STATS

STATS

The Magazine for Students of Statistics

Spring 2002 • Number 34

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

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

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

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

Features

Departments

Editors' Column

Beth Chance Allan Rossman

One of the hottest recent trends is "data mining," a catchy phrase that certainly sounds as if it involves a lot of statistics. As Dick DeVeaux writes in our lead article of this issue, the term means many different things to different people. Dick offers not only an informative overview but also a personal perspective about how statisticians can offer a lot to, while also learning a lot from, data mining. One does not have to dig very deeply to find valuable nuggets of information in this article. We hope that his article will inspire students (and others) to learn more about this important topic and perhaps even go on to make contributions to its development.

Statistics has long been considered a career that requires a graduate degree, but more and more attention is being paid to career opportunities for students receiving degrees in statistics at the bachelor's level. Patti Costello and Lisa Kay offer a glimpse of the diversity of opportunities available by summarizing results from an alumni survey that they recently conducted for Eastern Kentucky University. You might be surprised at all the different options there are with a B.S. in statistics!

Continuing a tradition of the past several years of this magazine, we include a report on the AP Statistics exam from Roxy Peck, Chief Faculty Consultant of that program. Roxy notes that not only is the rapid growth in numbers of students taking this exam an encouraging sign, but the quality of students' performance improved last year as well. Roxy's article also provides some advice for students about communication, an important aspect of statistical analysis on which many students struggle.

Carrying backpacks full of books is a daily routine for most students, but carrying too much weight can actually lead to serious health problems. To investigate whether students on their campus tend to carry more than is recommended, four Cal Poly students weighed backpacks of their peers, in relation to students' body weight, and analyzed the results. We are pleased to present a report of their project, and we encourage other students to submit articles based on their statistics projects.

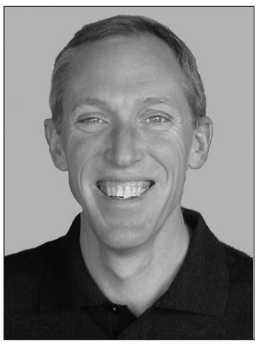
Chris Olsen provides more μ-*sings* for *STATS* readers in this issue, this time through a movie review. Chris calls our attention to an old, little-known movie called *Magic Town* in which Jimmy Stewart stars as a pollster seeking the Holy Grail of representative samples. Robin Lock again assumes his role as *The Statistical Sports Fan*, this time investigating the age-old question about whether offense or defense is more important for success in team sports. We also feature another article by Gretchen Davis, motivated by the difficulties AP Statistics students have had in recognizing paired data, in which she highlights two examples illustrating the advantages of a matched pairs design and the appropriate analysis of such data.

We also offer in this issue another installment of "Data Sleuth," in which we invite readers to solve mysteries based on the analysis of data. Patti Collings has provided two mysteries for this issue, one based on winning times in the Kentucky Derby and the other concerning ages of students in her undergraduate statistics classes. We invite readers to send us other mysteries for inclusion in future issues.

We also invite you to check some of the new features that we have installed on the website for STATS magazine (*www.amstat.org/publications/stats*). Not only does the site provide background information about the magazine and tables of contents for recent issues, but it also contains links to datasets and to other web resources discussed in the articles.

We both continue to be students of statistics as we continually learn new things, but it has been a while since either of us was a student in a formal sense. We would very much appreciate receiving your feedback, suggestions, and contributions, especially from our student readers.

Data Mining: A View from Down in the Pit



Richard D. De Veaux

Have you ever ordered a book at *amazon.com* or a CD at *cdnow.com* and then received an e-mail suggestion for another product to buy? Have you ever received coupons from your local supermarket based on your personal buying habits? Have you ever called a large mail order company and had the operator address you by your first name, ask how you liked the socks you ordered last month, and tell you about the specials you might be interested in? Have you wondered how such organizations use the data that they have acquired about you to make predictions about your habits and tastes? The answer lies in the important and growing, but vaguely defined, field called data mining.

In this article I hope to provide a short overview on what data mining is and how it differs, if in fact it does differ, from statistics. Because data mining is such a large field, it is impossible to cover all that it does and purports to do in a short article. And I certainly can't do justice to everyone who has contributed to it. Instead, I would just like to give a personal perspective of how data mining has changed the way I do data analysis and the kinds of problems I'm willing to tackle.

Dick De Veaux has degrees in Civil Engineering, Math, Dance Education and a PhD in Statistics. He is professor of Statistics at Williams College, an industry consultant for data mining, and is currently a visiting Professor at the Laboratoire de Probabilité et Statistique at the Université Paul Sabatier in Toulouse, France. Dick's other interests include singing (he is the bass in a Doo-wop quartet at Williams called the Diminished Faculty), cycling, and dance (he teaches modern dance at Williams during Winter Study). He is the co-author, with Paul Velleman of Cornell University, of a forthcoming introductory statistics textbook to be published by Addison Wesley in 2003.

What is Data Mining?

Data mining has been defined in almost as many ways as there are authors who have written about it. Because it sits at the interface between Machine Learning, Database Management, Data Visualization and Statistics (to name some of the fields), the definition changes with the perspective of the user. Here is a not so random sample of a few:

Data mining is …

> "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data."
> — Fayyad (Machine learning)

> "a knowledge discovery process of extracting previously unknown, actionable information from very large databases."
> — Zornes (Information technology [IT])

> "finding interesting structure (patterns, statistical models, relationships) in databases."
> —Fayyad, Chaduri and Bradley (Machine learning)

> "a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions."
> — Edelstein (Data warehousing, IT)

For others, data mining is not only the modeling and prediction steps, but a whole problem solving cycle. Funded by a European consortium, a group of data mining experts put together a methodology called CRISP-DM (CRoss-Industry Standard Process for Data Mining) in an attempt to provide a framework for data mining. In it are six distinct steps:

- Business understanding

- Data understanding
- Data preparation
- Modeling
- Evaluation
- Deployment

What's clear is that data mining remains a vaguely defined field. It covers a lot of what statisticians do, and it has become an important commercial endeavor. For this reason alone, statisticians and students of statistics should pay attention to those claiming to be data miners and to what they do.

Common to most of the definitions of data mining are several themes. One of the most important is that the results be useful. "Actionable, useful, knowledge discovery" are some of the words used in the data mining jargon. Sometimes, the goal of a data mining exercise is simply to produce a predictive model. However, useful may also mean interpretable. Many of the algorithms have models that are so complicated that they become uninterpretable black boxes and simply give a prediction. The user must decide which class of data mining algorithms to use depending on the goal.

Another data mining theme is the size of the databases. Indeed, many commercial and scientific databases truly are big. A transactional or scientific database may run into terabytes (1TB = 1000 gigabytes GB) of data. The UPS tracking database is reported to be about 16 TB large. By comparison, if all the books in the U.S. Library of Congress were digitized, they would amount to about 10TB. The National Climate Database (NOAA) is about 400TB.

Last, data mining is almost always a *team effort*. Rarely is one person responsible for the collection, storage, maintenance, extraction, cleaning and modeling of data. If you are going to be involved in a data mining project, you should be prepared to at least talk to and get friendly with people who have these other skills. You'll need their help. Otherwise you should be prepared to learn a lot about database management, SQL (Structured Query Language), and PERL (Practical Extraction and Report Language). These might be worth learning anyway, especially if you are going to specialize in problems with transactional databases. However, for the rest of this article, I will focus more on moderate size problems and concentrate on the part that most resembles statistical analysis.

"So, *is* data mining different from statistics?" This was the question asked to Jerry Friedman, a prominent statistician and data mining expert, at a recent conference after his data mining talk. Jerry paused and before answering it, asked if the questioner wanted the long answer or the short answer. Since the short answer was requested, Jerry's answer was, "No." The long answer might well have been a bit different, as I'll attempt to explain.

Why Do Data mining?

Data mining got its start in what is now known as customer relationship management (CRM). After creating large databases for keeping information such as transaction records, inventory, and billing history, companies began to realize that they might have an enormous amount of information that they could use not just to store transactions, but also to learn more about their customers. Using the information in the database, they can learn not only how to retain customers, but also to market to them more effectively and to create more opportunities for cross selling. By using predictive models, they hope to be able to suggest the products that you have the highest chance of buying, based on your past behavior and demographic information. With the blossoming of the Internet and the possibility of capturing even more data online in the late 90's, CRM became big business.

What Kinds of Problems?

CRM isn't the only application for data mining. In fact, interest is everywhere. Even hallowed traditional statistical areas like clinical trials are no longer safe from data mining. With the addition of ancillary genetic information on the participants in a trial, the possibility exists for adding this information to the models. Currently I am working on two projects. The first is a metal manufacturing problem involving cracking during the production process. Hundreds of process variables are measured during the molding process and the response variable is whether the molded piece is cracked or not. Since each cracked piece costs tens of thousands of dollars (these are *large* pieces of metal) to recast, and the cracking rate is annoyingly high, there is a lot of potential for savings. The second project comes from the insurance industry. Actuaries carefully price insurance policies. For example, the price of the same amount of liability insurance on an automobile depends on the age and sex of the principal driver, the state in which it's registered, and the driving record of the policy holder, among other things. We have data on tens of thousands of mature policies, policies that have reached the end of their term. For each we have many more predictors than were used to price the policy. For example, if it's a commercial vehicle, we might consider the type of industry it's used in, something not usually used to price the policy. There are dozens to hundreds of such extra predictors. Each of these policies either involved a loss (a claim was paid greater than the premium) or not, and we have information on the amount of loss as well. The challenge is to see whether we can predict which policies are more likely to result in losses. The signals here are small. Most of what leads to losses on insurance policies is random. But if there is any

predictability at all, the payoff can be enormous. An increase of a percent or two in profit for a billion dollar industry is not negligible.

Both of these are typical of the kinds of problems found in data mining. The data sets are large, the signal is small (if it's there at all), and the data are messy (they require a lot of pre-processing), but the potential payoffs are great. Interestingly, many of the techniques used for these problems are statistical models. As a profession, we should ask why there is such interest now in doing these kinds of analyses when we've had the capability for a long time. Is it just because "data mining" sounds more interesting than "statistical analysis"? It's food for thought.

What's Different? An Example.

How is data mining different from statistics? Let's start this discussion by considering an example. This example was used at The Fourth International Conference on Knowledge Discovery and Data Mining for a data mining competition known as the 1998 KDD Cup. The data are from one of the largest direct mail fund raising organizations, the Paralyzed Veterans of America (PVA). Here is the description of the problem from the KDD 98 Cup web site:

"Participants in the CUP will demonstrate the performance of their tool by analyzing the results of one of PVA's recent fund raising appeals. This mailing was dropped in June 1997 to a total of 3.5 million PVA donors. It included a gift "premium" of personalized name & address labels plus an assortment of 10 note cards and envelopes. All of the donors who received this mailing were acquired by PVA through premium-oriented appeals like this. The analysis data set will include:

- A subset of the 3.5 million donors sent this appeal
- A flag to indicate respondents to the appeal and the dollar amount of their donation
- PVA promotion and giving history
- Overlay demographics, including a mix of household and area level data.

