, the importance of statisticians in the legal system is only increasing in our information-rich (and ever-litigious) society. Our lead article in this issue comes from Ulric Lund, who began his statistics career with a consulting firm that specialized in litigation statistics. Ulric describes some of the behind-the-scenes legal contributions that statisticians make and offers suggestions for how to get involved in this type of work. He also illustrates the kinds of legal analyses done by statisticians with an example using a two-stage regression model to investigate public health expenditures attributable to the obesity problem in the U.S.

Field trips can be a fun learning experience in any course, and this is as true for statistics as any subject. Winston Richards and Jeffrey Wilson really know how to plan a memorable field trip for their students — they took their students to Trinidad! In their article they describe some of their students' experiences analyzing climatic data to model rainfall in Trinidad at various times of year.

One of the most stressful and frustrating aspects of being a student comes at the very end: making the transition from student to professional by conducting a job search. Sumithra Mandrekar and Jayawant Mandrekar offer advice for smooth navigation of this process. Their advice comes from recent personal experience, as they both recently completed their degrees and entered the job market. Two additional complications in their search, for which they also offer guidance, are that they are international students and are (happily!) married.

Brian Kotz embarked on a different career path from the Mandrekars, as he entered high school teaching after completing his undergraduate degree in statistics. Brian has taught statistics and other mathematics courses at two private high schools. In this issue's “Day in the Life” installment, he describes the variety of challenges that he faces in a typical day, shares his passion for his chosen career, and makes a compelling argument for the important work being done by high school teachers of statistics.

Robin Lock noticed a very unusual baseball streak — the Cleveland Indians recently scored exactly four runs in seven consecutive games. Always on the lookout for interesting items for his *Statistical Sports Fan* column, Robin proceeded to investigate how unusual this result really is.

Chris Olsen returns in this issue with a *μ-sing* about a book title that caught his eye. Find out if Chris thinks this book lives up to its claim of describing the greatest experiment ever conducted.

In this issue's installment of *Data Sleuth*, we invite you to solve a mystery about the declining proportion of male births in several countries, inspired by an exercise from the wonderfully-named textbook *The Statistical Sleuth*.

Statistical Adventures in Litigation

Ulric Lund

Introduction

As a student of statistics, you probably are keenly interested in career options upon graduation. You likely are aware of the many opportunities in areas such as biostatistics, pharmacology, public health, agriculture, marketing and manufacturing, working in industry or for a government agency. How about using your statistical background to support attorneys in the courtroom? Indeed, there is an exciting career choice for statisticians to be found in the legal arena. In what follows, I will describe my experience working as a statistical consultant for attorneys involved in litigation, and also present an interesting statistical application in the context of a hypothetical court case.

The Litigation Statistician

Attorneys retain expert witnesses from appropriate disciplines to support their side's arguments in a court case. A statistician may be employed during the course of a trial if an attorney's arguments require substantial data analysis or statistical modeling. The following are examples of cases that would require a statistician as an expert witness:

- Employment discrimination: Is a plaintiff's salary unfairly low due to his or her age, gender, or race?
- Census adjustment: Since the U.S. Census inevitably undercounts a fraction of the population, should it be adjusted for this undercount using sampling methods?
- Product liability: Do breast implants increase the risk of disease such as breast cancer or connective tissue disease?
- Tobacco industry: How much money should the tobacco industry pay to compensate the government for increased Medicaid or Medicare expenditures?
- Asbestos industry: How much money are asbestos workers entitled to for their increased incidence of lung cancer?
- Presidential elections: Who really won the 2000 Presidential Election in Florida?
- Antitrust, collusive agreements: What is the economic damage to a company due to a consortium of other companies' anticompetitive activity?

The educational background required for a statistical consultant working in the legal arena varies. Statisticians serving as expert witnesses typically will have a doctorate in the field. However, there are supporting roles to be played as well, creating job opportunities for graduates of all levels, including those with Master's and Bachelor's degrees. Statistical consulting firms retained by attorneys typically need to have support statisticians and statistical programmers. This is the type of work with which I am familiar, having worked in this role for two years shortly after completing my doctorate, and for which I will outline some of the responsibilities below.

In the pre-trial phase of litigation there is a process called *discovery*, during which both sides of a case turn over information and documents to the court and to the opposing side. Among these documents are the expert witness reports – reports prepared by the expert witnesses retained by the attorneys. These reports contain the scientific reasoning and analysis that the experts will testify to later during the court proceeding. At this stage, a statistician may assist the attorney in analyzing

Ulric Lund (ulund@calpoly.edu) is currently an Assistant Professor in the Statistics Department at California Polytechnic State University, San Luis Obispo. Prior to taking this position, he worked for two years at Environmental Risk Analysis, a statistical consulting firm in the San Francisco Bay area, and taught for one year at Western Washington University. His studies were completed along the California coastline, obtaining his B.A. in Mathematics from University of California, San Diego, and his Ph.D. in statistics from University of California, Santa Barbara.

the opposing side's expert witness reports in addition to developing their own side's reports.

Much of my time was spent critiquing the opposing side's expert witness reports. For this, a solid background in applied and theoretical statistics was essential since the types of analyses that I came across were varied and could be quite sophisticated. I had to be proficient in the statistical theory to understand the arguments of the opposing side's experts, and to be able to assess the validity and accuracy of the analysis put forth.

Even if the statistical arguments in an expert witness report are correct, their execution, usually involving computer programs, is still suspect. Any computer code used to generate results presented in an expert witness report must also be turned over to the opposing side during discovery. So, once I understood the theoretical arguments, I would proceed to check the accompanying computer code. This is where the detective-work really heated up, and where sometimes errors were uncovered. Familiarity with commonly used statistical packages (e.g. SAS, S-PLUS, and STATA) was helpful for this aspect of the job.

Another part of the discovery phase is the *deposition*. When the opposing side's attorney deposes an expert witness, it can often be a brutally confrontational encounter. This is routinely done, however, to clarify various aspects of the expert's report before the case goes to trial. Again, support statisticians may be involved in thoroughly preparing their expert witness for his or her deposition to reduce the chance of unexpected, and thus unwelcome, questions. Statisticians are also involved in assisting the attorney prior to the deposition to help formulate lines of inquiry used to obtain answers to unresolved questions about the expert's report, and to point out flaws and weaknesses of the expert's analysis. Following the depositions, statisticians may be asked to analyze transcriptions of the depositions, again trying to find inconsistencies, inaccuracies, and erroneous statements made by the opposing side's statistics expert. Depending on the attorneys' personalities, and again due to the adversarial environment, these can be quite entertaining to read, especially if you like courtroom dramas. You won't read anything like “Where were you on the night of August 17, 1997?” But, you may read something like “I am a bit at sea as to how you computed the monetary damages due to the plaintiff.”

As you can see, there is quite a bit of statistical sleuthing involved in this occupation. But, in addition to analytic ability, good communication skills are essential. The primary task in this line of work is to communicate statistical concepts and results to the attorney. Much of my time was spent preparing and assisting in the presentation of material at meetings with the client, informing him or her about the results of our analysis, and even giving statistical tutorials. Statistical concepts needed to be described in an elementary fashion, especially since the analyses were also to be presented to a jury at some time. The amount of tutoring the attorney needs is highly variable. I was fortunate enough to deal primarily with statistically seasoned attorneys, but this is not always the case.

The cases I worked on often employed multiple linear regression, and one concept that was quite important was the interpretation of the regression coefficients: in particular, the fact that (in the absence of interaction terms) a regression coefficient represents the predicted change in the response variable for a one-unit increase in the explanatory variable — regardless of the values of other explanatory variables. For example, in a multiple linear regression model predicting employee salary from age, gender and race, the estimated regression coefficient for the age variable predicts how much salary changes for a one-unit increase in age, regardless of gender and race. An expert witness presenting a simplistic model, as I suggest here, would be open for attack by a well-informed attorney from the opposing side. It is unlikely that salary is affected by age equally across genders and different races. The attorney would need to be informed of this weakness in the expert's model, and be aware that interaction terms should be explored as well.

The work environment is very fast paced. It was not uncommon to receive last minute requests for analyses, which made for high-pressure situations. Not only did I need to produce the results quickly, but they also needed to be accurate in case they were presented in court or in an expert witness report. These two goals of speed and accuracy are naturally conflicting, and provide for a good deal of stress. However, the fast paced work environment, working with a group of statisticians and attorneys towards a common goal, and surviving the adversarial nature of litigation, created the same sense of camaraderie experienced in team sports.

Even when statistical analyses play a large role in the litigation, the two sides' attorneys make other important legal arguments as the case proceeds, which may render the hard work of the statisticians irrelevant. It is also not uncommon for timetables to be switched and deadlines to be postponed due to legal wrangling. Several cases I worked on were either settled or thrown out of court by the judge without ever going to trial. The fruits of my hard labor often never saw the light of day in front of a jury, which was a bit dissatisfying.

But overall, I thoroughly enjoyed my work as a statistical analyst consulting for attorneys, and perhaps it could be a career choice for you as well. To prepare for this type of work, you will want to have a strong background in theoretical and applied statistics, as well as good computational abilities. Good communication skills are necessary for interacting with the attorneys, and if you are instrumental in preparing expert witness reports. Teaching or teaching assistant experience is

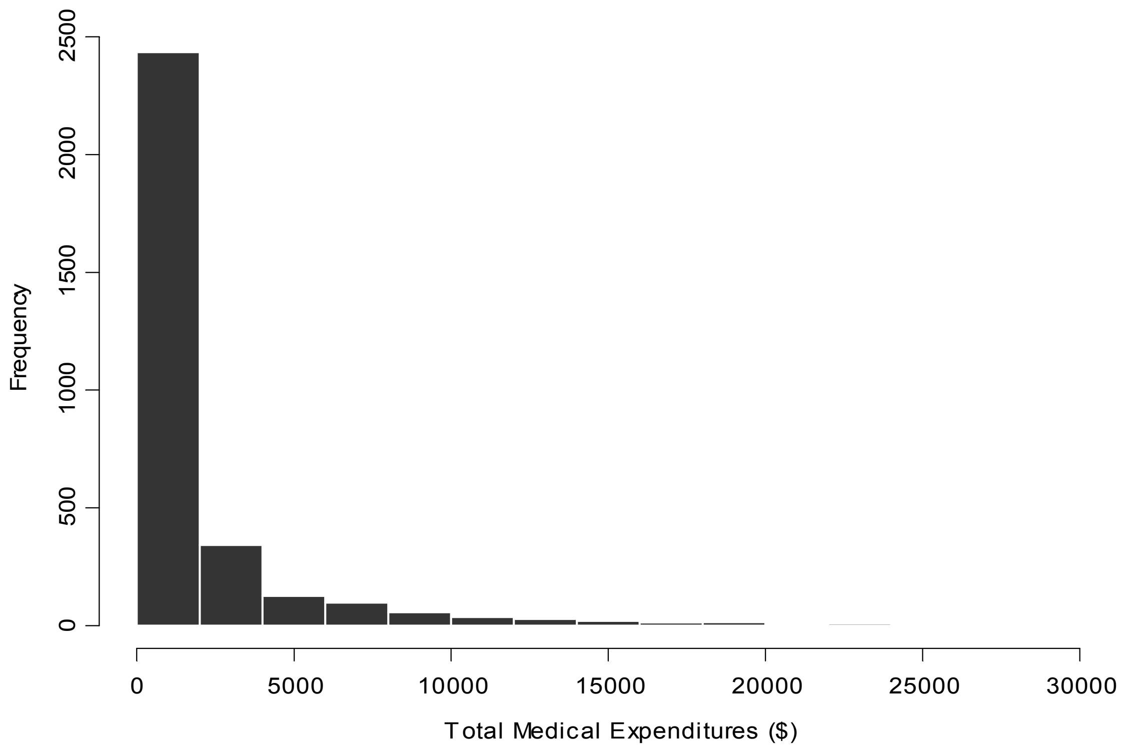


Figure 1. Histogram of Expenditures below $30,000 (MEPS, 1997)

quite helpful for this aspect of the job. Oh, and by the way, the monetary compensation can be quite lucrative! More information about the role of the statistician in the courtroom can be found in texts such as those by Good (2001) and DeGroot, Fienberg, and Kadane (1986).

Now, let's consider an example of a statistical analysis in the litigation setting. Out of respect to my previous employer, and to ensure the confidentiality of his clients, I will not present any details regarding specific cases that I worked on, or the analyses for these cases. However, I will present an analysis of existing data that could be used in a hypothetical lawsuit, and examine a statistical technique appropriate in that context, the two-part regression model.

Data Analysis Example in a Litigation Setting

It is not at all uncommon in litigation for a plaintiff to seek reimbursement for monetary damages due to a defendant's misconduct or negligence. For example, there has been recent legal activity involving a class-action lawsuit against McDonald's, claiming that the restaurant chain is responsible for the obesity and subsequent health problems of the claimants. Though the presiding judge has dismissed this initial lawsuit, the plaintiffs' attorney has indicated that he will be back to try again.