The objective of the analysis will be to identify response to this mailing — a classification or discrimination problem."

The entrants were given a data set with 100,000 rows and 481 predictors (currently housed at *http://kdd.ics.uci.edu/databases/kddcup98/kddcup98.html*). The objective was to build a model and select those past donors *most likely* to return a donation from the current appeal.

One of the ways data mining problems are different is that it's very hard to know how to start. A fact of data mining life is that most of your effort will go into preparing the data before you can do anything that statisticians might consider fun. It's instructive to spend a few minutes looking at the definitions of some of the predictors in the file: *http://kdd.ics.uci.edu/databases/kddcup98/epsilon_mirror/cup98dic.txt* .

For example, here are some of the variables describing the number of children in the household:

CHILD03	Presence of Children age 0-3 B = Both, F = Female, M = Male
CHILD07	Presence of Children age 4-7
CHILD12	Presence of Children age 8-12
CHILD18	Presence of Children age 13-18

So, a household with a 4 year old boy and an 11 year old girl would have the following four outcomes:

CHILD03	CHILD07	CHILD12	CHILD18
	M	F	

Generating four categorical variables, two of which will have missing values, screams for variable re-expression. Is it best to use number of children and mean age? It's not clear. It depends of the application. If you're selling children's building blocks, the answer will be different than if you are soliciting contributions for disabled veterans. Think about the amount of time needed to get 481 variables like this into shape to do anything.

After you've checked the variables for definitional differences, missing value codes, variable type, inconsistencies, and errors (remember you're doing this for possibly hundreds of variables), you have too many variables to start doing exploration in the usual way. I was taught that the first three rules of data analysis are:

1. Draw a picture
2. Draw a picture
3. Draw a picture

If you have 481 predictor variables, it's not clear how to do that. Even plotting the distribution of each variable becomes too much, and pairwise scatterplots are impossible. There are clever graphical packages available now that enable you to plot a good handful of variables at once, either by projection or slicing through the data with the use of slider bars. But even then, you have to know which dozen or so variables you want to look at.

Instead, models can often be used for the types of exploration that we traditionally did first graphically. This exploratory data modeling (EDM) seems to be at odds with traditional statistical analysis. But with so many predictors, we have few other choices. I often use a decision tree as a first step. In such a model, the data are recursively split into groups based on an optimal separation of the response variable. Looking at which variables the tree uses first can give us a reasonably sized subset of predictors to explore. For example, the first split of the response variable for the PVA data was on a variable called RFA_2. This variable contained

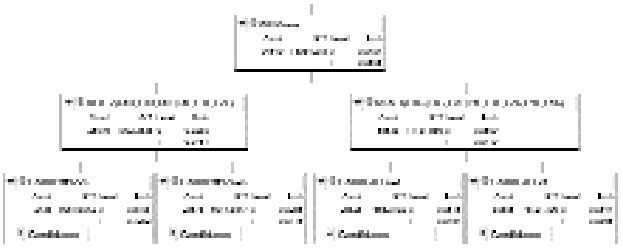


Figure 1: First Two Levels of Decision Tree for PVA Data

information on the past giving behavior of the donor. Figure 1 shows the first two levels of the decision tree (from a *Beta* version of *JMP version 5.0*).

What this model indicates is that the most important variables have to do with past giving and past mailings. RFA stands for recency, frequency and amount. Marketing scientists often use this 3 dimensional summary of buying (or in this case, giving) history, so it's not surprising that it appears as the most important variable. The other two have to do with how many other mailings were made and responded to. CARDPM12 contains the number of card promotions received in the last 12 months. Notice that the group with CARDPM12 < 5 received fewer than 5 promotions in the last 12 months and were nearly twice as likely (.1202 vs. .0715) to respond now. CARDGIFT measures the number of lifetime gifts the donor has given. The group with CARDGIFT ≥ 4 has given more often in the past and is somewhat more likely to respond now (.0484 vs. .0356). The next two levels would show the variables LASTDATE (indicating the date of the last gift) and HVP2, a demographic variable indicating what percentage of the census tract has home value over $150,000. More recent donors are more likely to respond now. Those in wealthier neighborhoods are also more likely. All but the last variable have to do with recency, frequency and amount given.

But the point here is not to use this model as it is. It's wrong (as are all models), and it's not even clear yet whether it's useful as a predictive model. However, it does give us an idea of where to start looking. There are 481 potential predictors, more than a dozen concerning the ethnic makeup in the census tract of the donor alone. What this simple tree suggests is that concentrating on the past behavior of both the donor and the organization's interaction with the donor might be a good place to start. After selecting a dozen (or two dozen) predictors, one could start applying traditional exploratory methods to those variables before moving back to modeling again.

Another common occurrence in the preliminary stages of data mining is what I call the "unusable predictor phenomenon." The very first time I ran a tree model for the PVA response, I got 100% accuracy with one predictor. I was suspicious, and it turned out to be for good reason. Whenever you get results too good to be true, they are. It turns out that in the database there are actually two response variables, Target_B which indicates *whether* the client responded and Target_D which shows the *amount* donated. I hadn't seen that Target _D was in the database, so it was in my first set of predictors. Not surprisingly, if you know how *much* someone donates, you have a pretty good idea of *whether* they donated. So it's perfect model, but totally unusable.

In the metal cracking problem, it turned out that the composition of the alloy used was the most important variable in predicting cracking rates. Unfortunately, the engineers and chemists were well aware of that. In fact, they were most interested in the high cracking rate alloys. This problem of unusable predictors being identified as most important occurs in nearly *every* data mining situation I've encountered.

Sometimes the problem is slightly more subtle. There can be information in what seems like a useless predictor. Once, in predicting the probability of a credit card default, we found that the account number of the card holder was an important predictor. Useless? Not completely. Because the account numbers had been given sequentially, the lower numbers belonged to customers who had held accounts the longest. The customers acquired later had a much higher probability of default. Using "time on book" turned out to be the predictor of interest, but it was masked by the account number because of the high correlation.

At first, this approach seems backward from the way most of us were taught. When possible, we should explore the data set graphically and think carefully about the model before proceeding. But with more predictors than one could possibly look at, using models as exploratory tools can be a useful way to start looking at a data mining problem.

What Kinds of Models?

Data mining models almost always remain exploratory in the sense that we don't take great care to specify error structures or even the functional form of the model. Although statisticians generally heed George Box's famous "All models are wrong, but some are useful," this is even more true for data mining models. The emphasis in data mining is more on usefulness and prediction and less on getting the model "right."

I'm often asked if this bothers me. Couldn't we get better results if we invested the time to think carefully about the model and error structure? Of course we could. Certainly there are many areas, time series and spatial statistics come immediately to mind, where ignoring fairly basic dependencies in the error structure can get you into big trouble. You have to ask yourself in the context of the problem at hand whether it's worth the extra effort. Suppose you're trying to decide which customers on your mailing list should get the Spring Catalog and a simple predictive model can tell you

which 60% of your customers are most likely to purchase. If this simultaneously increases your selling percentage by 15% and saves 40% in mailing costs, you may not want to spend a month trying to come up with a "better" model. It's a question of degree and perspective.

The Algorithms

Much of the technical discussion about data mining concerns the algorithms used for modeling data. When general data mining software first became available, the packages usually included, at a minimum, decision trees of various kinds, neural networks, K-nearest neighbor (KNN) methods, K-means, clustering methods and association rules. With the exception of trees and some clustering methods, these were largely unknown by statisticians, and so the line between data mining and statistics software was fairly clear. Now, general data mining software usually includes regression, especially logistic regression. At the same time, statisticians have also begun enlarging their toolboxes a bit, so the line has become fuzzier. Since the methods least studied by students of statistics are KNN methods, K-means, and neural networks, I will try to give a short overview of these methods.

All three methods are essentially model-free. For that reason, they can't be used to understand the nature of the relationships between predictors and response. But, these *black box* methods are often among the best performers for prediction in real data problems. Both K-means and KNN methods are primarily classifiers. That is, they are used to predict a discrete response variable— the class to which the new observation belongs. But, once that is done, a continuous response can also be estimated, for example, by taking the mean of the response for all observations in that class. Neural networks are used for both classification and regression.

Both K-means and KNN methods are fairly simple and base their predictions on neighboring points. So, they need some type of distance measure to decide which points are close. If you have all continuous predictors and if they are all on the same scale this is relatively easy to do. But in a real data problem like the PVA example, it's not clear at all how to compare someone's age with which magazines they buy. Often, most of the work in using these methods is in defining the distance measure.

K-nearest Neighbors

In its simplest guise as a classifier, KNN works as follows. Suppose we want to predict the category of a target point x_0. We simply find the *k* points *closest* to x_0. Then we take a majority vote of their category and use that to predict the category of x_0. KNN methods are the backbone of most software that recommends other purchases for you to make based on your past purchases, such as you'd find on *amazon.com* or *cdnow.com*.

Here's a simple example that gives the spirit of how K-nearest neighbor methods work. Suppose we want to find what movie to recommend to someone and we start with the following database:

	Star Wars	Batman	Rambo	Chocolat	Whispers	Andre
Lyle	y					y
Ellen	y			y		y
Fred	y	y				
Dean	y	y	y			
Jason	y			y	y	

We have six movies in our catalog and 5 customers. The movies are *Star Wars*, *Batman, Rambo, Chocolat, Cries and Whispers,* and *My Dinner with Andre.* For each combination of customer and movie, we have a *y* if the customer has bought that movie. In a real application, most of the table would, of course, be empty (imagine the database for *amazon.com*).

A new customer Karen enters our store and tell us that she loved *My Dinner with Andre*. What movie should we recommend to her? Using KNN, we could use the metric defined by the number of movies two people have in common. So, the closest neighbors to Karen are Lyle and Ellen. Now, what movies did they buy in common? Right, *Star Wars*. So, should we recommend that Karen buy *Star Wars?* There are two problems here. First, everyone has already seen *Star Wars*, so there's not much point in recommending it. Second, we didn't really use the data. Any movie would have given us the same recommendation. However, if we ignore *Star Wars,* we're left with *Chocolat* to recommend. We could even go further, and include Jason in the neighborhood. Then we can recommend *Cries and Whispers* as well. This simple example illustrates that much of the work in implementing KNN methods is in choosing the appropriate metric and in discounting frequent choices.