In 1998 forty-six states and the District of Columbia signed the so-called Master Settlement Agreement with five major tobacco companies, ending 4 years of courtroom battles between the states and the tobacco industry. Part of the Settlement provides that states be reimbursed for excess medical expenditures due to smoking. It is thus not too farfetched that sometime in the future a similar suit could be filed against our nation's fast-food giants. It is conceivable that states' attorneys general could try to recover excess medical expenditures attributable to obesity from the fast-food industry.

Obesity is a growing health concern for the United States, with the rate of obesity steadily increasing in the last decade. In 2001 the obesity rate was estimated to be about 21%, with over 60% of the population being either overweight or obese (Mokdad et al., 2001). These figures are likely even higher today. Furthermore, researchers have shown that obesity is a more serious health problem in terms of health status and health care use than smoking and problem alcohol use (Sturm, 2002). In light of this evidence, an obesity lawsuit as suggested above may not be that unrealistic.

In the context of this hypothetical lawsuit against the fast-food industry, we will examine the so-called two-part regression model, which could be employed to assess the fraction of the U.S. population's medical expenditures that is attributable to obesity. The two-part model is a tool that incorporates two commonly employed statistical models: multiple linear regression and logistic regression. Discussions regarding the two-part regression model can be found in the statistical literature, and, since it is frequently used in the context of economics, it is also referenced in econometrics journals. See for example Duan et al. (1983), Melenberg and Van Soest (1996), Mullahy (1998), Miller, Ernst,

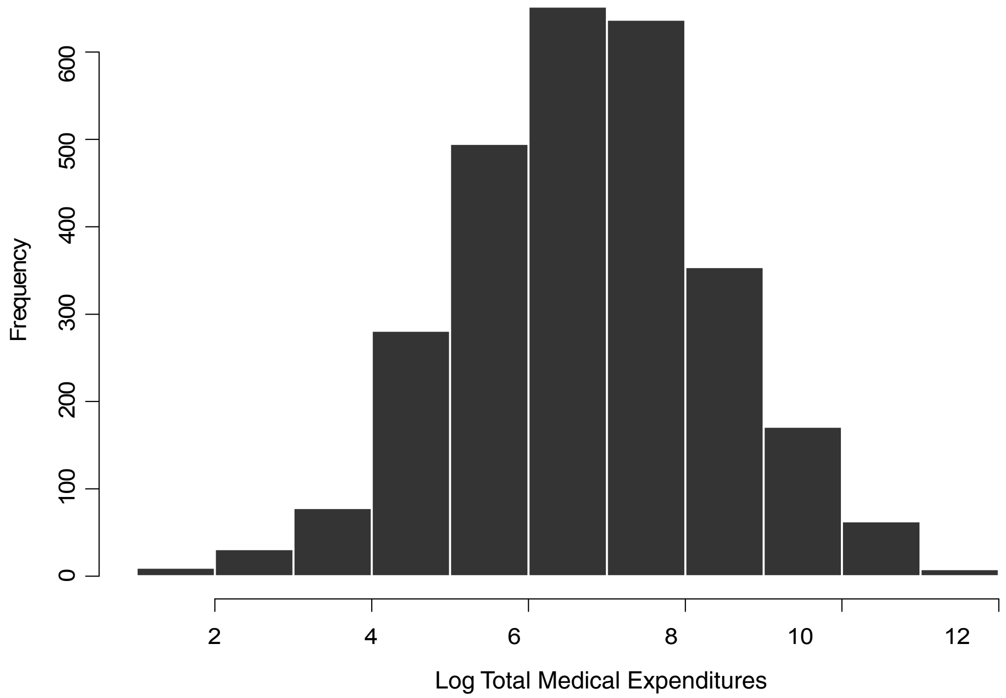


Figure 2. Histogram of Natural Log of Positive Expenditures (MEPS, 1997)

and Collin (1999), Zhou, Stroupe, and Tierney (2001), and Rubin (2001).

Our initial aim is to estimate the relationship between obesity and medical expenditures and then to estimate what proportion of medical expenditures are attributable to obesity. Our first instinct may be to employ a multiple linear regression model in which one of the predictor variables is an indicator variable for obesity:

$$Obese = \begin{cases} 0, & \text{if subject is not obese} \\ 1, & \text{if subject is obese} \end{cases} \qquad (1)$$

To see why this approach is problematic, let us examine a possible data source for this analysis. Combining information from two nationally representative surveys administered by the federal government, the Medical Expenditure Panel Survey (MEPS) and the National Health Interview Survey (NHIS), it is possible to obtain the healthcare related data we need, variables including height, weight, age, gender, smoking status, etc., as well as total annual medical expenditures. More information regarding these two data sources can be found on the Internet (National Center for Health Statistics, 2003; Agency for Healthcare Research and Quality, 2003).

We can combine subjects' health status information from the 1996 NHIS and their medical expenditures from the 1997 MEPS. Figure 1 provides a histogram of the data, restricted to expenditures below $30,000. There are two important features to these data — features that are commonly encountered when dealing with expenditure data. First, as seen in the histogram, the data are highly skewed to the right. When dealing with highly skewed data, statisticians often try to employ transformations to normalize the distribution. A log transformation often works well for this. However, the data set does not lend itself to the log transformation because there are a large number of observations for which the expenditure is $0 (13.5% of the observations). Note however, that if we focus on the observations that are strictly positive, Figure 2 shows that the natural log of the expenditures exhibit a textbook, bell-shaped pattern. So, what we need to do is somehow deal separately with the observations that have $0 expenditures and those that have positive expenditures. That is, we need a more sophisticated approach than just using a multiple linear regression model on its own.

The rationale behind the two-part regression model is to have a logistic regression model predict the binary outcome of whether a respondent has a positive medical expenditure, and then to use a multiple linear regression model to predict the magnitude of positive expenditure for those subjects that do have a positive expense. Since we are now modeling the magnitude of positive (non-zero) medical expenditures, we can use the log transformation to eliminate the skewness of the data.

For illustrative purposes, we will use the combination of MEPS and NHIS to estimate the parameters of a two-part regression model using age, gender, and obesity as predictor variables. Obesity can be defined in terms of body mass index (BMI). This is a measure that takes both height and weight into account:

$$\text{BMI} = \frac{\text{Weight in Kilograms}}{(\text{Height in Meters})^2} \qquad (2)$$

The National Institutes of Health suggests that a body mass index of 30 or more indicates obesity. For each subject in the sample we compute the body mass index and use this to determine the value of the *Obese* indicator variable as defined in (1) above. We do not want to just use BMI as the predictor variable, because our interest lies in the fraction of medical expenditures attributable to the medical condition *obesity*.

In addition to the *Obese* variable, we will also use the explanatory variables *Age*, measured in years, and *Gender*, another indicator variable coded 0 for males and 1 for females. To make the analysis more rigorous, many other explanatory variables should be considered, for instance smoking status, insurance status, and income. But for our purposes we will keep it simple with our three explanatory variables.

Now for the first of the two regression models, letting $p = P(Expend > 0)$, which can be thought of the expected value of an indicator variable for positive expenditure, we will use a logistic regression model to predict the log-odds of having a positive expenditure. The model assumes that the logit of p is linearly related to the explanatory variables:

$$\log\frac{p}{1-p} = \beta_0 + \beta_1 Age + \beta_2 Gender + \beta_3 Obese. \qquad (3)$$

Then, for the second part of the two-part model, we will use multiple linear regression to model the natural log of the positive expenditures. So, if we restrict ourselves to subjects with *Expend* > 0, we can obtain least squares estimates for the regression coefficients in

$$\log \oplus Expend = \beta_0 + \beta_1 Age + \beta_2 Gender + \beta_3 Obese + \varepsilon. \qquad (4)$$

Using MEPS and NHIS data to estimate the regression coefficients in (3) and (4) above, the resulting estimated regression models are:

$$\log\frac{p}{1-p} = 0.06 + 0.03 \cdot Age + 1.04 \cdot Gender - 0.02 \cdot Obese. \qquad (5)$$

$$\log\left[\oplus Expend\right] = 4.65 + 0.04 \cdot Age + 0.31 \cdot Gender + 0.25 \cdot Obese. \qquad (6)$$

In both estimated models *Gender* has a positive regression coefficient, indicating that females are more likely to incur an expenditure than males, and that when they have an expenditure, on average it is higher than for males. The *Age* variable also has positive estimated regression coefficients, indicating that the models predict a greater likelihood of expenditure and a greater magnitude of positive expenditure with increasing age. On the other hand, the logistic regression model estimates that obese individuals have lower log-odds (and thus lower likelihood) of incurring an expenditure, but that once an obese individual does incur an expense, the magnitude of the expense is on average higher than for a non-obese individual. This is reflected in the negative slope for *Obese* in the logistic regression and the positive slope for *Obese* in the multiple linear regression.

Note that our model does not have any interaction terms. This means that we are assuming the effect of age is the same regardless of gender and obesity status. Also, the effect of being obese is assumed to be the same regardless of age and gender. As explained in the previous section, this is a limitation of our model that would need to be addressed if we were interested in using our results in the courtroom.

Equations (5) and (6) can be back-transformed to obtain predicted values of the probability of expenditure and magnitude of positive expenditure:

$$\bar{p} = \frac{1}{1+e^{-(0.06+0.03 \cdot Age+1.04 \cdot Gender-0.02 \cdot Obese)}} \qquad (7)$$

$$\oplus Expend = e^{4.65+0.04 \cdot Age+0.31 \cdot Gender+0.25 \cdot Obese}. \qquad (8)$$

Using the relationship between explanatory variables and medical expenditure estimated from MEPS and NHIS above, we are now in position to estimate the proportion of medical expenditures in the United States that are attributable to obesity. We will use the two-part regression model to obtain predicted expected expenditures "as-is," and in a hypothetical world in which nobody is obese, "as-not-obese." The predicted expected expenditure for a subject is equal to his or her predicted probability of expenditure times the predicted magnitude of his or her positive expenditure:

$$\textit{Predicted Expected Expend} = \bar{p} \cdot (\oplus Expend) \qquad (9)$$

For each subject in the data set we can obtain a predicted expected expenditure "as-is" and "as-not-obese." For example, for a 45 year old, female subject with body mass index 32 (obese), the "as-is" computations are:

$$\bar{p} = \frac{1}{1+e^{-(0.06+0.03 \cdot 45+1.04 \cdot 1-0.02 \cdot 1)}} = 0.9191. \qquad (10)$$

$$\oplus Expend = e^{4.65+0.04 \cdot 45+0.31 \cdot 1+0.25 \cdot 1} = \$1,352.89. \qquad (11)$$

$$\textit{Predicted Expected Expend} = (0.9191)(\$1,352.89) = \$1,243.44. \qquad (12)$$

"As-not-obese," the obesity indicator variable gets turned off (equals 0), and the figures change to:

$$1 \tag{13}$$

$$\text{Expend} = e^{\text{[illegible]}} = \$1,053.63 \tag{14}$$

$$\textit{Predicted Expected Expend} = (0.9206)(\$1,053.63) = \$969.97. \tag{15}$$

We repeat these calculations for all subjects in the data set, getting expected expenditures in the real world and then also in a world without obesity.

Due to the design of the MEPS and NHIS surveys, each subject in the sample actually represents a certain number of people in the population. The number of people in the U.S. population that a subject in the sample represents is given by a subject's *sampling weight*. We can estimate the total medical expenditures for the entire U.S. population by taking the weighted sum of the expected expenditures of all subjects, weighting by the sampling weights. To estimate the fraction of all medical expenditures in the U.S. attributable to obesity, we compute the difference between the sum of the expected expenditures “as-is” and those “as-not-obese.” The obesity attributable fraction is therefore defined as:

$$\frac{\sum(\text{Expected Expend "As - Is"}) - \sum(\text{Expected Expend "As - Not - Obese"})}{\sum(\text{Expected Expend " As - Is"})} \tag{16}$$

where, again, the summations above are actually weighted sums.

Computing the necessary sums in (16), we find that our population's expected medical expenditures “as-is” are estimated at \$16.3 billion, and “as-not-obese” at \$14.5 billion, a difference of \$1.8 billion, or 5.51%. That is, the obesity attributable fraction of medical expenditures is estimated to be 5.51%. Even though this is a rather crude estimate since we didn't account for other reasonable explanatory variables in our regression models and did not include interaction terms, our estimated fraction is close to what other researchers have published. Finkelstein et al. (2003), for example, computed an attributable fraction of 5.3%—and they used an even fancier four-part model.

Of course, we would want to have a confidence interval for the attributable fraction as well. This is no simple task, since the sampling distribution of the attributable fraction is not of a simple nature. Though we will not pursue it here, it is possible to use a resampling method such as the jackknife or bootstrap to obtain the confidence interval.

Returning to the context of our hypothetical litigation against the fast-food industry, should these companies be responsible for reimbursing the government for 5.3% of all medical expenditures? The plaintiff's attorney may certainly argue so. However, there are some obvious aspects of the analysis that could be attacked by the defense. The most apparent flaw is that we have not accounted for possible confounding factors. We are using observational data, not data obtained from an experimental design. There may be other differences between obese and non-obese subjects that could explain the discrepancy in medical expenditures. For instance, if obese individuals are also more likely to be smokers, and if smoking also leads to higher medical expenditures, smoking status could be an important confounding variable. In this case, the 5.3% would be an overestimate of the true obesity attributable fraction. Including smoking as an explanatory variable in the model would help, but we could never include all possible confounding variables.

Another flaw in the data is that the NHIS obtains height and weight measurements using self-response – that is, subjects report their own height and weight. Thus, the prevalence of obesity estimated from the NHIS data is likely an underestimate, and the 5.3% is likely an underestimate of the obesity attributable fraction.

Even if obesity is responsible for 5.3% of healthcare expenditures, how much of this 5.3% is caused by fast-food? Attorneys for the fast-food industry would argue that the obesity epidemic in the United States is not caused by our nation's addiction to French fries and Big Mac's, but rather by our own sedentary, couch potato, lifestyle. Further statistical analyses would be necessary to quantify the percentage of medical expenditures caused directly by the fast-food industry.

Above are some of the statistical arguments that the attorneys would wrestle with. There would also be non-statistical legal issues to deal with: personal responsibility; whether the fast-food industry adequately publicizes the nutritional content of its food; whether the industry irresponsibly advertises to children, to name a few. As you can see, there would be many facets to this litigation, but one thing is certain: the attorneys would definitely need the help of a competent statistician. Perhaps that could be you!

References

Agency for Healthcare Research and Quality (2003), “Medical Expenditure Panel Survey Homepage,” on the web at www.meps.ahrq.gov.

DeGroot, M. H., Fienberg, S. E., and Kadane, J. B. (eds.) (1986), *Statistics and the Law*, New York: John Wiley and Sons.

Duan, N., Manning, W. G., Morris, C. N., and Newhouse, J. P. (1983), “A Comparison of Alternative Models for the Demand of Medical Care,” *Journal of Business and Economic Statistics*, 1, 115–126.

Finkelstein, E. A., Fiebelkorn, I. C., and Wang, G. (2003), “National Medical Spending Attributable to Overweight and Obesity: How Much, and Who's Paying,” *Health Affairs*, 22(4), 219–226.

Good, P. I. (2001), *Applying Statistics in the Courtroom: A New Approach for Attorneys and Expert Witnesses*, New York: Chapman and Hall/CRC.

Melenberg, B. and Van Soest, A. (1996), "Parametric and Semi-parametric Modeling of Vacation Expenditures," *Journal of Applied Econometrics*, 11, 59–76.

Miller, V. P., Ernst, C., and Collin, F. (1999), "Smoking Attributable Medical Care Costs in the USA," *Social Science and Medicine*, 48, 375–391.

Mokdad, A. H., Ford, E. S., Bowman, B. A., Dietz, W. H., Vinicor, F., Bales, V. S., and Marks, J. S. (2001), "Prevalence of Obesity, Diabetes, and Obesity-Related Health Risk Factors," *Journal of the American Medical Association*, 289, 76–79.

Mullahy, J. (1998), "Much Ado About Two: Reconsidering Retransformation and the Two-part Model in Health Econometrics," *Journal of Health Economics*, 17, 247–281.

National Center for Health Statistics (2003), "National Health Interview Survey," on the web at www.cdc.gov/nchs/nhis.htm.

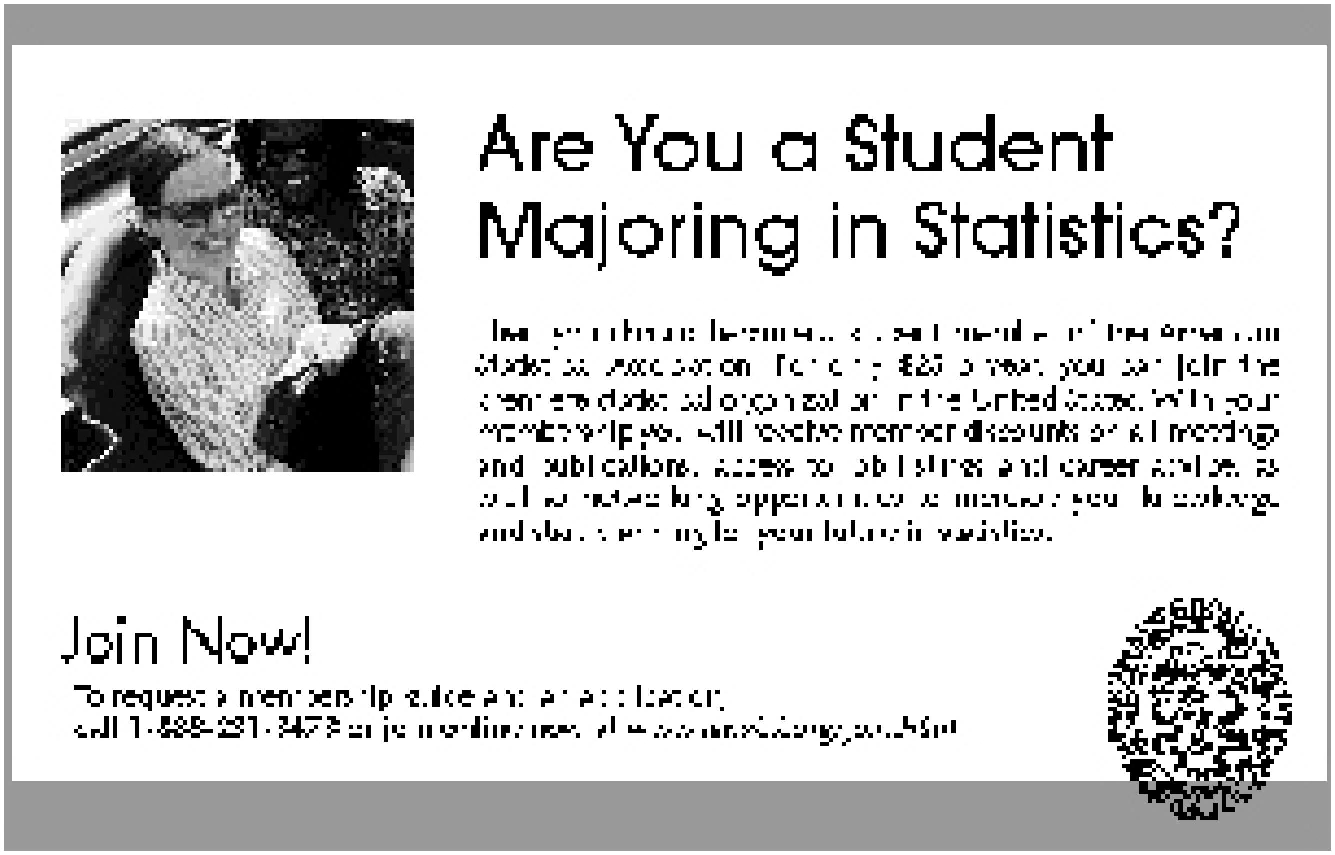
Rubin, D. B. (2001), "Estimating the Causal Effects of Smoking," *Statistics in Medicine*, 20, 1395–1414.

Sturm, R. (2002), "The Effects of Obesity, Smoking, and Drinking on Medical Problems and Costs," *Health Affairs*, 21(2), 245–253.

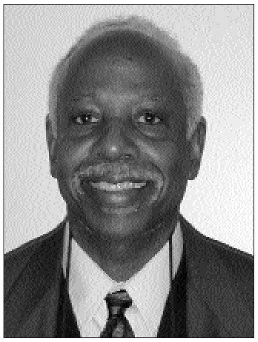
Zhou, X., Stroupe, K. T., and Tierney, W. M. (2001), "Regression Analysis of Health Care Charges with Heteroscedasticity," *Applied Statistics*, 50(3), 303–312.

Web Resources

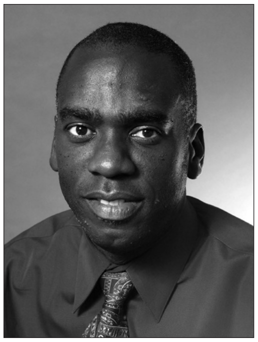
- Agency for Healthcare Research and Quality *www.meps.ahrq.gov*
- National Center for Health Statistics *www.cdc.gov/nchs/nhis.htm*



A Rainy Day in Trinidad



Winston A. Richards



Jeffrey R. Wilson

Introduction

For the past several years at the Pennsylvania State University, Richards and hydrologist Scott Huebner have taught a course entitled *Environmental Issues in Developing Countries*. For the field experience related to this course, students have taken trips during the spring break to Barbados, Jamaica or Trinidad, focusing on issues related to water and sewerage treatment facilities as well as beach erosion and water quality. The trip is usually designed so that the first stop on the island is the Central Statistical Office, where students examine environmental data and look for related problems. A key objective of the course is to have students witness first-hand how the collection and modeling of real data are formulated for statistical analyses to assist the local government.

On one of these trips during the month of February 1993, we noted among the data reported by the Central Statistical Office of Trinidad and Tobago a series of monthly rainfall amounts that encompassed the 115 years from 1863–1977. At that moment Richards was reminded of a tourist whom he had met on an earlier (November 1991) visit to Trinidad. This tourist reported that on his first two-week visit to Trinidad it had rained repeatedly. He was so disappointed that he did not return to Trinidad and Tobago for the next ten years. After students heard this story, they agreed to take a closer look at the data in order to investigate the statement made by the tourist. This article describes how the students used the rainfall data to apply some of their new knowledge about probability distributions and modeling. Does it really rain a lot in November in Trinidad? How does this compare to the rainfall amounts in February?

Data Collection

It was important that the students obtain knowledge and historical information on the collection of the data, so the students were taken to the Water and Sewerage Authority (WASA) of Trinidad and Tobago where they obtained the following information from an officer who was familiar with the data. The students were informed that a Casella Pot gauge (an instrument for recording rainfall amounts) was placed on the Botanical Gardens premises, which is located in the north of the island in close proximity to the Trinidad and Tobago zoo. Because of the close proximity of the gauge to the nearby zoo, zookeepers were paid by the government to collect these readings. The officer felt that the dedication and civic mindedness of gauge readers were more responsible for the maintenance of the series of readings than the monetary compensation provided by the government. Nevertheless, there were some difficulties reported with the data collection: meters were not always read on the last day of the month, so some "months" included data for more than 31 days, and evaporation could not be measured because there was no evaporation station in the vicinity of the Casella Pot gauge.

The students were informed that a Lambrecht automatic recorder (an instrument weighing about 15.2 kg for measuring humidity, temperature, atmospheric pressure, precipitation, wind and other climatic values) still existed at the WASA headquarters and could be used to correct for the "long months" through a ratio method. We were informed that such a procedure would considerably minimize the errors due to "overlap" in the series of readings. The effects of such a procedure are related to the distance between Lambrecht and Casella Pot gauges. Moreover, the Botanical Gardens gauge was inland and somewhat shaded, while the WASA gauge was close to the sea. Thus it was obvious that the 115 years of rainfall data had a certain

Winston A. Richards (ugu@psu.edu) is an Associate Professor of Mathematics & Statistics at Penn State Harrisburg and a summer visiting Professor at Stanford University's Statistics Department. He is a Fellow of the American Statistical Association and has been President of the Harrisburg Chapter of the ASA.His research area is Exact Distribution theory and he has an interest in Environmental issues In Developing countries. He received his BS and MS in Mathematics from Marquette University and his M.A. and Ph.D. from The University of Western Ontario.

Jeffrey R. Wilson (jeffrey.wilson@asu.edu) is a biostatistician and the Director of the School of Health Administration and Policy in the W. P. Carey School of Business at Arizona State University. He received his B.A. in Mathematics from the University of the West Indies, and a Master's and Ph.D. in Statistics from Iowa State University. His major research interest is generalized linear models, categorical data, sampling issues and analysis of overdispersed data. He has received the Golden Key award for teaching and is a recipient of grants from NSF, NIH and

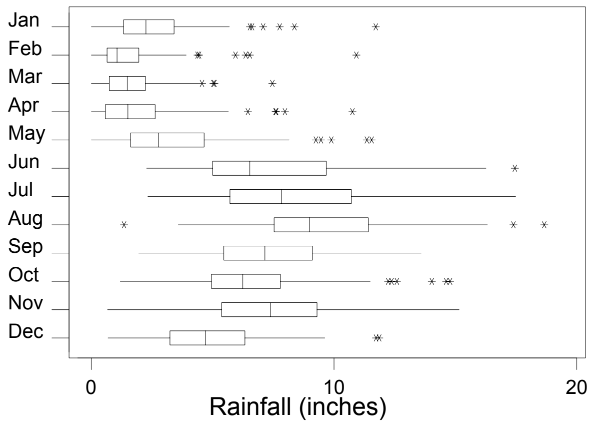


Figure 1: Boxplots for Rainfall for Jan–Dec 1863–1977

degree of measurement error. The meeting with the officer proved fruitful for the students, as they were made aware of some of the difficulties associated with the collection of data and thus the need to make the necessary adjustments in modeling.

Why Model?

There are several reasons to identify a probability model for rainfall in Trinidad and Tobago. We sought to generate a model based on the data collected through 1977. Such a model can be used to predict rainfall events after 1977, assuming that climatic conditions do not change.