K-means

The idea behind K-means is to split the data into K groups. We start with K centers, usually picked at random. Then for each point we identify which cluster it belongs to using an appropriate distance metric. Once we have the memberships, we can recompute the means of these points, obtaining new centers. The two steps are alternated until they converge. The algorithm is iterative, but fast. To predict the category of a new point, you find the cluster it is closest to and then take majority vote within the cluster. For a continuous response, you could take the average as the prediction. Of course, there are many ways to choose the centers of the groups, the assignment to groups, and the numbers of groups, but the basic idea is this group assignment.

This method is especially fast at prediction for a new point because all that's required is to find the right group and to report the relevant statistic. Of course, with a mix of continuous and categorical predictors, finding an appropriate distance measure is still a challenge.

Neural Networks

Neural networks comprise a large field of research, but the ones usually used for data analysis, the so-called multilayer perceptron or back propagation network, are fairly straightforward. The multilayer perceptron is really just a very flexible non-linear regression model. To understand this, it's helpful to start by thinking of a multiple regression problem and by reminding ourselves of two things that we statisticians do all the time.

In any regression problem, we can create new variables from original variables. This includes simple transformations of the original variables or combinations of them. In the language of data mining, this is called *feature creation*. So, for example, if we have both the price and the earnings of a company, the ratio of price to earnings is a new feature. Especially common features are linear combinations of variables. In methods such as principal component regression, factor analysis, multidimensional scaling and even such exotic techniques as projection pursuit regression, the model is based on linear combinations of the original predictors.

Suppose we create k new linear combinations of the inputs (the predictors) and denote them as $z_1, \ldots z_K$. We'll call these the *hidden nodes* of the network. Then we transform each of these nodes by what's called a sigmoidal function (sigmoidal means S-shaped). Usually the logistic function is used for this. Now, we perform a linear regression on these transformed variables, and that predicts the response. If we want the response to be limited to a number between 0 and 1, as we might for a classification problem, we can also apply a sigmoidal transformation to this prediction. (Then we can interpret the response as a probability of membership.)

What do we get? We get a very flexible model with lots of parameters that can fit just about anything. Why? Because we haven't limited the number of new features k. In fact, k can be larger than the number of original variables if you like. The coefficients making up the linear combinations and the regression coefficients in the next stage are all parameters to be estimated. We can easily generate models with as many parameters as we want. In doing so, we can fit any training set of data perfectly, a point not lost on neural network salespeople. Of course, the trick is in keeping the network from fitting the data *too well* – overfitting. Overfitting is a major problem of neural networks and computer intensive methods in general. An overfit model fits the data at hand, the *training data* well, but won't generalize well out of sample since the model has fit the noise as well as the signal.

But with proper care to avoid overfitting, neural networks provide generally reliable predictions in many situations. I usually use neural networks as a benchmark in the search for simpler, more interpretable models. I, like many statisticians, always hope that I can find a model with three predictors, maybe a 2-way interaction of two of them, and some simple smooth functions of them that describe the response. If I find such a simple model and it does as well, then I'm happy. Not only will I have good predictions, but I can also explain the model to my client. If, however, I try it on a test set of some sort and it performs only a fraction as well as the neural network, then I know I have work to do. Rarely do I directly use the neural network for the kinds of applications I work on, because almost invariably, the client wants to know "what's going on" inside the model. Because of the large number of parameters, there are many similar models that can generate the same predictions. The individual coefficients of any particular model become meaningless, and the model is uninterpretable. However, variable importance measures and some recent profiling software has become available for some neural network packages that does enable one to get insights. (For more details on neural networks, see De Veaux and Ungar (1996).)

Model Assessment

Traditional statistical models are nice to work with because of the asymptotically based theory that usually accompanies them to produce confidence intervals and test hypotheses. For trees, neural networks and K-nearest neighbor methods, no such theory exists. So, we have to rely on the data themselves to test the model. The standard way to do this is by some sort of cross-validation.

When the data set is very large, we often simply split (randomly) the data into two parts, one for training (i.e., *building*) the model, and one for testing. We save the test data set until we have our preliminary model built based on the training data and then see how well it performs on the test data. For a continuous response, residual sum of squares or R^2 (the square of the correlation between predicted and actual) is often used. For a classification problem, there is a choice among a variety of measures based on how often the correct class is predicted. Sometimes these are weighted by the different costs of the types of wrong prediction.

If the data set is not large, then we have to get around the problem by K-fold cross-validation. We split the data into K parts, using all but one each time to train the model and the other as a test set. So if $K = 10$, our data might look like this:

1	2	3	4	5	6	7	8	9	10
Train	Train	Train	Train	Train	Train	TEST	Train	Train	Train

Here we've shown the 7th time. We've used all but the 7th part of the data to train the model. We'll assess how well it works by predicting on the 7th part. We repeat this 10 times and average the performance across the 10 test sets. This method is often used for choosing a parameter of the model, like the number of nodes to use for a decision tree.

One caveat to worry about is that the *real* prediction error is almost always going to be larger than the one you calculated from cross-validation. This is because by randomly selecting within the same data set, you'll almost always wind up with a more homogeneous data set than you'll encounter when you go to actually deploy the model.

Data Mining Myths

One of the reasons for the success of data mining (or at least the success of data mining software) is the allure that with the purchase of one package, one can avoid all the unpleasant things that statisticians warn against. Some of the myths of data mining include the beliefs that

- data mining models can take care of bad data
- data mining eliminates the need to understand the data or the business problem
- data mining algorithms automatically find interesting patterns in data
- data mining eliminates the need to know any statistics.

Statisticians entering into a data mining project should realize that some of the people on the project teams firmly believe these myths. That may be why they signed up for the project. But data mining is a team effort. Unless you are prepared to be a database and data warehouse expert, a PERL and SQL programmer, a data preparation expert, and a subject matter expert, you're going to need help. On the other hand, data mining projects provide a great way for statisticians to get involved with large projects. We need to leverage the momentum of the interest generated by data mining to contribute effectively. If we don't get involved with data mining projects, other people will step in to do our work.

Summary

Is data mining the same as statistics? Well, yes and no. If we open up our toolboxes a little bit to some of the exciting things in machine learning and artificial intelligence, we'll find that much of it is familiar. But it also encompasses problem solving and deployment as well as just building a predictive model. For that reason it tends to be a team effort requiring communication and interpersonal as well as technical skills. But, it can provide enormous payoffs for the statistician who is willing to jump into the fray.

Further Reading

A good introduction to data mining from a marketing perspective:

Berry, M. and Linoff, G. (1997), *Data Mining Techniques*, New York: John Wiley & Sons, Inc.

Both of these are must reads:

Cleveland, W. S. (1994), *The Elements of Graphing Data* (revised edition), Summit, NJ: Hobart Press.

Wainer, H. (1997), *Visual Revelations*, New York: Copernicus.

A different slant on some of the more popular data mining tools:

Dhar, V. and Stein, R. (1997), *Seven Methods for Transforming Corporate Data into Business Intelligence*, Upper Saddle River, NJ: Prentice Hall, Inc.

A more academic sampling of the kinds of problems encountered by data mining:

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (eds.) (1996), *Advances in Knowledge Discovery and Data Mining*. AAAI/MIT Press.

A practical introduction to data preparation. One of the only books on the subject:

Pyle, D. (1999), *Data Preparation for Data Mining*, San Francisco: Morgan Kaufmann.

For the statistics student, an excellent overview of the methodology of data mining:

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning: Data mining, Inference and Prediction*, New York: Spring Verlag.

An excellent in-depth analysis of K nearest neighbor methods:

Friedman, J. (1997), "On Bias, Variance, 0/1 Loss, and the Curse of Dimensionality," *Data Mining and Knowledge Discovery*, 1(1), 55-77.

A practical introduction to the challenges of data mining:

Edelstein, H. (2001), "Introduction to Data Mining and Knowledge Discovery"(Third Edition), available from http://www.twocrows.com.

References

De Veaux, R. and Ungar, L. H. (1996), "A Brief Introduction to Neural Networks," available at http://www.williams.edu/Mathematics/rdeveaux/papers/amstat.ps .

Web Resources

Here are two websites full of other links and useful information:

http://www.twocrows.com/dm-links.htm (Two Crows)

http://www.kdnuggets.com/ (KD Nuggets)

Where Do All of the Undergraduate Statistics Majors Go?

Lisa W. Kay Patricia S. Costello

Dick Scheaffer (2001), while serving last year as President of the American Statistical Association (ASA), wrote, "Statistics degrees awarded in the United States in the 1999–2000 academic year (self-reported data from Amstat Online) amounted to about 1100 bachelor's degrees, 1600 master's degrees, and 460 doctorates." Where do all of the B.S. statisticians go after graduation? At Eastern Kentucky University, which offers a bachelor's degree in Statistics, we surveyed our alumni to provide some insight into the answer to this question. Our findings may prove interesting to students interested in job opportunities for bachelor's degree statisticians and to faculty who teach in these programs.

Background

Eastern Kentucky University (EKU) is a regional four-year institution with a current enrollment of 14,762 students. The Department of Mathematics and Statistics at EKU offers the only B.S. in Statistics in Kentucky. Figure 1 displays the number of graduates completing the requirements for a B.S. in Statistics from 1982 to 2001. Some of these students earned a second major in another field.

Patti Costello is an Associate Professor at Eastern Kentucky University, where she has taught statistics since 1982. She loves to tell students in Freshmen Orientation that some of them are Statistics majors—they just don't know it yet! She is also interested in international adoption and has adopted two children from South Korea. Her e-mail address is Patti.Costello@eku.edu.

Lisa Kay is an Assistant Professor at Eastern Kentucky University. She was an undergraduate at EKU and is one of the two Assistant Professors with Ph.D.s mentioned in the article's Graduate School section. Her three daughters and two cats keep her very busy. Her e-mail address is Lisa.Kay@eku.edu.

Figure 1. Number of Graduates of EKU's B.S. Statistics Program from 1982 to 2001

Placing the numbers for the department in a national context is illuminating. Very few schools that offer B.S. degrees in Statistics have large enrollments. Bryce, Gould, Notz, and Peck (2001) presented the data in Table 1, which led them to conclude that "Most undergraduate programs in statistics are small."

Table 1. Bachelor's Degrees in Statistics Awarded in 1996–1997

Number of Degrees Awarded	Number of Institutions
Less than 5	37
5–9	26
10–14	5
15–19	6
20 or more	4

Most students who earn the B.S. degree in Statistics at EKU take 24 credit hours comprised of the following statistics courses: Applied Statistics I and II, Applied Probability, Sampling Theory, Nonparametric Statistics, Statistical Methods Using SAS, and Mathematical Statistics I and II. In addition, the students are required to

take 12 hours of calculus, 3 hours of linear algebra and matrices, 3 hours of a programming language, and 3 hours of an upper-division mathematics or computer science elective, for a total of 45 credit hours. We encourage students to take courses or to minor in other fields such as insurance, business, manufacturing and technology, and biology. This agrees with a recommendation from the American Statistical Association's Curriculum guidelines for Undergraduate Programs in Statistical Science, given at *http://www.amstat.org/education/Curriculum_Guidelines.html*, which states, "Because statistics is a methodological discipline, statistics programs should include some depth in an area of application." We are also making some changes, such as the addition of an undergraduate course in experimental design, in order to more closely follow these guidelines.

Alumni Survey

During the 2000–2001 academic year, the B.S. Statistics degree at EKU underwent program review. As part of that review, we conducted a survey of alumni. On April 6, 2000, we sent a letter to all of the EKU alumni who had graduated since December 1982. The letter described the program review and provided some personal information about our current Statistics faculty. Enclosed was a survey that asked each graduate to provide information about current and previous jobs, information about any graduate work completed, opinion of the strengths of the statistics program, and suggestions for improvement of the program. In January 2001 and in February 2001, we sent a follow-up letter and survey to those alumni who had not yet responded. Of the 92 alumni who graduated between December 1982 and August 2000, 58 returned the survey. (There were 18 for whom no current address was known. Thus, from the mailing to 74 alumni, the response rate was 78%.) A complete listing of our alumni's current jobs, previous jobs, and graduate school experience is given at *http://www.stats.eku.edu/statalums*.

It is important to remember that these results may not generalize to all alumni due to this voluntary response. In conducting our survey, we found it very useful to begin with the alumni relations office. We also did some detective work that involved contacting some of our alumni's parents and asking other alumni for their help in finding current addresses. It also helped that we had kept a list of names of our students who had graduated over the years. It was very rewarding to hear from alumni that we hadn't heard from in years.

Graduate School

The survey asked alumni if they had attended graduate school. If they had, they were asked for the name of the school, degree received, and graduation date.

Of the 58 alumni who returned the questionnaire, 20 of them (34.5%) have earned a graduate degree of some kind. Eleven of the alumni who returned the survey, or 19.0%, have earned either Master's degrees or Ph.D.s in Statistics. Two of these alumni with Ph.D.s are currently Assistant Professors. The remaining 15.5% who received graduate degrees earned Master's degrees in Education, Mathematical Sciences, Computer Science, or Business. (This is not surprising because some of these alumni had minors or second majors in these fields.) Five of the alumni are currently enrolled in graduate school, and one of these was expecting to receive his Ph.D. in Cognitive Systems Engineering in May 2001. (In addition, we are aware of three more EKU graduates who took many of the department's statistics courses, but who did not graduate as statistics majors, and still went on to earn Ph.D.s in Statistics.) Of the seven EKU statistics majors who received their B.S. degrees in May 2001, two are currently attending graduate school, one in statistics and one in mathematics.

Job Opportunities

Fewer than half of our graduates pursued graduate school, so one of the primary purposes of the survey was to find out if graduates who did not pursue graduate school were able to find jobs as statisticians. According to Ritter, Starbuck, and Hogg (2001), "There are very few positions exclusively for BS-Statisticians. More common are positions for which BS-Statisticians qualify, but for which statistics is only one of several appropriate types of preparation." The survey asked alumni to give their present employers, their job titles, and whether or not they use statistics in their positions. Similar information was requested for previous jobs.

Survey results indicate that our statistics alumni are able to find good jobs that enable them to use their statistical education. Of the 58 alumni who returned the survey, 45 (77.6%) said that they use statistics in their current jobs, while 49 (84.5%) said that they either use statistics in their current jobs or have used statistics in their previous jobs or both.

Many of the alumni work as SAS programmers. Graduates were asked, "What are the strengths of the BS degree in Statistics at EKU? What did you learn that has been the most beneficial to you?" Nineteen (32.8% of responding alumni) stated that the course STA 575, Statistical Methods Using SAS, was the most beneficial course that they took. One alumnus stated, "The SAS course (STA 575 I believe) has proven to be *the most valuable* course that I took while attending Eastern. I can think of 3 instances where I received the opportunity to interview for a position based upon my experience with that class." Another graduate responded, "Most beneficial was being able to obtain a job very quickly within the Lexington [Kentucky] area

using my 'SAS' skills that I learned at EKU."

STA 575, Statistical Methods Using SAS, has proven to be a valuable learning tool for alumni who want to work as SAS programmers. The first half of the course covers many of the topics tested in the SAS Core Certification, Level 1 exam. (For more information on this exam, see *www.sas/com/service/edu/certify/intro.html*.) In particular, we teach our students how to access data files, work with SAS data sets, and manage data. By the middle of the course, they have a pretty good understanding of the SAS Data step and how to handle SAS programming errors. We try to use "real" data as much as possible so the students can learn that "real" data is usually very "messy." The second half of the course focuses on various SAS procedures that are used to analyze data. Many employers who hire SAS programmers want to hire people who have previous SAS experience. Taking this course gives our students SAS experience and qualifies them for entry-level SAS programming jobs.

EKU Statistics alumni also use their statistical education to work as biostatisticians, actuarial analysts, quality assurance engineers, survey statisticians, forecast analysts, marketing analysts, and teachers at both the high school and college levels. Some are employed at government agencies such as NIOSH and the Census Bureau, while others work at contract research organizations (CROs) such as Kendle International, StatProbe Inc, and Pharmaceutical Outcomes Research Inc. According to Kogut (1997), "There is a growing trend for many statistical consulting jobs to be offered via contract research organizations (CROs), especially in the pharmaceutical industry. . . . CRO employment can provide intensive and varied work experience for the statistician beginning a career." One of the alumni stated that "between ClinTrials Research & StatProbe, we have hired approximately 10 Stats graduates from EKU. We have had great success with the students that we have hired. . . . The Pharmaceutical/BioTech and Contract Research organizations are always in need for these types of individuals." Some of the alumni work directly for pharmaceutical companies, such as Eli Lilly and Bristol-Myers Squibb Company. Several EKU alumni have also worked at Procter and Gamble Company.

EKU has two alumni who work at Toyota Motor Company in Georgetown, Kentucky, as well as alumni who perform actuarial work for Anthem Blue Cross and Blue Shield and Ohio Casualty Insurance Company. One graduate who works as an actuary commented, "If I didn't choose the statistics major, I would not have been able to pass Part 2 of the Casualty Actuary Society exams in a timely manner. This exam focused exclusively on statistics." A number of our alumni are now managers.

At least ten of our alumni (17.2% of those who filled out the survey) teach at the middle school, high school, or college level. Their statistics training has been invaluable to them as well. One alumnus who teaches at a high school stated,

> Three years ago, I was asked to develop a Probability & Statistics course for seniors. I have doubled the enrollment in 3 years. By having my Statistics degree, in addition to Mathematics Teaching, it has made it easier for me to implement the Statistics strand of the KY Core Content into my classroom teaching. It also has allowed me to serve as an 'informal' reference person within my department.

As more statistics is included in the pre-college curriculum, the need for teachers with statistical skills will become even greater.

Importance of Statistical Skills

Another indication that graduates of a bachelor's degree statistics program gain skills that lead to career opportunities is that many of our alumni contact the department when their employers have job openings. As one alumnus noted, "Employers were glad I had SAS experience, but seemed more impressed by the Statistics Degree. There are very few programmers with both SAS experience and Statistics in the workplace."

Other comments gathered from the alumni surveys also indicate a need for graduates with statistical skills:

> "The pharmaceutical industry is in need of people with the skill sets of statistics graduates. There are many jobs available for students with a BS degree in Statistics, and it appears to keep growing with all of the data being collected by companies."

> "SAS jobs are everywhere and there have never been enough people to fill all the positions at any of the places I've worked in."

> "Entry-level jobs at [company's name] are hard to fill. We could probably hire all of EKU's graduates."

> "The BS degree in Statistics at EKU has given me a solid foundation on which to build my career. It has given me so many open doors to choose from."

> "Without my statistics background, I wouldn't be where I am today earning top $$$. I love my job and each project is a new challenge."

The *Occupational Outlook Handbook* from the Bureau of Labor Statistics at *http://stats.bls.gov/oco/ ocos045.htm*

suggests, "Job opportunities should remain favorable for individuals with statistical degrees, although many of these positions will not carry an explicit job title of statistician." (For more information about a wide variety of career possibilities, see *http://www.amstat.org/careers* and the September 2001 issue of *Amstat News*, which contains many useful articles.) While there may not be many jobs that actually bear the title *statistician* available to B.S. statisticians, many employers are finding such graduates quite prepared to do statistical work. Some employers wrote letters of support for our program review. One Kentucky employer, who prior to retirement ran his own statistical consulting company, stated, "Many of my first professional employees were EKU graduates of your undergraduate statistics program. To be honest, I was pleasantly surprised at the high caliber of these graduates." Another letter of support from the Head of the Statistics Section at a company that employs our graduates included the following statements:

> It is very difficult for our department to recruit and hire employees who have a BS degree in statistics. This is because there are not many universities in the US that offer this degree. And for those universities that do offer the degree, very few students graduate with that degree each year. Often we hire students with [a] BS in mathematics and we train them but this is certainly less than optimal for us.

Conclusion

An undergraduate degree in statistics can prepare students well for graduate school in statistics as well as other fields. EKU alumni have been successful in graduate school in several areas.

There is also a good job market for statistics graduates. According to the web site of the Bureau of Labor Statistics at *http://stats.bls.gov/oco/ocos045.htm*, "More employment opportunities are becoming available to well qualified statisticians with bachelor's degrees." While a position with the actual title of statistician might require a graduate degree, there are many job possibilities for those with degrees in statistics that do not carry the title of statistician.

The results of our alumni survey and research for our program review point to several suggestions for bachelor's level statisticians who are looking for career opportunities:

- There are lots of job opportunities for B.S. statisticians, but people seeking these jobs need to look beyond just those jobs bearing the statistician title.
- SAS programming experience is very valuable and can create many job possibilities.
- A second major or a minor in another field is an asset to B.S. statisticians.
- Reaching out to alumni from one's alma mater can lead to more job opportunities.

Acknowledgment

The authors would like to acknowledge the contributions of Onecia Gibson and Ron Pierce to the program review document, which provided some material for this article.

References

Bryce, G. R., Gould, R., Notz, W.I., and Peck, R. L. (2001), "Curriculum Guidelines for Bachelor of Science Degrees in Statistical Science," *The American Statistician*, 55(1), 14-18.