For the students this is an exercise in applications, but for some Trinidadians there may be much more involved. For example, farmers in Trinidad must rely on a certain amount of rainy and sunny days for their profits to be fully realized. They need to adapt to the environment if they are going to get optimum yield from their crops. A model of the rainfall distribution for that country would therefore be extremely useful to the farmer. Over the years, certain anomalies in climatic conditions have been noted, often referred to as the Greenhouse Effect. These anomalies have included excessive rainfall and recurring drought. Modeling the rainfall data might help to identify outliers associated with rainfall and enable the meteorologist to monitor the trend.

Discovering Seasonality Through Exploring Data

The first step in identifying a probability model is to explore the sample data through graphical and numerical procedures. Students generated boxplots of rainfall amounts per month over the period 1863-1977 (Figure 1). From the graphs it appeared to the students that the rainy season extended from approximately May to December, while the dry season spanned December to May. Further, they noted from the boxplots that the monthly rainfall amounts for the first ten months of the year were skewed to the right.

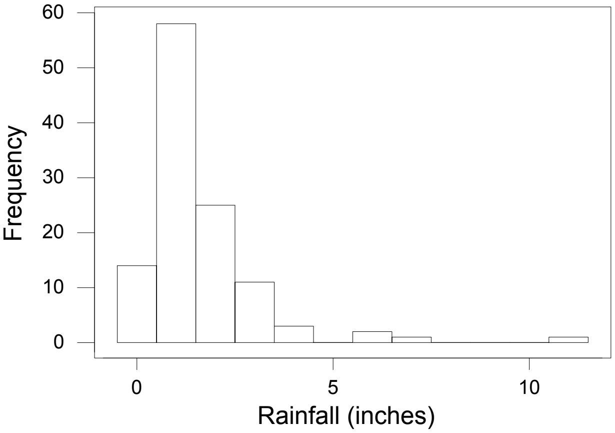


Figure 2: Histogram of February Rainfall for Trinidad (1863–1977)

Next, from observing separate histograms for the rainfall for all February months (the month that we were there, Figure 2) and for all November months (the month the tourist was there, Figure 3), the students gained some degree of belief that November's distribution was roughly symmetric and February's rainfall was skewed to the right.

Determining the Theoretical Models

As the students understood the history and relevance of the data, they were prepared to state and investigate their questions of interest: "What was the nature of the underlying probability distributions that were responsible for the pattern of rainfall for February and November over the 115 years?" We suggested that they try one of two probability distributions: the gamma probability distribution and the normal probability distribution. Approximately half of the students tried the gamma and the other half tried the normal. Below we present the results comparing the February rainfall to a gamma distribution and the November rainfall to a normal distribution.

To estimate the parameters for these probability

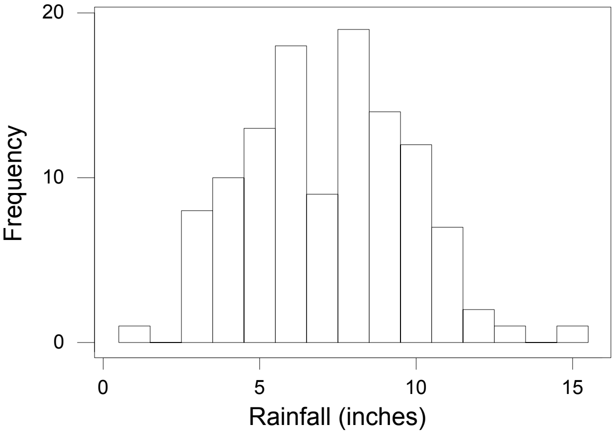


Figure 3: Histogram of November Rainfall for Trinidad (1863–1977)

models, students were instructed to use one of two methods: maximum likelihood or the method of moments. The consistency, unbiasedness, and the asymptotic properties of maximum likelihood estimates were made known to the students (Rao, 1973). The ease of computing method of moments estimators was also pointed out.

For the gamma distribution the shape parameter α and the scale parameter β have method of moments estimators:

$$\bar{\alpha}_m = n\bar{x}^2 / (n-1)s^2;$$

and

$$\bar{\beta}_m = \sum (x_i - \bar{x})^2 / n\bar{x} ,$$

respectively, where $\bar{x}$ denotes the sample mean, s^2 denotes the sample variance, and x_i denotes the i^{th} observation. The maximum likelihood estimators for α and β are obtained from the solution of the following equations:

$$\left[\sum \log(x_i) / n \right] - \log(\bar{x}) = \Psi(\alpha) - \log(\alpha)$$

and

$$\beta = \bar{x} / \alpha ,$$

where $\Psi(\cdot)$ is the digamma function, (Ross, 1981). The students obtained the solutions to these equations iteratively through the use of the software package Derive (Kutzler, 1996).

Analysis of the fit of these probability models (gamma and normal) using the maximum likelihood estimates was conducted using the Pearson χ^2 goodness-of-fit test and the Anderson-Darling statistic A^2 (D'Agostino and Stephens, 1986). The Anderson-Darling statistic A^2 is based on the empirical cumulative distribution function (ECDF) and is a measure of the vertical difference between the empirical distribution function and the hypothesized distribution function. This statistic measures how much the set of data points under consideration deviates from the hypothesized distribution. This Anderson-Darling procedure makes use of all of the information contained in the sample and therefore is more powerful than the χ^2 goodness-of-fit test, which groups the data.

February Rainfall

When the data for 115 February months were examined, Minitab produced the class intervals and frequencies shown in Table 1. Having months with zero rainfall is problematic for the gamma distribution, so we replaced all three data values of 0 with .005, the maximum value possible since rainfall amounts were measured in increments of .01 inches. The maximum likelihood estimates were $\hat{\alpha}$= 1.26 and $\hat{\beta}$= 1.21. Using these values as the parameters of the gamma distribution, we obtained the expected frequencies in each interval. The resulting table contains cell frequencies less than five, so one has to be careful applying these chi-square tests for the fit. To address this, we combined the last five cells in the table, leaving four cells. The computed chi-square statistic (with one degree of freedom) was 1.405 (p-value = 0.236), failing to reject the gamma distribution hypothesis at the 5% level of significance.

Table 1. Observed and Expected Gamma Frequencies for Rainfall Data in February Months (1863–1977)

Class Intervals	Frequencies	Probability	Expected frequencies
[0.0.0.75)	37	0.3346	39.63
[0.75,2.25)	56	0.4308	49.54
[2.25,3.75)	15	0.1528	17.57
[3.75,5.25)	3	0.0495	5.69
[5.25,6.75)	3	0.0155	1.78
[6.75,8.25)	0	0.0048	0.55
[8.25,9.75)	0	0.0015	0.17
[9.75,11.25)	1	0.0004	0.05

However, using different class intervals (Table 2), the results change slightly. The computed chi-square statistic with 1 df was .529, p-value = 0.467. These findings highlight an important problem associated with the application of the chi-squared goodness of fit test: when data are grouped, information is lost. The method of grouping also determines how much information is lost.

The Anderson-Darling statistic for comparing these data to a gamma distribution can be computed as A^2 = 1.476. However, it is not straightforward to obtain a p-value in this case. On the other hand, we can use Minitab's built-in test for normality, which reports the Anderson-Darling statistic as A^2 = 7.562 and the p-value as .000. See the normal probability plot in Figure 4. Thus, we have strong evidence that a normal distribution is not an appropriate model for the February rainfall data.

November Rainfall

When comparing to a normal distribution, we again use Minitab's built-in test and compare this to the

Table 2. Observed and Expected Gamma Frequencies for Rainfall Data in February Months (1863–1977)

Class Intervals	Frequencies	Probability	Expected frequencies
[0.00,1.50)	72	0.6107	70.24
[1.50,3.00)	31	0.2616	30.09
[3.00,4.50)	8	0.0875	10.06
[4.50,6.00)	1	0.0278	3.19
[6.00,7.50)	2	0.0086	0.99
[7.50,9.00)	0	0.0026	0.30
[9.00,10.50)	0	0.0008	0.09
[10.5,12.00)	1	0.002	0.03

Table 3. Observed and Expected Normal Frequencies for Rainfall Data in November Months (1863–1977)

Class Intervals	Frequencies	Probability	Expected frequencies
[0.0,0.26)	4	0.0391	4.50
[0.26,5.20)	23	0.1831	21.06
[5.20,7.80)	38	0.3694	42.48
[7.80,10.40)	39	0.2987	34.35
[10.4,13.00)	9	0.0967	11.12
[13.00,15.60)	2	0.0124	1.43

results from Pearson's chi-square statistic. We used Minitab's default intervals (Table 3), but then we grouped together the last two cells. The maximum likelihood estimates were

$$\hat{\mu} = 7.196, \hat{\sigma} = 2.621.$$

The Anderson-Darling statistic was $A^2 = 0.368$ (p-value = 0.425) and the chi-square statistic was 1.471 (p-value = .479 with 2 degrees of freedom). See the normal probability plot in Figure 5. Thus, this normal distribution was not rejected at the 5% level of significance, and the students established that the rainfall amounts from 115 November months are reasonably approximated by a normal probability distribution.

Extension

Considering the data for 115 Novembers and assuming a normal population, then the estimated mean rainfall was 7.196" and its estimated standard deviation was 2.621". The ninety-fifth percentile based on the normal distribution is found to be 11.51". Thus the probability that two or more rainfalls in November over the next 16 years (1978–1993) would exceed 11.51" of rain can be computed from the binomial distribution to yield a value of 0.1892.

Examining the actual records for rainfall in Trinidad from 1978–1993 found the following November rainfall amounts: 1.49", 8.47", **13.16**", 7.39", 5.83", 3.38", 8.58", 7.88", 8.78", 3.806", 8.826", 9.29", 5.39", **12.18**", 7.67", 11.14" respectively. The rainfall exceeded 11.51" in 1980 and 1991 (as shown in bold). The records showed that the rainfall in 1991 when the tourist was present was 12.18" of rainfall, which was larger than the 95th percentile for November months based on a series of data that ended in 1977.

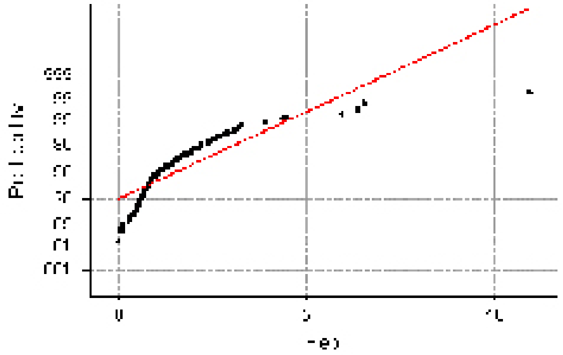


Figure 4: Normal Probability Plot for February Rainfall

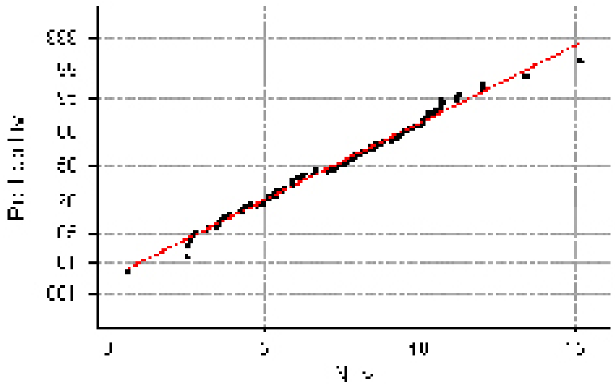


Figure 5: Normal Probability Plot for November

Assuming that the order of occurrence of the annual November rainfall was random, the students concluded that the number of November months until one had a rainfall that exceeded 11.51" (95th percentile) would have a geometric probability distribution. Thus one would have to wait on average 1/0.05 or 20 years with a standard deviation of 19.49 years to see a rainfall beyond 11.51". From the records in the 16 years that followed from 1977, the wait was only 3 years and then 11 years for excessive rainfall.

Conclusion

Trinidad is a small island of about 1864 square miles located close to the equator. It has two seasons: wet and dry. The tourist was clearly present during the rainy season. However, what he experienced in 1991 does not provide him necessarily with enough evidence to say that it rains a lot in Trinidad. Maybe the 12.18 inches of rainfall in that November month was normal based on a series starting in 1970 and expanding to 1999. We do know that in November 1991 the rainfall is somewhat unexpected based on a series of 115 years and an extrapolation of 14 years.

References

Agresti, A. (1980), *Categorical Data Analysis*. New York: John Wiley & Sons.

D'Agostino, R.B., and Stephens, M.A. (1986), *Goodness-of-fit techniques*, New York, NY: Marcel Dekker, Inc.

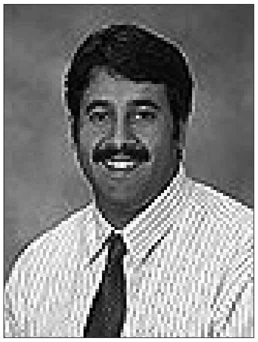
Kutzler, B. (1996), *Introduction to Derive for Windows*, Honolulu, Hawaii, Edward Enterprises.