Kogut, S. (1997), "Statistical Consulting at a Contract Research Organization," *The Statistical Consultant*, 14(2), 2-6.

Ritter, M.A., Starbuck, R. and Hogg, R.V. (2001), "Advice from Prospective Employers on Training BS Statisticians," *The American Statistician*, 55(1), 213.

Scheaffer, R. (2001), "In a World of Data, Statisticians Count," *Amstat News*, 291, 2-3.

Web Resources

American Statistical Association Career Center: *http://www.amstat.org/careers*

American Statistical Association Curriculum Guidelines for Undergraduate Programs in Statistical Science: *http://www.amstat.org/education/Curriculum_Guidelines.html*

Occupational Outlook Handbook: *http://stats.bls.gov/oco/ocos045.htm*

SAS Institute: *www.sas/com/service/edu/certify/ intro.html*

Results of EKU Statistics Alumni Survey: *http:// www.stats.eku.edu/statalums*

The Statistical Consultant: *www.amstat.org/sections/ cnsl/*

AP Statistics Turns 5!

A Report on the 2001 Exam: Questions, Performance, and

Roxy Peck

AP Statistics enjoyed its fifth birthday in June of 2001, and there was much to celebrate, including continued growth and improved student performance. Approximately 42,000 students took the exam last spring, and with an estimated 71,000 students enrolled in AP Statistics around the country last fall, we are preparing for about 49,000 exams in 2002.

The AP exam challenges students to demonstrate their ability to apply appropriate statistical methods as well as sound statistical reasoning and clear communication. The exam consists of a multiple-choice section and a free response section. The free response section is made up of five open-ended questions and one longer investigative task that requires integration and synthesis of multiple concepts.

Two hundred university and high school statistics teachers graded the free response section of the exam during a weeklong reading that was held during June 2001 at the University of Nebraska. Each free response question was assigned a score from 0 to 4 according to holistic rubrics. For those interested, the free response questions and the scoring rubrics can be found at APCentral, a web site for teachers, at *www.apcentral.collegeboard.com* (visitors to this site must register, but there is no charge for registration).

The free response questions in 2001 covered topics from the four basic content areas of the AP Statistics course description: sampling and experimental design, descriptive methods, inferential methods, and probability. A brief description of the content of each question and the corresponding expectations for student responses appears below, followed by a discussion of one of the largest difficulties students still face on the exam: communication.

Roxy Peck is Chief Faculty Consultant for the Advanced Placement Statistics Program. She is Associate Dean of the College of Science and Mathematics and Professor of Statistics at Cal Poly, San Luis Obispo.

The 2001 Free Response Questions

Question 1 assessed the student's understanding of numerical measures of center and spread and the concept of outliers. To receive full credit for this problem, the student was expected to describe a plausible procedure using summary statistics to check for outliers, to use that procedure to check for outliers in a data set for which summary statistics are given, and to comment on a statement from a newspaper report concerning the relative position of one particular observation.

Question 2 evaluated the student's ability to use information provided in the form of a probability distribution to make a recommendation on which of two brands of copy machines a company should purchase. Several different approaches to this problem were reasonable—an approach based on comparing the expected cost of machine B to the fixed cost of machine A, an approach based on computing the probability that the cost of machine B would exceed the fixed cost of machine A, or an approach based on simulation.

Question 3 evaluated the student's ability to design and carry out a simulation to estimate the distribution of the number of prizewinners in a weekly radio contest. To receive full credit on this question, the student was expected to describe how he or she would carry out the simulation and then conduct three trials using the given table of random numbers.

Question 4 assessed the student's understanding of some basic principles of experimental design, including randomization and blocking. Students were asked to explain the purpose of these strategies relative to a particular context.

Question 5 evaluated whether the student could carry out a test of significance and state conclusions in context. To receive full credit on this question, the student was expected to state hypotheses, identify an appropriate test procedure, check any necessary assumptions, compute the value of the test statistic and

the associated p-value (or rejection region), and then, based on the result of the test, give an appropriate conclusion in context.

Question 6 was the exam's investigative task. As such, its purpose was to evaluate the student's understanding in several course topic areas and to assess ability to integrate statistical ideas and apply them in a novel way. This year's investigative task involved using graphical displays to compare two groups, inference about the slope of a regression line, and using bivariate data given for each of two groups to reason about the group membership of a new data point. It was a very rich problem with a number of different possible reasonable approaches to the classification issue.

Exam Performance

Overall student performance improved compared to the past two years, with higher scores on the multiple-choice section of the exam and more consistent performance across the six questions of the free response section. Composite scores (multiple choice + free response) are used to determine the score reported to the student (5, 4, 3, 2, or 1). Table 1 gives the distribution of reported scores for the five years that the exam has been given. The increased percentage of 5's and 4's this year reflects the improved student performance on this year's exam.

Table 1. Reported Score (Percentage) Distribution

Reported Score	1997	1998	1999	2000	2001
5	15.7	13.7	11.1	9.7	11.4
4	22.1	21.4	20.3	21.6	23.4
3	24.4	24.6	25.8	22.4	24.9
2	19.7	18.6	20.9	20.6	19.1
1	18.0	21.8	21.9	25.7	21.2

Exam performance this year (and in past years) was strongest in the area of describing data and weakest in the area of statistical inference. This was apparent in both the free response inference questions as well as in the multiple-choice questions dealing with inference. In general, students were much stronger on the mechanical and computational aspects of problems than on parts that required interpretation or conceptual understanding.

Importance of Communication

In particular, communication of results continues to be a weakness. It is important for students to realize that communication plays a crucial role in statistical analysis. In fact, interpretation of results and the relationship of the calculations to the particular context is a large part of what distinguishes the discipline of statistics from the discipline of mathematics (see, e.g., Cobb and Moore, 1997). In statistics, meaning comes from context, and it is the interpretation of the analysis in context that is the ultimate desired outcome.

A look at how the free response questions on the AP Statistics exam are scored shows that communication is weighted equally with statistical knowledge. (For more detail on how communication is rated, see the *Teachers' Guide* (Watkins, Roberts, Olsen, and Scheaffer, 1997).) Thus, it is not possible to achieve top scores on these questions if communication is weak. While the *Teachers' Guide* provides general guidelines for five different levels of communication, below are some concrete examples of student responses, taken from Peck (2002).

Sample 1: Poor Communication

"Matrix used on calc. $X^2 = .51686$, $p = .99996$ results in an insignificant p-value. No effective conclusion can be made that there is an association between the active ingredient in the 2 brands of pills and the pharmacy." (2001 exam, Question 5)

Comments: This student may know something, but you can't really tell it from this response! The student has chosen an inappropriate test, with no indication of the rationale behind the choice. Further, the results and conclusions for the test that was performed are incomplete and poorly communicated, resulting in a very low score.

Sample 2: Weak Communication

"It isn't an outlier because it is within the IQR." (2001 exam, Questions 1c)

Comments: This response isn't worded correctly (within IQR doesn't make sense since the IQR is a single number) but we can sort of tell what the student means by this comment. This was considered a minimal response.

Sample 3: Adequate Communication

"10 inches of rainfall is not outstanding at all. The mean is 14.941 with a standard deviation 6.747. That implies that 67% of the data is between 14.941 ± 6.747. A z score can be used to determine the exact probability." (Calculations followed) (2001 exam, Question 1c)

Comments: This reasoning is incorrect (the distribution was not normal), but it is clear what the student is doing, and the response shows some understanding. This was considered a developing response.

Sample 4: Good Communication

"Since we want at least a pain relief of 50, drug B at 400 milligrams would be better than any of the strengths of drug A because the plot is of averages. For drug A, some patients might have had no pain relief and other 100—so it averaged 50. But drug B at 400 milligrams shows that most of the time the relief would be above 50 therefore bringing the average up to 90." (2000 exam, Question 1c)

Comments: This isn't the answer that was expected, but the idea is well communicated. The student shows understanding of the distinction between averages and individual measurements and clearly understands the information provided by the graph. This response received a good score for this problem, even though the scoring rubric was looking for an answer that justified the choice of drug A at a low dose.

Conclusions

The AP Statistics program continues to thrive and grow. This year's encouraging news is the observed improvement in student performance over the past two years. Communication is an area where students still struggle. It is important for students to understand the necessity of communication skills to the discipline of statistics and that these skills can also improve with frequent practice. As AP Statistics becomes more established in our high schools and as students gain confidence and experience, we look forward to continued improvement in the years to come.

Acknowledgment

Parts of this paper are taken from the article "Calculations Aren't Enough! The Importance of Communication in AP Statistics," copyright © 2002 by collegeboard.com, Inc. Reproduced with permission. All rights reserved. *www.collegeboard.com*.

References

Cobb, G. W. & Moore, D. S. (1997), "Mathematics, Statistics, and Teaching," *American Mathematical Monthly*, 104, 801–823.

Peck, R. (2002), "Calculations Aren't Enough! The Importance of Communication in AP Statistics," APCentral (*www.apcentral.collegeboard.com*).

Watkins, A., Roberts, R., Olsen, C. & Scheaffer, R. (1997), *Teachers Guide for AP Statistics*, New York: The College Board.

Developing Good Communication Skills

What can you do to learn to communicate more effectively when writing about statistics?

- Pay attention to your explanations and interpretations throughout the course. Communication is an important component in early course topics as well as when you are learning about inferential methods.
- Never give answers that are "mechanics only." Always provide a final conclusion in "every day language." This should be understandable to an audience who is not familiar with statistics or with the technology you used to perform the analysis.
- In particular, never rely on "calculator talk" as an explanation of what is being done: "I entered x:20 n:400 and .95 on my TI-83. 1 prop z-int printout is .02864, .07136" (a sample student response on question 6a of the 2000 exam) is NOT a good explanation of what is being done and why!
- Similarly, don't limit your response to just statistical terminology. For example, don't stop after saying "reject the null hypothesis." Go on to explain what this conclusion means in the context of the data and question provided.
- Reading carefully is just as important as writing carefully. Use the discussions in your textbook as a model of how to communicate statistical ideas. Reading problems carefully is important too, especially in determining exactly what type of analysis is appropriate.
- Practice writing about statistical concepts and processes as well as about analyses. Many students have difficulty describing processes—how to tell if there are outliers in a data set, how to carry out a simulation, etc. Try doing this in general terms, beyond a particular example.

Student Projects

Oh My Aching Back!

A Statistical Analysis of Backpack Weights

Jenni Mintz, Jessica Mintz, Katrina Moore, Kim Schuh

In high school, the requirement of bringing a textbook to class daily caused many complaints to erupt. Students always complained about having bad backs. We wanted to know if college students also suffered from the problems of carrying too many books in their backpacks, so we decided to do an observational study to answer some questions. We read that previous research (e.g., *www.sw.org/news/options/november/backpack_study*) suggested that a student should carry no more than 10–15% of his or her body weight. We decided to see if on average Cal Poly students carry less than 10% of their body weights in their backpacks.

Summary of Data Collection

The population of interest was all Cal Poly students, and we decided to get a sample of 100 students. Our sampling frame included students from The Avenue (a campus dining facility), The Park (a campus fast-food facility), the University Union, and the Reserve Room at the Kennedy Library. Twenty-five students were sampled from each location, during the course of four days and four different times of the day: morning, lunch time, late afternoon, and evening.

Our sampling technique involved the four of us meeting at a location. One of us was a designated talker who introduced the survey to the participants. A second person read the scale, and a third gathered the surveys. We used the scales to measure backpack weight to the nearest pound, and we asked students to report their body weight on the survey. We continued until 25 surveys were taken for each location. We tried to reduce sampling bias by asking people randomly and did not allow extraneous variables to get in the way. Even if people looked really busy or antisocial, we still asked them to get as much of a true representation as possible.

One possible remaining source of bias is the small detail that people can lie about their weight, thus making the correlation between how much weight they are carrying in their backpack and their weight incorrect. However, because the questionnaire was anonymous, hopefully the participants did not do this. There were also a few people (five) who declined to participate in the survey. It is possible that this could have biased the results because those who declined may have had more in their backpacks.

We believe the sample was representative of the entire Cal Poly student population. We received numerous survey results from various years and both sexes. The variety in locations, days, and times of the day lets us assume that the conducted survey was done in a representative way.

Summary and Interpretation of Data

Figures 1-3 present a histogram, boxplot, and descriptive statistics of the ratios of backpack weight to body weight.

Jenni Mintz is a journalism sophomore, Jessica Mintz is a psychology freshman, Katrina Moore is a child development junior, and Kim Schuh is a graphic communications freshman at Cal Poly in San Luis Obispo, California.

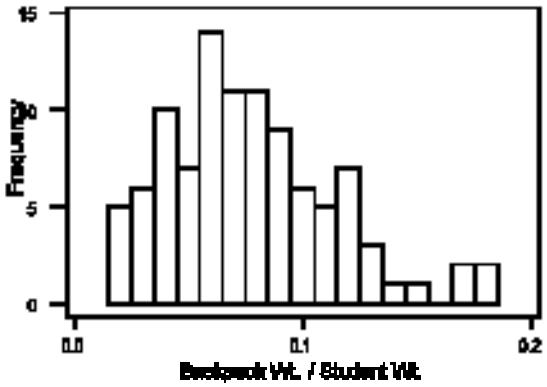


Figure 1. Histogram of Backpack Weight to Body Weight Ratios

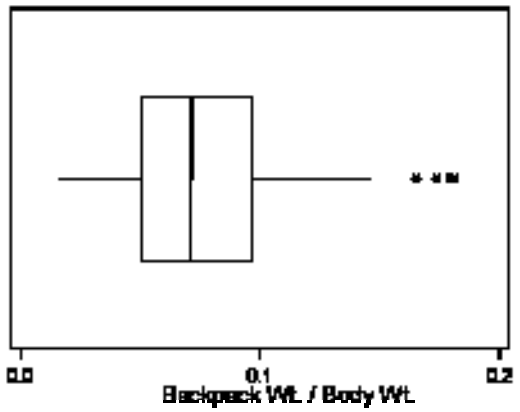


Figure 2. Boxplot of Backpack Weight to Body Weight Ratios

The graphs of our quantitative data show much spread with numerous peaks and valleys throughout. The data appear to be skewed to the right with four high outliers who carry more than 16% of their body weight in their backpacks. The mean of the ratios is 0.077 and the median is 0.071. The range of our data is 0.165, standard deviation is 0.037, and interquartile range is 0.046.

Statistical Inference

To answer our research question, "Does the typical Cal Poly student carry less than 10% of his or her body weight in his or her backpack?" we used the one sample *t* procedure. Let μ equal the mean ratio between the weight of the backpack and the student's weight. Then our hypotheses are

H_0: $\mu = .10$ The mean ratio between the weight of backpack and student's weight equals .10.

H_a: $\mu < .10$ The mean ratio between the weight of backpack and student's weight is less than .10.

Note that we considered our observations to be a simple random sample and $n \geq 30$, so the *t* test is valid. Using our output from Minitab, we found the test statistic equals –6.24 and the p-value is < 0.001. With such a small p-value, we consider our results to be statistically significant, meaning that getting a sample mean as small as ours would not likely happen by chance if the population mean was .10. Therefore, we reject the null hypothesis and support the alternative hypothesis that the mean ratio between the weight of backpack and student's weight is less than .10.

We can find a confidence interval to estimate the mean ratio of backpack weight to body weight in the population. A 95% confidence interval for μ is (0.0699, 0.0844). We are 95% confident that μ is between 0.0699 and 0.0844.

Alternatively, we could also count up the number of students who carry less than 10% (those will be our successes) and we can find a confidence interval to see what proportion *p* of the Cal Poly population carries less than ten percent of their body weight in their backpacks. Since we have at least 5 successes and 5 failures, we used a one-sample z interval. The sample proportion is $\hat{p} = .76$ and the 95% confidence interval for *p* is (.676, .844). We are 95% confident that between .676 and .844 of all Cal Poly students carry less than 10% of their body weights in their backpacks.

Interpretation/Explanation of Results

We found that the mean weight carried by Cal Poly students tends to be less than .10 of their body weight. In the *t*-test of significance we found the p-value to be quite small, which provided significant evidence for the alternative hypothesis. We also found with a 95% confidence level that between .676 and .844 of all Cal Poly students carry less than 10% of their body weights in their backpacks.

Reasons for what we observed include the following: Some people take lighter loads on certain days and perhaps, by chance, we surveyed those individuals on

Figure 3. Descriptive Statistics of Backpack Weight to Body Weight Ratios

Variable	n	mean	std dev	Minimum	Q1	median	Q3	Maximum
ratios	100	0.077	0.071	0.016	0.050	0.037	0.096	0.181

During our collaboration on the backpack analysis project, we were struck by the sheer amount of time it took to collect the data, do our analysis, and draw conclusions. This process has given us a new appreciation for various experiments and observational studies since we've worked "behind the scenes."

The most interesting part of our project was by far the data collection. Many students complied quickly and expressed interest in our study, hoping to see the results published in the college

their particularly light day. The ten percent statistic came from an article about high school teens carrying too much weight around campus. Since college students are more independent and most classes are lecture-based, the need to bring the book diminishes. Also, college students may be more apt to put on weight, so even if they were carrying the same amount of weight in their backpack (which is improbable), the proportion would be less since their own personal weight has increased.

Recommendations

We recommend that any student carrying too much weight in his or her backpack should follow these guidelines (courtesy of www.*gobroomecounty.com/press/082701h.html*):

- Use a hip strap for heavier weights.
- Use a backpack with padded, wide straps and a padded back.
- Use both of the backpack's straps, firmly tightened, to hold the backpack 2 inches above your waist.
- Engage in exercises to better condition your back muscles.
- Use correct lifting techniques. Bend with both knees when picking up a heavy backpack.
- Neatly pack your backpack and try to keep the items in place.
- Try to make frequent trips to your locker to remove items from your backpack you do not need.
- Consider purchasing a backpack with wheels.

Future Questions

How does the weight of the backpack itself affect the total weight carried by the students? If we limited our sampling frame to high school students, would the data be significantly different? Would the week in the quarter influence the amount of weight contained in a student's backpack? Do males or females tend to carry more weight in their backpacks? How does time of day or location on campus affect the contents of a backpack?

Editors' Note: We have made these students' data, including more variables than are discussed in this article, available at the STATS website:
http://www.amstat.org/publications/stats/data.html

Data Sleuth

This feature invites you to solve mysteries involving data. We hope that you will find this feature to be fun and enlightening, and we encourage you to send us your own submissions of data mysteries.

Mystery 1: Kentucky Derby Debates

Contributed by Patti B. Collings, Brigham Young University

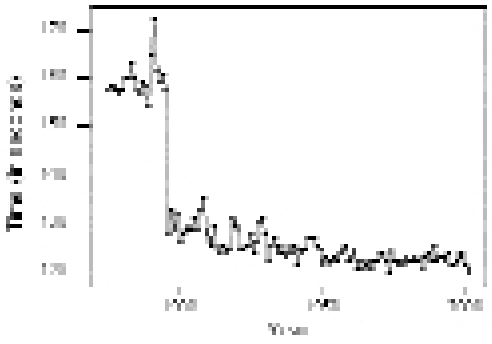
The Kentucky Derby is an annual horse race held since 1875. Figure 1 displays a time plot of the winning times (in seconds) in the Kentucky Derby by year, and Figure 2 displays the average speed (in seconds per mile) of the winning horse each year.

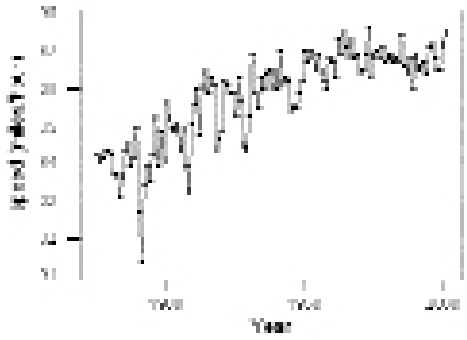
Question 1: In both plots, what is the overall trend in the data? How do you account for this?

Question 2: In both plots, why do you think there is so much variability from one year to the next?

Question 3: What unusual characteristic do you observe in the overall trend for the winning times? Does the same characteristic appear in the winning speeds? How might you explain this?

(The solutions appear on p. 25.)





Mystery 2: Who's Going to School?

Contributed by Patti B. Collings, Brigham Young University

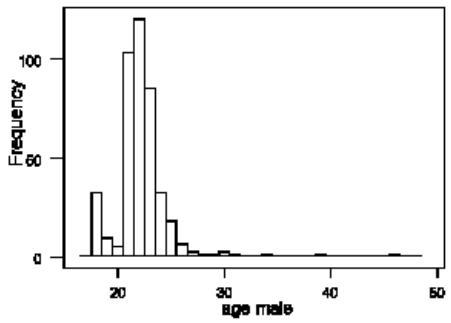
The histograms below are ages of male and female students taking introductory statistics at Brigham Young University during a recent semester. Brigham Young University is a private, church-owned university in Provo, Utah.