Minitab, Inc. (1993), *Minitab for Windows*, Release 9, State College, PA.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications*, *2nd ed*. New York: John Wiley & Sons.

Ross, S. M. (1981), *Introduction to Probability Models*,

Preparing for Life After Graduate School:



Sumithra J. Mandrekar **Jayawant N. Mandrekar**

Overview

A question lingers on every student's mind towards the end of their graduate school life — What next? As recent graduates, neither of us was an exception to this saga. Based on our personal experience, we aim to provide some valuable job-hunting hints for fresh biostatistics graduates and the dos and don'ts of the process. Throughout this article, we also provide references to the websites and the books that helped us in our job search and that, we believe, can serve as guiding tools for some readers as well.

Introduction: Who Are We?

Young upcoming biostatisticians with a hope for a successful future…

It was the summer of 1997 when both of us joined the Ohio State University (OSU) with dreams and aspirations of becoming successful biostatisticians (a dream that many of us carefully nurture and care for over several years). In this sense, we were no different than any of our fellow wide-eyed students who joined with us that year. We all began the long and arduous journey of a typical graduate school life, with its share of sunny and cloudy days. Some of us continued until the end, whereas others decided that it was time to leave school and enjoy the remainder of their lives without stress and with a few more dollars in their wallets!

Now, without going into the usual well-known "not-so-memorable" details of the course load, exams, teaching and research that span a few years, we will jump straight into the last few months before graduation when the tension is at its peak. But before that, a few facts about us: both of us graduated with Ph.D.'s in the same year (2002) within a three-month interval (as OSU follows a quarter system), were (and are still) happily married, and were international students.

Selection or Elimination?

Looking for a job is a selection process for both the employer and the candidate, but it is also an elimination process for the two sides — so beware!

About a year prior to "tentative" graduation time, we first short-listed our career interests. The ASA's Career Center webpage has informative articles and valuable information on different career options available in the statistics and biostatistics fields. This is the first and the most critical step whose importance cannot be over-emphasized. There is nothing worse than an aimless job search. Having said that, it is always possible that one might have missed out on some avenues of opportunity and might happen to stumble upon others while looking for the ones in mind. In our case, we both decided that we wanted to work for a research institution with lots of opportunities to consult and collaborate with physicians, with some room for personal research. Again, in narrowing down to this decision, we also had to check if research institutions accepted international students, although this may not be a constraint to several of you reading this article. The organization must be willing to process work permits and other visa-related documents for international students. This issue, if applicable, can become the determining factor of your job search.

As the next step, we sifted through the list of all available career opportunities that fit our interests and noted their corresponding deadlines for application and selection. This is not to say that one should not apply to those jobs where your available dates may be beyond the required starting dates for the job, because in many instances the organization is willing to wait for a candi-

Sumithra J. Mandrekar is Research Associate (Lead Statistician), Cancer Center Statistics, Mayo Clinic, Rochester, MN 55905 (email: mandrekar.sumithra@ mayo.edu). She has an interdisciplinary Ph.D. in the areas of Statistics, Psychology and Internal Medicine. She has an interest in statistical modeling and inference and at Mayo Clinic, she works on the design and analysis of cancer clinical trials. Jayawant N. Mandrekar is Research Associate (Lead Statistician), Section of Biostatistics, Mayo Clinic, Rochester, MN 55905 (email: mandrekar.jay@mayo.edu). His statistical areas of interest include survival and categorical data analysis. At Mayo Clinic, he offers statistical support to the divisions of radiology, general internal medicine and infectious diseases among others. He is also the statistician and voting member of the General Clinical Research Center (GCRC).

date if they feel that his/her profile is a perfect match for them. A useful source of the statistics job related advertisements is *Amstat News*, the membership magazine of the American Statistical Association. A special careers issue of *Amstat News* in September 2000 gave an overall picture, making it a good starting point to learn about about all issues regarding careers in statistics. Some websites that have information on up-to-date biostatistics job opportunities include include ASA JobWeb (academic and industry), University of Florida's list of job announcements (academic and industry), and *monster.com* (industry). The Statistics Department websites of several universities have postings of the current openings in their university as well as useful links and pointers to other websites and postings.

The Joint Statistical Meetings (JSM) held every year in August are yet another very good resource. Several employers from academia and industry screen potential candidates at these meetings. It may be useful to attend these meetings even a year before you actually start your job search as it gives a good overall picture of the kind of jobs that are available and some of the skills and experience that each organization looks for in a candidate. The annual meetings of the American Public Health Association (APHA) and the International Biometrics Society Eastern North American Region (ENAR) are also useful resources, but to a lesser extent than the JSM. More information on these conferences can be obtained at the ASA, APHA, and ENAR websites.

One more possible obstacle in our case — are there two openings in the same organization or are there similar opportunities in the nearby vicinity? The first issue is difficult to resolve when an organization has only one opening at a given time, and a couple, both with the same educational background, applies at the same time. In such a case, it may be beneficial to discuss with the employer whether there would be any future openings when they can consider the other person. Sometimes the rules of an organization might prohibit couples from working in the same division or even the same organization. In such situations, it is useful if there are other opportunities in the same town or city. Many times, this can become a constraining factor, but in our case, since we already had several other restrictions (listed in the preceding paragraphs) we decided not to insist on this ideal situation. So, we both had the entire repertoire of jobs in front of us — we were two independent random variables when it came to the job search!

Several job listings matched our interests and so began the next phase of our journey: the formal application process.

The Application Packet

Good reference letters are powerful “selling” agents!

Before putting together an application, it is important to have a strong idea of the expected graduation date, and three to five reliable references (make sure that both you and your reference are in an acceptable comfort zone regarding your application). Often times, a candidate considers job positions that have slightly different emphases, and your references should be willing to highlight your skills and work ethics pertinent to that specific job profile, at the same time keeping in mind your career goals and professional interests. The other key components of a typical application include a cover letter, a statement of purpose (carefully thought out and tailor made for each application), a professional resume, and a current up-to-date transcript. Some useful tips on preparing cover letters can be found in Hansen and Hansen (2001) and Toropov (1998). The Toropov references are readily available in most university and public libraries. It is also not expensive to own one. A new copy costs around $20, and used ones can be purchased from between $5 to $15 from websites like Amazon and Career/LifeSkills Resources Inc. Today there is a lot more emphasis on interpersonal and communication skills than a decade ago, but these books cover many basic aspects and give useful tips that are still very relevant and important.

The importance of the statement of purpose (or SOP, as it is popularly known) cannot be over-emphasized. It is “the” tool that enables an employer to judge your focus and interest in their job. A well presented logical SOP, together with a detailed but not painfully elaborate resume, are the two weapons a candidate possesses to successfully launch an application. It may be beneficial to have some friends or faculty review your resume and SOP, as they may be able to provide unbiased opinions on not only the content but also the order and style of the presentation. A good source of “statistical” resume templates and references to books that give guidelines on preparing resumes, cover letters and SOP is O'Brien (2000). Other sources include Eischen and Eischen (2000) and the *monster.com* website, both of which have general guidelines for building a resume, not limited to the statistics field.

Many people (particularly past students) had advised us that the single most important item in an application packet is the reference letter. Let us think about this a little further: why do you choose someone as your reference? Probably the most obvious reason is because you have worked with them a lot and most likely have enjoyed working with that person. Now, it makes perfect sense to put a lot of weight on a reference letter as it sheds light on your working abilities, efficiency, and all that good stuff that an employer is typically looking for in a potential candidate.

With all this in place, the application packets are ready to be fired out the door to the different places. You can also post your resumes online at some websites like the ASA JobWeb (academic and industry) and *monster.com* (industry). Several pharmaceutical companies and clinical and/or contract research organizations also have the provision of submitting resumes online to their job advertisements and/or posting them on their website.

Who Is Interested In Us?

Suddenly, the days seem longer and the anxiety keeps mounting…

A typical selection process at an organization begins as soon as they receive a few applications. They do not necessarily wait until after the deadline to sort and screen the likely candidates. We do not know the route of the applications in all the organizations, but can speak about the few where we applied.

Typically, a receipt notification was sent to us as soon as the organization received our application. With that, began the waiting game. This is the time when many intelligent minds get an opportunity to chew and digest our SOP, resume, and the reference letters, and sometimes it may lead to fiery roundtable discussions and extended committee meetings, discussions that make them decide if their organization would benefit from our qualifications. Sometimes several weeks passed with no communication. Finally (much to our relief) some places responded with the good news (which is obviously an interview call) and others with the bad news (where we were graciously rejected). If there are positive responses from many places, then careful scheduling is required by the candidate so that everyone is kept happy.

At this juncture, there is always the dilemma of whether to keep all of the potential employers on the same page with full information about other jobs that you have applied for. From our experience, we think that it is a good idea if everyone is updated about everyone else, because in some sense knowing that you are in “demand” increases your “market” value!

You don't always hear from those from whom you most “want” to hear, but nevertheless, you move on to the next step — interviews.

The D-Day

Now that you have come to the bridge, cross it with confidence and caution…

Preparing for an interview is a long, drawn out process. It is not sufficient to sharpen just your academic knowledge: you have to be an allrounder. Besides being purely technical, many interviews also focus on behavioral aspects and communication skills. Therefore, it is beneficial to prepare some examples and illustrations in advance that can highlight your interpersonal skills and work relations. Although many may argue that appearances are deceptive, we think that it is imperative to be well dressed as it creates an impression of well preparedness and at the same time enhances your self-confidence. The saying, first impression is the best impression, fits well in this situation. Some useful references on preparing for interviews are Eischen and Eishen (2000), Toropov (1996), and Kennedy (1996).

When presentations are a part of the interview, they often become the focal point of the decision-making process, and hence dedicated preparation is essential. Many times a candidate is judged not by the sheer “quantity” and “technicality” of the presentation content, but also the quality, which includes the style, ease, timeliness and accessibility of the material presented. It is important to capture the attention of the audience and keep them interested for most of the presentation. Efficient handling of audience questions is vital as it re-emphasizes your strong hold on the presentation topic and also highlights your listening capacities. A well-answered question deserves credit, but a badly answered question has the ability to supercede that and wipe out the previously earned credit. So, when in doubt or in a “no clue” situation, it is never wrong to say, “I don't know” or “I need to think on this a little more,” rather than attempting to answer. On the same note, remember that the questioner may not always have the answer!

Being attentive and alert and at the same time calm and composed are essential during the one-on-one or panel interview sessions (Krannich and Krannich, 1999). It is always desirable to have sufficient background about the organization and its expectations for the new person beyond what is stated in the job advertisement. Often times towards the end of an interview, an opportunity is given to the candidate to ask questions during which time the homework comes in handy. Further, if possible, always read about the technical background of the interviewers (if known *a priori*) as that gives a fairly good idea about the work done at the organization, as well as the opportunity to ask questions intelligently.

At the end of the day, an “exit” interview summarizes all the information and the logistics for you, and then it is back to the waiting game.

Final Countdown

Life is full of choices; be sure to weigh all your options before making the final decision…

The first thing to do after getting back from an interview is to send a personal thank you note to each and everyone with whom you interviewed, went out for lunch or dinner, or who helped organize the interview trip and schedule for you. Krannich and Krannich (1997) have a good collection of sample follow-up letters.

Typically, most potential employers specify a time line within which they expect to have interviewed and made a decision about all the possible candidates. However, it is not uncommon for a candidate to contact the employer in case of other pending offers on which a decision has to be reached. In any case, you should never feel pressured or be forced to make a decision.

Once you have received an offer, it is helpful to discuss your plans with the people who sent you the acceptance letters and make them aware of your expectations (in case you are waiting to hear from some place else). At the same time, do not put all your eggs in one basket. Keep your options open until you have all the outcome information from all the places you interviewed with. The final decision should make you feel comfortable and satisfied and if you are lucky, it can very well be your "dream" job. Some tips on negotiating an offer are outlined in Simon (1998).

Coming to a decision on which offer to accept is a balancing act. Job satisfaction, good compensation, potential for career growth, and geographical location are some issues that can drive the decision-making process. It may be helpful to get opinions from faculty and recent graduates on these issues. One thing worth mentioning at this juncture is that between the two of us, we had worked as summer interns at several places, which included university settings, pharmaceutical companies and research organizations. One of us also worked as a summer intern at Mayo Clinic, and so we had an opportunity to get a glimpse of the work environment *a priori*. From our personal experience, therefore, we think that it is beneficial for graduate students to do at least a couple of internships or more in different types of organizations to get a flavor for the real work environment and job profile.

At Mayo Clinic, there is a healthy mixture of consulting and collaborative clinical research, teaching and personal research components. It would therefore be appropriate for us to say that all of this meticulous planning and preparation paid off, and we did end up with our "dream jobs"!

References

Eischen, C. W. and Eischen, L. A. (2000), *Resumes, Cover Letters & Interviewing: Setting the Stage for Success*, Cincinnati, Ohio: South-Western College Publishers.

Hansen, K. and Hansen, R. S. (2001), *Dynamic Cover Letters: How to Write the Letter that Gets You the Job*, Berkeley, California: Ten Speed Press.

Kennedy, J. L. (1996), *Job Interviews for Dummies*, Foster City, California: IDG Books Worldwide.

Krannich, R. L. and Krannich, C. R. (1997), *201 Dynamite Job Search Letters*, Manassas Park, Virginia: Impact Publications.

Krannich, C. R. and Krannich, R. L. (1999), *101 Dynamite Answers to Interview Questions: Sell Your Strengths!* Manassas Park, Virginia: Impact Publications.

O'Brien, R. G. (2000), "Applying for a Job: Your Curriculum Vitae and Cover Letter," *Amstat News,* 279:15–20.

Simon, M. (1998), *Negotiate Your Job Offer: A Step-By-Step Guide to a Win-Win Situation,* New York: John Wiley and Sons.

Toropov, B. (1996), *Last Minute Interview Tips*, Franklin Lakes, New Jersey: Career Press.

Toropov, B. (1998), *Last Minute Cover Letters*, Franklin Lakes, New Jersey: Career Press.

Web Resources

Amazon Bookstore: *www.amazon.com*

American Public Health Association: *www.apha.org*

American Statistical Association (ASA): *www.amstat.org*

ASA Career Center: *www.amstat.org/careers/*

ASA JobWeb: *jobs.amstat.org*

ASA Meetings: *www.amstat.org/meetings*

Career/LifeSkills Resources: *www.career-lifeskills.com*

International Biometric Society Eastern North America Region: *www.enar.org*

Monster Job Search Services: *www.monster.com*

University of Florida's List of Statistics Job Announcements: *www.stat.ufl.edu/ vlib/jobs.html*

A Day in the Life

AP Statistics Teacher

Brian Kotz

Introduction

I absolutely *love* teaching statistics. Period. No qualifying clauses, no conditional statements—I simply love teaching statistics. I honestly cannot think of another job that I would find as satisfying or that would be better suited to my personality. While I actually enjoy teaching many high school courses, statistics is my particular passion. The discipline is not only interesting, accessible, and directly useful to our students, but it also allows for humorous and hands-on learning by its very nature. Given the kinds of learning and teaching that statistics requires, the high school environment with its schedule, structure, curricular emphasis, and student age group can provide a great climate for statistics education. As I begin my ninth year of teaching high school, I consider myself fortunate to be working with motivated, impressionable students in a field that I feel so excited about. If you are considering teaching statistics at the secondary school level, I hope I will be able to impart some of this enthusiasm while also showing you an honest account of the life of an AP Statistics teacher.

Background

I majored in statistics as an undergraduate, and I consequently selected courses dealing with mathematics teaching, data analysis, and probability in pursuing my Master's in education. Once I decided to become a teacher, I gravitated toward secondary school teaching for several reasons. For one thing, I do not have a Ph.D. Secondly, I had a very positive experience when I attended my secondary school, and I was inspired by many of those teachers. Thirdly, a teacher can be engaged in many areas outside of the classroom in a high school, and I enjoy this additional level of interaction with my students and community. Lastly, my mother has been a teacher for over 35 years, and she has always felt that the high school years in a student's life are critical in the development of both intellect and character. With all of this in mind, I began my statistics teaching career with a sense of purpose, a unique background, and the support of several resources and colleagues. For my first seven years of teaching, I worked at a private coeducational boarding and day school of slightly over 500 students in grades 8–12 in New Jersey. I am now in my second year of teaching at a private day school for girls in grades 4–12 in Washington, DC.

A Typical Day

Many activities and responsibilities are part of a typical high school teaching day. To be honest, the requirements of high school faculty can be as diverse as the schools themselves. Over the years, I have been asked to coach, to chaperone, to run a residence hall in a boarding school, and even to chair a department. But there have also been times and situations where my responsibilities were strictly limited to the classes I was teaching. Again, each school is different, and the needs of each school are different. Four classes has been a standard teaching load in the schools where I have worked, but some schools expect as few as three and as many as six from their teachers. Also, schools will at times reduce a teaching load because of a teacher's other responsibilities. I should note here that while I will be channeling my remarks toward statistics teaching, I think it would be rare that a high school teacher would teach exclusively statistics classes. In addition to my statistics classes I have taught grades 8–12 in algebra, geometry, calculus, and even some discrete math. Again, it depends on needs, enrollments, teacher strengths, and so many other variables.

The Hours before the First Class

On a typical teaching day, I arrive between 7:00 and 7:30 AM. I check mail and e-mail, make copies, meet

Brian Kotz (bkotz@cathedral.org) is a teacher at the National Cathedral School in Washington, DC. He received his undergraduate degree from Harvard (Statistics, 1990) and his Masters in Mathematics Education from Rutgers (Ed. M., 2001). He is originally from West Virginia and is a third-generation teacher. After working as a computer systems consultant, Brian began his teaching career at The Peddie School in Hightstown, New Jersey in 1995.

with students if needed, answer last minute questions from "walk-in" students about an upcoming test or last night's homework, and otherwise prepare for my classes. If one or more students come in early to discuss a statistics topic, we spend most of the time talking about *communication* and *context*.

High school students are just like any other group in that they have varying levels of confidence in their abilities to write well, to speak in public, to ask questions, and to answer questions. When I was a statistics student, my professors and mentors required that I not only understand the theoretical aspects of my studies, but that I also be able to convey my understanding and appreciation of the topics through writing. Analyses could not simply be a summary of numbers and equations. I had to explain how I obtained my results, explain exactly what the slope of an equation meant, develop and explain graphs that supported my findings, etc. I had to write about other real statistical analyses in a comprehensive manner, critiquing the collection of the study's data, the analysis, and the presentation of results. I also had to develop recommendations for improved analysis and discuss these recommendations in a persuasive report or presentation. In some ways, I had to learn how to ask good questions and how to give good answers.

Although it can be occasionally frustrating for the students (and particularly first thing in the morning), I require them to state their questions carefully and with the same standards that they ask of me when I am explaining a topic. The trick is to accomplish this without discouraging the asking of a question on their part. Thinking ahead to the future outside of our classroom, they will have to convey their ideas to the general population, and possibly without the use of technical language. Furthermore, I also take the students to task with their writing. When they feel they are done with an analysis, I often ask them to "remove themselves" mentally from their work and see if the writing makes sense from another person's perspective.

The high school environment affords me the opportunity to work with a student's communication skills on a very detailed level. With small section sizes, frequent class meetings, and the occasional open time before or during the class day, we are able to devote a great deal of time to effective writing and presentation, particularly early on in the school year. Although there is always pressure to complete the entire AP syllabus before the May exam, we still have the time and structure to work on the fine points of communication on a daily basis.

In keeping with the importance of communication, having students understand the proper context of the numbers and tools they are dealing with is equally important. With the technology available today, computing a confidence interval is not very difficult. Likewise, I believe that the actual computation of a confidence interval is a small part of what the course should be about. Therefore interpretation, rather than computation, is stressed in our classes. In a class that has a daily homework requirement, time for reflection, and fewer than 20 students, we are often able to have students drive the conversations about appropriate context. Both in the classroom and during a morning help session, I am able to take on a role that is more "facilitator" than "source of the one right answer." Using the confidence interval example once more, how does a student explain to her peers what a certain 95% confidence interval means? How do you define "confidence"? Can you give me an analogy, or an example, or a way to explain it through a simulation? What do you mean by "95% of the time" — and is that really what you mean? Your remarks imply that a parameter can move — is that true? What assumptions are you making? What is the population for which this interval is appropriate? Can you generalize to other groups? When students attempt to discuss context in this way, it not only personalizes the experience and the material, but it also helps them to understand the material on their own terms rather than through standard wording in a textbook. Again, I feel that the structure of the high school schedule and environment provides a great opportunity for student personalization, connection, and understanding. It is my hope that students see the course as a unifying experience that is applicable to much of the world around them, and that they leave a help session or class as critical thinkers and effective communicators, not just as "number crunchers" with things to memorize.

The Classes

In my current school, there are six or seven class periods of generally 50 minutes in length, but class frequency, schedule, and length vary by school. By 8:00 AM or 9:00 AM, I am teaching my first class. Usually I will have two or three classes in a row followed by a break, so my classes are over in the early afternoon. For the past few years, I have taught one or two classes of 9th grade geometry in addition to one or two classes of 12th grade statistics, both AP and non-AP.

I conduct my statistics classes using various styles and formats. With each topic or concept, we usually perform an activity or exercise. The use of technology such as programmable calculators, statistical software, animated displays, and other tools allows for effective student exploration and helps to accommodate various learning styles and approaches. More importantly, a student's own work is sometimes the best tool I have in convincing the non-believer about many topics such as the Central Limit Theorem, levels of significance, conditional probability, and others. Contrived data sets are convenient, but to pique student interest and to obtain student buy-in, analyzing data that have just been generated via simulations can be a blessing.

My teaching has changed over the years because of advances in educational technology and greater research into students' learning of mathematics and statistics. Again, considering this from a high school perspective, the ability to work closely with a student as he or she develops an understanding of complex concepts is priceless. I can tell a student that deleting an outlier will affect the mean of a data set, and the student can write this fact down. But having a student actually discover this fact based on her own data and calculations or by seeing an interactive visual display that immediately demonstrates this point really drives the lesson home. In the high school setting, I believe that teachers can readily customize the presentation of the material to the strengths and characteristics of the students and not simply deliver the same course or use the exact same methods over and over again, year after year. Furthermore, as the students can receive a high level of attention in the classroom, the teacher can ensure that the students' understanding of the underlying statistical theory is *supported* by the technology rather than *replaced* by technology. In my opinion, as the technology for statistics teaching and learning has become more and more user friendly, high school students actually have become more apt to investigate their own conjectures — and they subsequently treat technology more as a tool for learning than as a thing to learn.

I also try to make the class entertaining and enjoyable. As I said earlier, statistics is an exciting discipline that by its very nature lends itself to interesting, useful, humorous, hands-on learning methods. Random number tables and coins can be useful, but candy tastes better! Small, round, multi-colored chocolate or fruity candies which have markings on one side (I am deliberately avoiding product names here) are delicious and can be used to teach proportions, elementary probability, basic inferential statistics, binomial distributions, chi-square tests for homogeneity, and so on. Computer and calculator simulations can bring out unexpected competition and interest for many students. While the classroom should remain serious and student-centered, I believe that an occasional song, a humorous demonstration, a pun on the word “mean,” or a cartoon regarding true-false tests or improper extrapolation can actually improve the students' attention, memory, understanding, and enjoyment of the material — as well as mine.

After Classes

With my classes over around 1:45, I have some time to take care of a few things before my last school commitments. Occasionally, I will have a brief meeting with a fellow teacher or an administrator regarding a student or school matter. Sometimes, students who cannot come by early in the morning may visit for extra help. Otherwise, I begin grading, planning for the next day, or checking various statistics web sites. Tools such as listservs and the Internet have become such an important part of teaching statistics today, and I have benefited greatly from these resources. I strongly recommend that teachers subscribe to the AP Statistics listserv where one can see questions, issues, and thoughtful responses from teachers from various levels of academia and from all over the world.

My usual academic day concludes with two important activities that allow for the community to come together in meaningful ways. First, at around 2:30 twice a week (or more frequently if needed), I meet with my advisees. In the private schools where I have worked, I have served as an advisor for roughly 8 to 10 students each year. In this capacity, I act as a point of contact for the students, their teachers, and their parents in matters such as course selection, community service requirements, academic performance, and discipline. Then, in the last 30 to 45 minutes of each day, there is time set aside for chapels (twice a week), assemblies (once or twice a week), club meetings (as needed), and a weekly department meeting. The academic portion of the school day ends around 3:30.

After the Academic Day and Outside of the Classroom

Although classes may be over, there can be some days that do not end at 3:30. For example, there are afterschool faculty meetings roughly every other week in which we discuss student performance and behavior, general pedagogy and curriculum, and other school matters. Also, students are genuinely appreciative when they see a teacher come and watch an afterschool athletic event or performance. When I serve as a coach, administrator, or member of a particular committee or project, the 3:30–5:30 PM time slot is frequently assimilated into the rest of the school day.

Many high school teachers choose to involve themselves outside of the classroom, and some schools structure non-academic responsibilities as part of a teacher's job requirement and compensation. In the schools where I have taught, I have had the opportunity to participate in several non-academic facets of school life where I got to know the students well and could contribute my time and effort in meaningful ways. Students always need adult leaders and sponsors for their clubs, publications, and community service organizations, and some teachers also choose to become sponsors of an entire grade and help with the students' governing, fund-raising, and general esprit-de-corps. If a teacher has had success or interest in a particular sport or has played at the intercollegiate level, coaching opportunities may exist.

Athletics and student organizations aside, there are also numerous academic and administrative opportunities for teachers in many high schools. To begin with, there are the usual positions of principal, assistant principal, dean, and department chair. Depending on the school and regulations, some of these may require or recommend coursework in administration or supervision or advanced degrees. Also, committees and groups regarding school policies, pedagogy, and the implementation of initiatives are common (particularly in private schools) and are always appreciative of faculty help and input. Due in part to my interests in how students learn statistics, I have joined a committee at school that helps to inform faculty about the latest research regarding general learning and teaching.