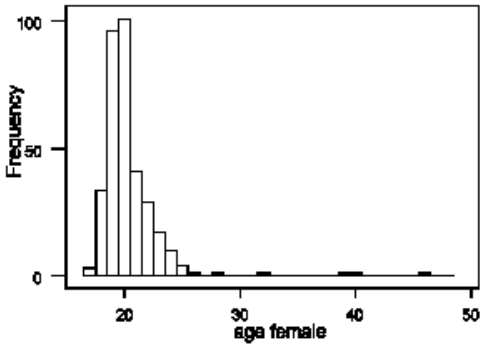
Question 1: What comparisons would you draw between these two distributions?

Question 2: Why do you think the standard deviation of the males (2.46 years) is less than that of the females (2.69 years) while, the average age of the males (22.14 years) is higher than the females (20.38 years)?

Question 3: Compare the shapes of the two histograms. What unusual characteristic do you observe? What explanation can you offer for the difference in the two histograms?

(The solutions appear on p. 25.)

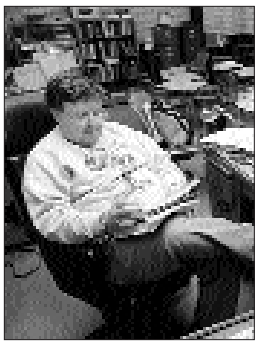




μ-sings

Statistics Goes to the Movies

Magic Town



Chris Olsen

OK, so let's consider our recent movie-going. Who have been the best characters? Judging from the buzz, two of the best must be Gladiator Russell Crowe on his impossible (but successful) quest to impose mortality on the Roman Emperor, and Russell Crowe as Mathematician/ Economist John Nash on his quest to solve a problem that nobody else has even thought of.

Apparently, these Holy Grail quest movies are in. Now, upon reflection doesn't it seem to you that these characters have the stellar qualities of statisticians? Rugged and indomitable characters, faced with events not of their making and beyond experimental control, yet proceeding undaunted and unbowed?

Wouldn't it be something if there were a movie about a statistical-type person as the undaunted hero or heroine? Possibly Jodie Foster as the pioneering statistician who solves the Behrens-Fisher problem, or Harrison Ford putting the final touches on the formal logic underlying Fisher's theory of fiducial probability? Well, it turns out that there is such a movie — *Magic Town*, starring James Stewart and Jane Wyman, which hit the streets in 1947. (The home video took a little longer…)

Stewart's character is a public opinion expert, and at the beginning of the movie he is searching for his personal holy grail — the perfect sample. Just think, he opines; if one could find The Representative Sample, one could spend a lot less time and effort and $$$ attempting to gauge the public's desires and opinions. This, of course, would turn into big profits for his polling agency. As luck and the script writers of Hollywood would have it, Stewart stumbles onto a gold mine when comparing national election results to some polls taken in a town called Grandview, state unknown, but suspiciously Midwestern in character. The national percentages exactly match the percentages for Grandview, in characteristic after characteristic! Now, I know this sounds pretty suspicious to post-Watergate statisticians, and one must admit that sampling error does seem to have passed Grandview by. However, if you are comfortable suspending reality enough to go to the *Harry Potter* and *Lord of the Rings* movies, you can certainly give this one a little break too.

Now, where was I? Oh, yes — the perfect sample. Stewart recognizes this golden opportunity, and he and his minions sneak into Grandview under cover, posing as insurance salesmen. (See? Statistics everywhere!) This ruse is necessary, of course, so that the investigator doesn't alter the situation he is observing as he observes it. However, Jane Wyman's character, the headstrong editor of the local paper, is going to present some problems. She is of the opinion that the town needs to get out of its rut and attract some industry. This "rut" is the one that seems rather wonderful — and profitable — to Stewart, and pollster/insurance man and newspaper editor quickly stake out opposite directions for the town.

The plot, as they say, then thickens. Your faithful reviewer must note that after starting out at a very high level of audience excitement generated by the discovery of The Representative Sample and the attempts to keep the townsfolk naïve, the movie degenerates into subplots having to do more with the ethics of deception, and of course the burgeoning romantic relationship between the two protagonists. This reviewer would have liked to see the sampling issues developed further, possibly with Stewart and Wyman uncovering a manipulative evil genius at work, played by Vincent Price (a sort of 40's Anthony Hopkins), thus explaining the lack of sampling error. However, the makers of the film saw it differently, and Stewart falls irrevocably in love with both town and editor. Romance wins again; by the end of the movie everybody who deserves one gets a happy ending.

The home video may be difficult to find, but the search is well worth it. *Magic Town* is a fast moving (for the 40's) film, and a cheap (for the 00's) date. Sample it at your earliest convenience, preferably with a significant other who wonders just what it is that statisticians do.

[*Magic Town*: Republic Pictures Home Video: Run time

The Statistical Sports Fan

Offense and Defense as Predictors of Team Success

Robin Lock

In the 1999 National Football League season, the St. Louis Rams led the league in scoring (32.9 points per game) and went on to win the Super Bowl Championship. The next year, the Baltimore Ravens set an NFL record for the fewest points allowed in a 16 game season (10.3 points per game) and also won the Super Bowl. These accomplishments illustrate the extremes of a long-standing debate in many sports over the relative importance of offense vs. defense in producing a winning team. In this article, we examine data from a recent season for professional teams in football (NFL), basketball (NBA), hockey (NHL) and baseball (MLB) to investigate how a team's winning percentage is related to its offensive ability (points scored) and defensive ability (points allowed).

The Data

We use data from all teams for the 2001 regular season (which started in 2000 for the NHL and NBA). Table 1 displays a subset of the data for teams in one division of each sport. Offensive production is measured by the average points/goals/runs scored by each team per game and the defensive performance is the average points/goals/runs allowed per game. The key response variable (winning pct.) is the number of wins divided by the number of games played (x 100). The only tie games occurred in the NHL (15% of the games) and are counted as one-half a win in computing the winning pct. This deviates slightly from the NHL's new scoring system that awards teams a point for a tie game or losing a game in overtime (and two points for a win). We give no credit for overtime losses in order to preserve 50% as the average winning percentage. Figure 1 shows dotplots of the winning percentages for all teams in each sport. Note that the most variability (and granularity) occurs in the NFL, which has only 16 games in the regular season, while the least variability occurs in MLB with its 162 game schedule. The NBA and NHL both play 82 game schedules for their regular seasons. The data for all four leagues may be obtained from the STATS website *www.amstat.org/ publications/stats/data.html*.

Table 2 shows the correlations in each league between the winning percentages and both the offensive and defensive scoring rates. For the NFL and MLB we see that the correlation with winning percentage is stronger for Defense than for Offense, whereas in the MBA and NHL the Offense has a slightly stronger correlation with winning percentage. But are any of these differences statistically significant? We'll investigate that question by considering two different models for predicting winning percentage using the offensive and defensive scoring rates.

Table 1: Scoring (Per Game) and Winning Pct. for a Sample of Teams

NFL-AFC East	NFLOff	NFLDef	NFLPct
New England	23.2	17.0	68.8
Miami	21.5	18.1	68.8
NY Jets	19.3	18.4	62.5
Indianapolis	25.8	30.4	37.5
Buffalo	16.6	26.3	18.8
NBA-Pacific	**NBAOff**	**NBADef**	**NBAPct**
LA Lakers	100.6	97.2	68.3
Sacramento	101.7	95.9	67.1
Phoenix	94.0	91.8	62.2
Portland	95.4	91.2	61.0
Seattle	97.3	97.3	53.7
LA Clippers	92.5	95.3	37.8
Golden State	92.5	101.5	20.7
NHL-Northwest	**NHLOff**	**NHLDef**	**NHLPct**
Colorado	3.29	2.34	69.5
Edmonton	2.96	2.71	54.9
Vancouver	2.91	2.90	50.6
Calgary	2.40	2.88	42.1
Minnesota	2.05	2.56	38.4
MLB-NL West	**MLBOff**	**MLBDef**	**MLBPct**
Arizona	5.05	4.18	56.8
San Francisco	4.93	4.62	55.6
Los Angeles	4.68	4.59	53.1
San Diego	4.87	5.01	48.8
Colorado	5.70	5.59	45.1

Model #1:
Winning Pct = $\beta_o + \beta_1$ Offense + β_2 Defense + ε

Results for each league of a multiple linear regression fit using two predictors (points scored per game and points allowed per game) are summarized in Table 3. If you'd like to predict the winning percentage for your favorite team, just multiply their current average points scored and points allowed (per game) by the respective coefficients and add on the intercept.

Not surprisingly since its 16-game season is much shorter than the other leagues, the NFL has the weakest percentage of variability in its winning percentages explained by the model. The signs and magnitudes of the coefficients are consistent with the correlations of Table 2 – slightly more weight given to the defensive statistics in the NFL and MLB while the reverse is true in the NBA and NHL. We should take care when interpreting these individual coefficients since the offensive and defensive variables may be correlated themselves. Interestingly, though, none of the sports show a statistically significant correlation in their samples between the offensive and defensive scoring variables. In fact, three leagues (NHL, r = -0.26, NFL, r = -0.23 and MLB, r = -0.11) show small negative sample correlations between these variables, while the NBA shows a positive sample correlation (r = 0.28). We will leave you to ponder whether this uniqueness of the NBA is related to the fact that both individual correlations are the weakest for the NBA, but the R^2 for the combined model is strongest for the NBA.

Figure 1: Winning

Model #2:
Winning Pct = $\beta_o + \beta_1$ (Offense fi Defense) + ε

For individual games, the winner is always determined precisely by the difference between the number of points a team scores and the number of points it allows. If we apply this reasoning to a season's worth of data, we might suggest using the *difference* between average points scored (per game) and average points allowed (per game) as a single predictor of a team's winning percentage. Table 4 summarizes the results of fitting this model as a simple linear regression to the data from each of the four sports. Scatterplots (with regression lines) showing these linear relationships are displayed in Figure 2. Note that the intercept is 50.0 in each case since the average offense must equal the average defense, hence the average difference must be zero and the intercept will be the average winning percentage. From the R^2 values we see that these models are nearly as effective as the more complicated versions in Model #1 and only require remembering a single parameter to apply.

Comparing Correlations

Returning to our original question, how might one determine if the offensive or defensive variable is a significantly better predictor of winning percentage? One approach involves comparing these two models. We may view Model #2 as a special case of Model #1 where the coefficients of the two predictors are constrained to be equal. This constrained model would be consistent with a null hypothesis that states that each predictor is equally important in the model (i.e., $\beta_2 = -\beta_1$ in the notation of that model). We would have evidence against that null hypothesis (and in favor of an alternative that more weight should be given to one of the predictors over the other) if the performance of the unconstrained model (Model #1) was significantly better than that of the constrained model (Model #2). Note that Model #1, with an extra free parameter, will always do somewhat better than Model #2 (if we measure a good fit via R^2 or the sum of squared errors), but we still need to ask whether that improvement is statistically significant. A quick comparison of the R^2 values between Table 3 and Table 4 should lead one to believe that very little predictive power is lost when we move to the more restricted model in any of the four sports. Thus we can informally conclude that neither offensive performance nor defensive ability is a substantially better predictor of a team's winning percentage.