As a statistics educator, I have been specifically and consistently tapped in two areas throughout my career. Whenever there is a survey that needs to be developed or analyzed, I am often asked to lend any expertise or analysis skills to the endeavor. I am more than happy to do this as I have seen many cases of hard work and well meaning research derailed by poorly executed data collection and analysis — and I don't want my students and colleagues to have to go through such heartbreak if I can help to avoid it. If the survey's subject matter, design, or analysis method is appropriate for discussion with students, I also have a ready-made real-world case study and data set for the class. Also, as the importance of quantitative literacy has become more and more apparent to high school teachers, I have often been asked to investigate how more statistics topics can be introduced and mastered prior to AP Statistics. Even with all of the many recommendations and standards out there today, this can be no easy task. While many feel that statistics should be a common thread throughout the K-12 curriculum, great care should be given to determining what is appropriate or relevant for which class.

In high school, the answer often lies in the teacher's comfort level with the material, the ability of the students, and the requirements of the various courses as mandated by the school or state. Consider least squares regression as an example. The interpretation of a linear model's slope as “rise over run” is appropriate for Algebra I. In context, we could even say “feet per second” or “passengers per aircraft” on real data. But can an Algebra I student who just uses the linear regression program on his or her calculator fully understand that these calculator outputs are *estimates* developed from an algorithm? Should we cover what “least-squares” means in 10th grade? If so, how should we do it for that age group? Should we talk about residual plots and patterns at this stage? Do we discuss influential points in 9th grade? How much do I use the calculator, and when? Will the students still be ready for their next math courses or future standardized aptitude tests if I devote a great deal of time to this? These are just some of the questions that may be posed to a high school statistics educator as schools try to develop and implement the best curricula and pedagogy for their students.

Evening

On some evenings, I tutor students from other schools. Whether it is general math or introductory statistics, and whether it is at the high school or college level, there seems to always be a need for tutors. I enjoy tutoring because not only do I feel as if I am helping someone, but I also enjoy seeing how other students comprehend statistics material and how other teachers have chosen to present the material.

In the evening, I work on any classroom preparations that I was unable to get to before leaving school. In addition to writing a test or walking through an exercise for the next day, I may update my grades or work on student comments. Most private schools ask a teacher to write a specific, one paragraph comment for each of his or her classroom students two or three times a year. These are sent home along with grades and are intended to provide a greater insight into a student's strengths, potential areas of improvement, and decorum. While this takes more time than standard grade computation and reporting, it allows the teacher to provide more meaningful feedback and assessment of a student's work.

Professional Development

Many high schools stress or require professional development and continuing education in one's field. Several high school teachers participate in many opportunities during the academic year, and some choose to devote a large part of their summer to both their own continuing education and the professional development of others. Nearly every summer since I started teaching, I have been involved in some activity that was directly related to my statistics teaching and/or to obtaining my Master's degree. I have also served as a reader for the AP Statistics exam for six years, and I would have to say that this surprisingly enjoyable and very useful professional development experience has had an enormous impact on how I have chosen to teach the material.

I am fortunate to be teaching in a discipline that takes its continuing education seriously and has provided several resources for high school teachers. The American Statistical Association, the College Board, textbook companies, and several other organizations have developed workshops, meetings, conferences, publications, web sites, and other opportunities for teachers to share ideas and to stay on the leading edge of statistics education. Some experienced teachers also offer full week workshops for new and experienced teachers each summer. As helpful as the listserv and other tools can be, I honestly feel that there is no substitute for a good statistics workshop or conference. In such venues, one sees how others are successfully teaching certain topics and what fellow educators consider to be important or essential. People can meet face-to-face, exchange ideas, share material, and listen to other teachers who share a similar passion for the teaching of the material and have had similar experiences. I personally benefit from these opportunities as I am removed from the day-to-day perspective of teaching and can consequently take time to reflect about the past year and to think about some of my larger course goals and methods in the company of other experienced professionals. I have seen first-hand how the leaders of the greater statistics education community are an extremely supportive and dedicated group of professionals who care a great deal about the teaching of statistics and are willing to help instructors in any way. This support can be a great comfort to a statistics teacher, and I have often tapped these colleagues for useful teaching ideas and general guidance.

Conclusion

In today's world, we are inundated with graphs, charts, polls, and figures that not only provide insight for day-to-day decision making but also provoke awareness of many issues. The citizen who neither understands this information nor questions how it was obtained or compiled is at a disadvantage. As educators, we need not wait until college to encourage our students to develop these needed skills.

As a teacher of statistics, I know that the material, skills, and applications that we cover will be directly relevant in my students' lives and careers regardless of the vocational or educational paths they choose. Many former students of mine representing a wide range of majors and interests often write back saying that their high school statistics experience proved extremely helpful in both their studies (undergraduate and beyond) and their careers. Statistics is one of the most important classes a student can take in his or her educational career, and many aspects of the high school environment can facilitate a strong and positive statistics experience both for students and teachers alike.

Although we may inspire a student to become an actuary, a biostatistician, or an applied mathematician along the way, ultimately with AP Statistics we are trying to develop intelligent, critical thinkers who will know more than just keystrokes on a calculator and can be informed, contributing citizens in the real world. High school teachers have a unique framework and opportunity to make this happen while contributing to our academic communities and the lives of our students in numerous, meaningful ways.

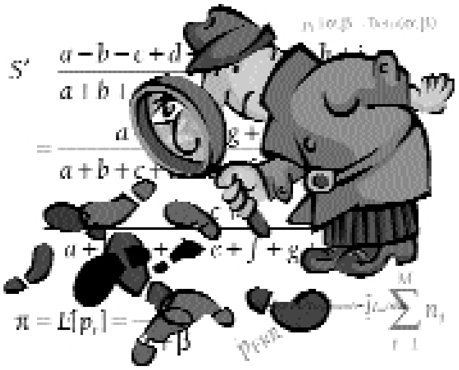
If you are truly interested in teaching introductory statistics to high school students, there is a school for you somewhere out there. As no two schools are alike, carefully consider the mission statement and objectives of the school or school system along with all of the responsibilities and rewards that will come with the specific job. As I stated at the beginning, I love what I do — and I really can't imagine doing anything else.

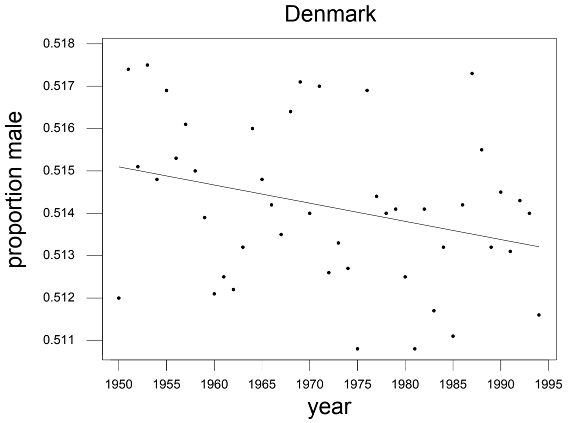
Web Resources

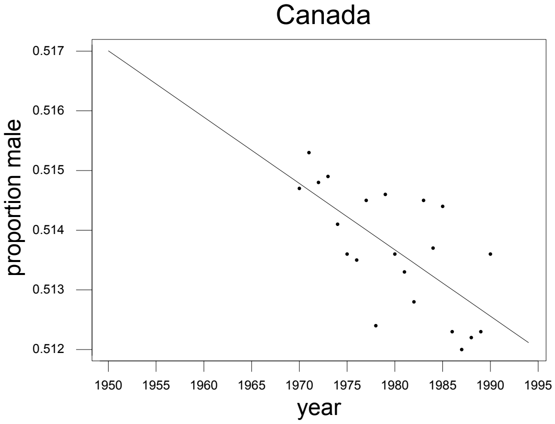
AP Statistics Listserv Archives: *mathforum.org/ epigone/ apstat-l*

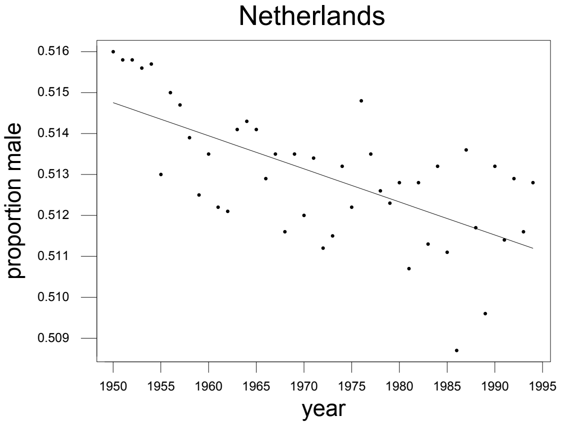
Data Sleuth

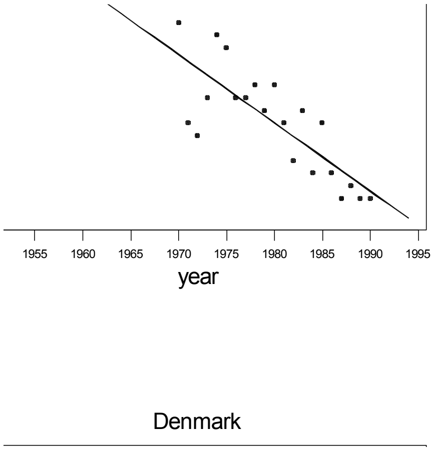
Where Have All the Boys Gone?











The scatterplots display the proportions of male births per year for four countries. Applying least squares regression to these data reveals the following estimates, standard errors, and t-statistics:

Denmark: $\textit{proportion} = .599 - .000043\,\textit{year}$; $\hat{\sigma} = .0018$, $t = -2.07$, $p = .044$
(.041) (.000021)

Netherlands: $\textit{proportion} = .672 - .000081\,\textit{year}$; $\hat{\sigma} = .0012$, $t = -5.71$, $p = .000$
(.041) (.000014)

Canada: $\textit{proportion} = .734 - .000111\,\textit{year}$; $\hat{\sigma} = .00077$, $t = -4.02$, $p = .001$
(.095) (.000028)

United States: $\textit{proportion} = .620 - .000054\,\textit{year}$; $\hat{\sigma} = .00026$, $t = -5.78$, $p = .000$
(.019) (.0000094)

Question #1: Explain why the United States can have the largest of the four t-statistics (in absolute value) even though its slope is only the third largest (in absolute value).

Question #2: Explain why the standard error for the regression slope is smaller for the United States than for Canada, even though the sample size is the same.

Question #3: Can you think of any reason why the standard deviations about the regression line might be different for the four countries? In particular, why does the United States have the smallest variation about the regression line?

Reference

Ramsey, F. and Schafer, D. (2002), *The Statistical Sleuth: A Course in Methods of Data Analysis, 2nd ed.*, Pacific Grove, CA: Duxbury Press, pp. 200–202.

(Solution on page 26)

The Statistical Sports Fan

A Streak Like No Other

Robin Lock

Some of the most fascinating records in sports involve streaks of consecutive events. Examples include UCLA's men's basketball team winning 88 consecutive NCAA games, Paul Cluxton making 94 free throws in a row for the Northern Kentucky Division II basketball team, and Oberlin College playing 44 consecutive football games without a win. One of the most famous streaks of all time was Joe DiMaggio's record of hitting safely in 56 consecutive baseball games during the 1941 season, a feat that Stephen Jay Gould (1989) has called "the greatest accomplishment in the history of baseball, if not all modern sport."

Here is a somewhat more obscure statistic: on July 20, 2003, the Cleveland Indians set a major league baseball record by scoring exactly four runs for the seventh consecutive game (see Table 1).

How likely is it for a team to have a streak of identical scores such as Cleveland achieved? Such a "matching" streak is a bit more complicated to model since the likelihood of the streak continuing at each stage depends on the score that initializes it. Streaks such as DiMaggio's can be analyzed with straightforward geometric probabilities, assuming independence and a constant probability of the streak continuing at each stage (see, for example, Short and Wasserman, 1989).

To investigate the probabilities of identical score streaks in baseball, we'll use a dataset containing scores for all regular season games played during the 2002 season. The data were obtained from the Retrosheet website (*www.retrosheet.org*), where the organizers are compiling computerized records for all historical Major League Baseball (MLB) games. Normally, each of the 30 MLB teams plays 162 games in a season, but occasionally a game that is postponed due to rain will not be made up if the outcome can't affect the season's standings. There were four such games in 2002, so the dataset has 4852 team scores from 2426 games. The distribution of those scores is summarized in Table 2.

Table 1: Scores during the Streak

7/11/03	Cleveland 4	Chicago White Sox 7
7/12/03	Cleveland 4	Chicago White Sox 2
7/13/03	Cleveland 4	Chicago White Sox 7
7/17/03	Cleveland 4	NY Yankees 5
7/18/03	Cleveland 4	NY Yankees 10

Probabilistic Approach

Let p_i represent the proportion of times a team scores exactly i runs in a game and assume that any pair of games are independent. The probability that a pair of scores match can be found by

$$\sum_i p_i^2$$

(think of a tree diagram where the outcomes of interest are branches that match). Using the proportions in Table 2 from the 2002 regular season as estimates for the score probabilities, the probability that two scores match is estimated to be 0.09804. The probability of a scoring streak ending after just a single game would then be 1 – .09804 = .90196. To get a particular score three times in a row would have a probability of p_i^3, so the overall probability of a streak of three (or more) games is

$$\sum_i p_i^3 .$$

Table 2: Frequency of Scores in the 2002 Regular MLB Season

Runs	0	1	2	3	4
Count	275	469	616	681	619
Proportion	0.057	0.097	0.127	0.140	0.1287
Runs	5	6	7	8	9
Count	566	472	337	257	176
Proportion	0.117	0.097	0.069	0.053	0.036
Runs	10	11	12+		Total
Count	136	91	157		4852
Proportion	0.028	0.0198	0.106		1.0000

The chance of a streak ending after *exactly* two games would then be

$$\sum_i p_i^2 - \sum_i p_i^3$$

and the general formula is

$$P(\text{Streak of exactly length } n) = \sum_i p_i^n - \sum_i p_i^{n+1}$$

Using the proportions from Table 2 as estimates for the p_i's, the probabilities of each streak length are summarized in Table 3. This analysis puts the odds of a single streak lasting the seven games accomplished by Cleveland at about 1 in 436,853. If we look for a streak of seven or more games, the probability jumps to about 0.00000262 or 1 in 381,561 streaks.

A Simulation Approach

An alternate method for estimating probabilities of streak lengths during a season is to simulate lots of seasons using all of the scores in the 2002 season as a pool to draw from. Using a computer package (Fathom), we generated seasons for 1000 teams by selecting 162 scores (at random) from the distribution of all 2002 regular season scores. For each of these 1000 "teams" we calculated the number of streaks in the season of length 1, 2, and so forth. The results of the simulation are summarized in Table 4.

Actual Streaks

Both the probabilistic approach and the simulation rely on a number of assumptions that may not be reasonable in practice. For example, the scoring rates probably differ between teams and may not follow the same distribution as the aggregated scores for all teams. Individual games may not be independent. Teams typically play blocks of three or four games in a row against the same opponent in the same city. No allowance is made in the models for whether a team is on the road or at home.

To investigate the degree to which the distribution of actual streaks matches the ones obtained by probabilistic calculation and simulation, we can use the 30 teams that played during the 2002 season to look at the frequencies of their scoring streaks (Table 5). The longest scoring streaks that year, four games each, were accomplished by Los Angeles, Milwaukee, Montreal, Toronto and St. Louis (twice).

Comparing the probabilities obtained in Tables 3, 4 and 5 shows strong agreement between the theoretical calculations, simulation results and actual scoring streaks. A chi-square statistic computed using the actual counts of Table 5 compared to counts based on the theoretical probabilities of Table 3 yields a very modest value of 1.53. So, although the simplifying assumptions may not hold exactly, the probabilities they produce fit the observed data well. Thus it would appear reasonable to estimate that the odds of a scoring streak lasting seven or more consecutive games to be about 1 in 381,561.

But is Cleveland's streak really so unusual? Certainly it's extremely rare for a particular streak to last that long. But professional baseball teams have been playing since 1876, including 170,950 games to the end of the 2002 season, generating a total of 341,900 scores (from league totals for each season at *baseball-reference.com*). If we assume that about 90% of all games are starts of new scoring streaks, there have been a bit more than 300,000 scoring streaks in baseball history. So one streak of at least seven consecutive games scoring the same number of runs during this time period is about what one should expect.

If we assume a probability of $p = 0.00000262$ for any single identical scoring streak to last 7 or more games, the chance of this occurring at least once among N independent streaks can be calculated with $1 - (1 - p)^N$. Based on the 2002 data, we can assume that a single team will have about 146 scoring streaks in a full season. Using these values, Table 6 gives some probabilities for a team to match or exceed Cleveland's streak in various time frames. For comparison purposes, Berry (1991) estimates the probability that a baseball player matches or exceeds DiMaggio's 56 game hitting streak during the next 50 seasons to be about 0.013.

Cleveland's streak ended on Monday, July 21, when Milton Bradley was thrown out at home plate with two outs in the bottom of the eighth inning trying to score the Indian's fourth run in what turned out to be a 4-3 loss to the Chicago White Sox.

Suggestions for Additional Explorations

Baseball is particularly well-suited for this type of exploration because extensive, detailed records are readily available. What about consecutive identical scoring streaks in other sports like ice hockey or soccer (where the distributions of scores are more concentrated on a few smaller values), or football or basketball (where even a pair of matches is probably very unlikely)? More traditional analyses can be applied to winning/losing streaks in various sports or other common streak records such as consecutive shutout innings for a pitcher (would you believe the independence assumption here?) or goal scoring games for a hockey player. Another variation to consider is streaks with a continuous time frame such as consecutive shutout time in hockey or soccer.

References

Berry, S. (1991), "The Summer of '41: A Probabilistic Analysis of DiMaggio's "Streak" and Williams's Average of .406," *Chance* 4(4), 8–11.

Gould, S. J. (1989), "The Streak of Streaks," *Chance* 2(2), 10–16.

Table 6: Chances of Seeing at Least One Identical Scoring Streak of 7 or More Games

Time	Probability
Single streak	0.00000262
Single season (one team)	0.000383
Single season (30 teams)	0.0114
Fifty seasons (30 teams)	0.437

Short, T. and Wasserman, L. (1989), "Should We Be Surprised by the Streak of Streaks?" *Chance* 2(2), 13.

Web Resources

- A repository of game logs for most major league baseball games: *www.retrosheet.org*
- Seasonal statistical summaries going back to 1876: *baseball-reference.com*

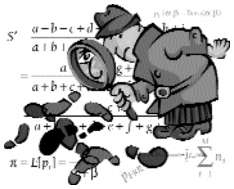
Solutions to Data Sleuth Mystery:

Question #1:

There are several factors that affect the *t* statistic: *t* = observed slope / standard error (slope). So even if the observed slope is smaller, the standard error could be even smaller, resulting in a larger *t* value. The standard error of the slope depends on things like the sample size and the variability in the explanatory variable.

Question #2:

The standard error of the slope also depends on $\hat{\sigma}$, the variability about the regression line for that country. Note that $\hat{\sigma}$ is much smaller for the U.S. than for Canada.



Question #3:

There could be less variability about the regression line in the U.S. than in Canada since there are a larger number of births in the U.S. A sample proportion is similar to a sample mean in that if it is based on more observations, there will be less variability from sample to sample. Presumably the larger number of births per year reduces the random fluctuations in the proportions from year to year.

μ-sings

The Greatest Experiment Ever Performed on Women

Chris Olsen

OK, I guess I'm going to have to confess. In the manner of Will Rogers, I never met a bookstore I didn't like. Some folks like antique shops, some are into clothes emporia, and some – on the other side of a vast cultural divide from moi — go to modern art museums. For me, that combination of a deer-in-the-headlight and Dracula-descending-on-a-blood-bank visage appears at the slightest indication of tomes for sale around the corner. Recently, though, my book affliction got me into some serious trouble. In hindsight, I suppose it was fate that intervened that day when a short dinner with my daughter combined with a quick raid on the bookstore next door. My hurried eyes (apparently blurred because of my hungry stomach) caught sight of what was clearly a statistics book on the "new non-fiction" table, its title: *The Greatest Experiment Ever Performed*.

Wow!, thought I to myself. As a statistics teacher always on the prowl for new examples, this appears to be a great find. Like the similarly titled *The Perfect Storm*, this book must spin an exciting story, probably combining a perfect placebo, a regal randomization technique, and above all, a superlative statistical analysis. I snapped it up without further ado and toddled off with my also hungry daughter. It was only later, <editor: please insert Phantom-of-the-Opera organ music here> when I picked it up to read while lying in bed, that my experimental design euphoria turned to a flashback of Edgar Allen Poe's short story, "The Pit and the Pendulum." I started reading — naturally enough — on the title page and discovered there was more to the title than I had seen, and even a subtitle.

"The Greatest Experiment Ever Performed…" (so far so good) … "…On Women: ..." <Swissshhhhh> … "…Exploding the Estrogen Myth." <SWISSSHHHHH!!!> Gulp!, thought I to myself — how did I get into THIS pit??? As a statistics teacher on the testosterone-based side of *homo sapiens*, not only did I have significantly fewer clues than one about whatever estrogen is, I was also seriously less than familiar with any of the purported estrogen myths! (However, the title did have the word "Exploding" in it, so I figured it might still be OK for a guy to read.) As it turns out, it was not only OK, it was terrific. Written by Barbara Seaman, an advanced science writing Fellow at Columbia University's School of Journalism, *Greatest Experiment* is a fascinating story of investigators, doctors, drug companies, and the conduct of scientific research.

On July 9, 2002, it was publicly reported that the safety-monitoring board had halted the Prempro arm of the Women's Health Initiative, a 5-year study involving over 16,000 women. Prempro, for the unititiated (i.e., men), is a drug that combines Premarin, an estrogen named for its source — pregnant mare urine (I am NOT making this up!) — and Provera, a synthetic progesterone which balances out some bad effects of estrogen for some menopausal women. The WHI study volunteers who had taken Prempro had fewer bone fractures and less colon cancer; however, they also had more breast cancer, heart attacks, strokes, pulmonary embolisms, and blood clots than those on a placebo. Now, one might suppose at first blush that science had once again triumphed – a randomized clinical trial performed in service to women's health. That clinical trial, however, is only a small part of the total story. By that day in 2002, drugs like Prempro – chemicals that mimic women's natural hormones — had already been used by hundreds of millions of women to stave off the effects of menopause.

Ms. Seaman's book introduces the development of menopause treatment over the course of the late 19th and 20th centuries, but the story really gets going in 1938, when a British biochemist published his formula for oral estrogen, in order to prevent Nazi Germany from cornering the world market. This cheap and powerful drug was snapped up by doctors and drug companies the world over, the genesis of what Seaman calls the "Greatest Experiment." In her words,

> "I call the marketing, prescribing, and sale of these drugs an experiment because, for all these years, they have been used, in the main, for what doctors and scientists hope or believe they can do, not for what they *know* the products can do. Medical policy

> on estrogens has been to 'shoot first and apologize later' — to prescribe the drugs for a certain health problem and then see if there is a positive result."

Granting Ms. Seaman her "Greatest Experiment" terminology, the history she presents certainly does not compare with the Nazi experiments of World War II, or to the Tuskeegee syphilis experiments in the United States, but she presents a cautionary tale very effectively. The marketing of Prempro as — in effect — a fountain of youth pill for women had no controlled studies or otherwise objective evidence to back up such a claim! To make a long story short, the touted advantages of Prempro, the most commonly prescribed form of hormone replacement therapy (HRT), were shown to be either nonexistent, or seriously open to question by the data in the WHI study. Based on these results and more recent studies, and a re-awakening to the — in hindsight — inexorable march of negative evidence, the *Journal of the American Medical Association,* the *New England Journal of Medicine*, and the British Journal, *Lancet*, have in no uncertain terms editorialized against the casual use of HRT.

It is here perhaps that we can see a potential role for us as students and teachers of statistics and interpreters of the scientific enterprise. A full 7 years before the halting of the Prempro study, *Time* magazine (June 26, 1995) ended a cover story about estrogen as follows: "As is so often the case in modern medicine, the most a patient can ask of her doctor is to lay out the risks, the benefits and the honest fact that the data are inadequate, and then let her make the choice." If that is so, we consumers of medical practice need to have more than a mantra-like definitional understanding of the differences among anecdotal evidence, observational studies and experiments. We need a vocabulary that will allow us to navigate the competing claims of medical studies, and a working knowledge of the role of statistics as a foundation for claims in medical studies. Ms. Seaman echoes *Time*'s sentiments: "What we all need to do now, patients as well as doctors, is to learn how to evaluate the *quality* of information."

So, in closing, let me offer some sage words of advice to all you statistical bookworms out there. First, most definitely carve out a weekend to read Ms. Seaman's well-crafted and well-documented work, *The Greatest Experiment Ever Performed on Women*. (Yes, you too, guys.) Second, resolve to reflect on the implications of our collective ignorance of things medical. Third, take your daughters out to dinner more often — you may learn a lot of fascinating stuff.

References

Commentary. (2003), "Breast Cancer and Hormone-Replacement Therapy: Up to General Practice to Pick up the Pieces," *The Lancet* 362:414–415.

Fletcher, S. W., and Colditz, G. A. (2002), "Failure of Estrogen Plus Progestin Therapy for Prevention," *The Journal of the American Medical Association* 288(3):366–367.

Herrington, D. M, and Howard, T. D. (2003), "From Presumed Benefit to Potential Harm — Hormone Therapy and Heart Disease," *The New England Journal of Medicine* 349(6):519–521.

Seaman, B. (2003), *The Greatest Experiment Ever Performed on Women,* New York: Hyperion Books. (ISBN: 0-7868-6853-8).