To do the comparison formally, we can do a partial F-test to determine whether the sums of squares explained by the unconstrained two-predictor model is a significant improvement over what is explained by the single variable model. Since the reduced model has one fewer degree of freedom, we would obtain a test statistic by dividing this amount of change in the sum of squares by the mean square error for Model #1 and comparing the result to an F-distribution with 1 and $n - 3$ degrees of freedom, where n is the number of teams in the league. The F-test statistics and p-values for these tests are given in Table 5. They show no significant improvement in any of the four leagues, thus confirming our informal decision.

Note: In applying this test to other situations, you will often want to standardize the predictor variables before constructing the two models. For example, if we used number of touchdowns (in football) as the offensive measure and number of yards allowed as the defensive measure, we should not expect the coefficients for the two predictors in Model #1 to be similar, even if both variables were equally effective predictors of winning percentage. However, if both predictors were converted to z-scores by subtracting their mean and dividing the result by their standard deviation, we could apply this technique to the transformed predictors. Since the offensive and defensive variables in our data must have the same mean and have very similar variances, we can dispense with the additional conversion to z-scores in fitting Model #2.

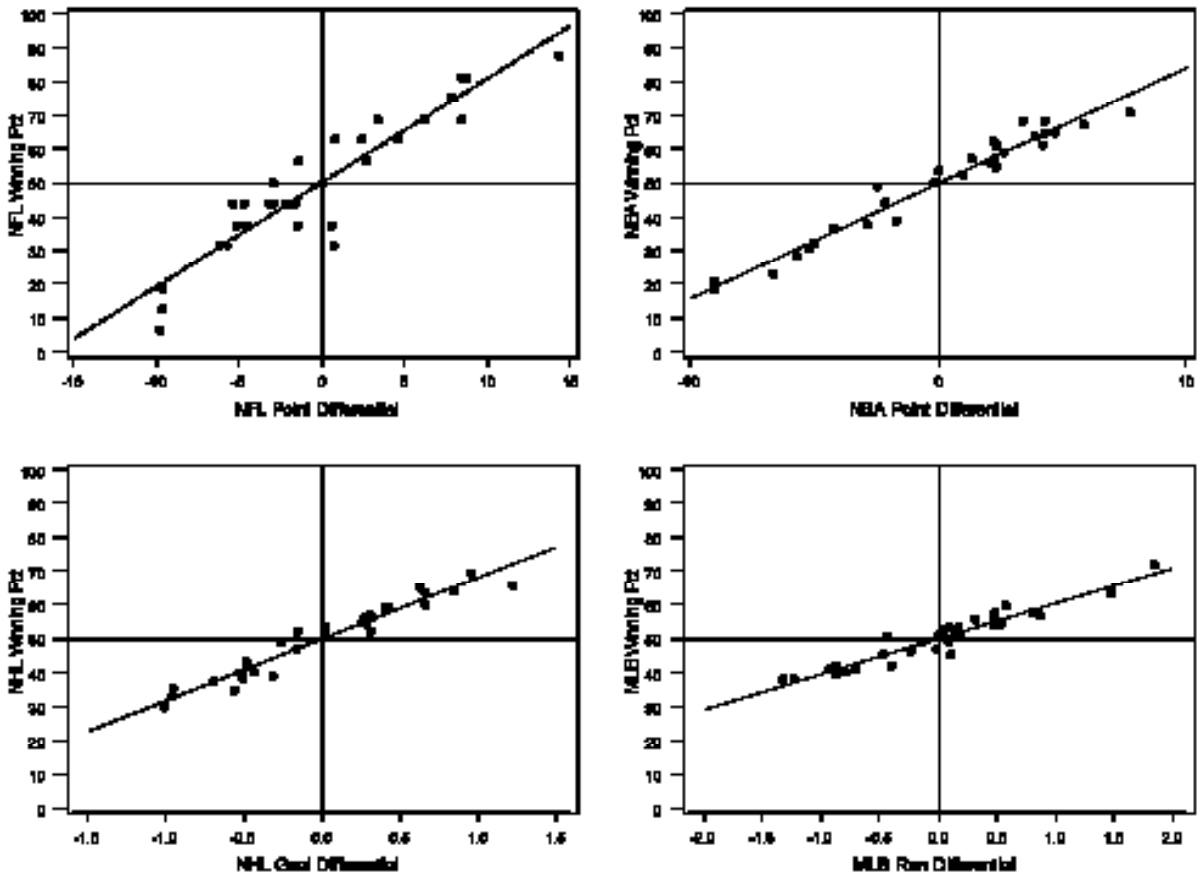


Figure 2. Predicting Winning Percentage Using Offense-Defense

Suggestions for Additional Investigations

What would a similar analysis show for another league (e.g., Women's NBA or Major League Soccer), different competitive levels (e.g., college or high school), or other seasons of the NFL, NBA, NHL, or MLB? Would combining the offensive and defensive ratings in some other way (e.g., a ratio of points scored to points allowed) provide a better single predictor than the difference? Can you find other measures of performance (e.g., turnover ratio in the NFL, shooting percentage in the NBA, save percentage in the NHL or team home runs in MLB) that might do a significantly better job than the basic points for and points against variables – or significantly enhance the predictive power of a model that already used those variables? Could a bootstrap technique be used as an alternate method to assess the difference in the offensive and defensive correlations with winning percentage (suggested by Tim Hesterberg)? Can you identify outlier teams in the data that did unusually well or poorly in winning percentage relative to their offensive and defensive statistics?

Conclusion

The data shows no significant difference in the ability of points scored (offense) or points allowed (defense) to predict winning percentages in any of the four major professional sports leagues. In each case, using the difference (offense – defense) provides a much better predictor than either variable individually – one that is nearly as effective as a multiple regression model that uses both variables. So a team that wants

Table 3. Estimated Coefficients for Model #1

League	Intercept	Offense	Defense	R^2
NFL	56.0	2.92	-3.22	86.4%
NBA	39.5	3.48	-3.37	95.8%
NHL	46.6	18.67	-17.43	94.1%
MLB	58.3	9.43	-11.2	91.8%

Table 4. Estimated Coefficients for Model #2

League	Intercept	Difference	R^2
NFL	50.0	3.08	86.3%
NBA	50.0	3.42	95.8%
NHL	50.0	18.07	94.0%
MLB	50.0	10.35	91.3%

Table 5. Partial F-tests for Difference between Offense and Defense

League	F-test statistic	p-value
NFL	0.24	0.625
NBA	0.28	0.602
NHL	0.30	0.591
MLB	1.69	0.205

to win needs to pay attention to both dimensions of the game.

Acknowledgments

Thanks to Jeff Witmer and Albyn Jones for independently suggesting the nested regression models approach for comparing these predictors and to Tim Hesterberg for suggesting the bootstrap.

Web Resources

Yahoo! Sports NFL standings at *http://sports.yahoo.com/nfl/standings.html*.

Information Please Almanac NHL standings at *http://www.infoplease.com/ipsa/A0881665.html*.

Doug's NBA & MLB Statistics Homepage at *http://home.rmi.net/~doug* (an especially good source for conveniently downloadable data from past seasons).

Data Sleuth Solutions

Mystery 1: Kentucky Derby Debates

Question 1: Overall, there is a downward trend in winning times. This is accounted for by improved racing horses from better breeding practices and also by better training of horses and jockeys.

Question 2: Variability is accounted for by different horses winning, different jockeys, conditions of the track at race time, weather, etc.

Question 3: There is a big drop in the winning times beginning in 1896. However, there is not a similar anomaly in the speeds. This happened because the length of the race was shortened from 1.5 miles to 1.25 miles.

Mystery 2: Who's Going to School?

Question 1: The shapes of both distributions are substantially skewed to the right. The males are older than the females by an average of almost two years. The ages of female students are a bit more variable.

Question 2: While males generally wait to take introductory statistics, more males take it between the ages of 21 and 23 than females who take it between the ages of 19 and 21.

Question 3: Very few males take introductory statistics at ages 19 and 20. Approximately 80% of the male students at Brigham Young University have served two-year missions. They usually began their service soon after their 19th birthday.

If you would like to explore these data further, the Kentucky Derby data can be found at *www.kentuckyderby.com*, and the age data can be found at the STATS web site (*www.amstat.org/publications/stats/data.html*).

A.P. Statistics

Some Thoughts on Paired Data

Gretchen Davis

On the 2001 Advanced Placement Statistics exam, students were asked to compare the active ingredient in a "name" brand drug and its generic counterpart. Many students did not realize that these observations were paired by pharmacy and that a paired *t*-test on the ten differences was therefore the appropriate procedure. We will consider several examples that may help students understand how to deal with dependent data. Our first example will be the "Soles of their Shoes" data from the classic text *Statistics for Experimenters*, written by Box, Hunter, and Hunter. Our second example will be the "Friday the 13th" data from the Data and Story Library (DASL), a statistical web site available at Carnegie Mellon University.

The Partial Story for ‡Soles of their Shoes·

Box, Hunter, and Hunter presented hypothetical data concerning researchers who wondered if a new less expensive synthetic material would be as long-lasting as the material that was currently used for the soles of sports shoes. Their experiment involved boys who were close in age. After the boys wore shoes with soles made of old and new material for a fixed period of time, the researchers measured the amount of wear-and-tear for both materials. Soles with high wear-and-tear numbers were less sturdy than soles with low numbers.

If we look at a back-to-back stemplot (Figure 1) for the amount of wear for the old and the new material, the results look similar. At first glance, it does not seem that one material is better than the other.

The average wear-and-tear for the old material is 10.63, which is a bit lower than the average of 11.04 for the new material. The standard deviations are close: 2.45 for the old and 2.52 for the new material. Thus, there is a bit more wear-and-tear on average and slightly more variation for the new material.

The researchers might have considered several basic designs for conducting this experiment. They could have selected twenty boys and randomly assigned ten to wear shoes with the new material and the other ten to wear shoes with the old material. If this were the case, we would then compare the two groups using an independent sample *t* test:

Old Material		New Material	
6	6	4	scale:
	7		13|6 means 13.6
82	8	8	
5	9	38	
987	10		
	11	238	
	12		
32	13	6	
3	14	02	
$\bar{x}_{old}$ =10.63, s_{old} =2.45		$\bar{x}_{new}$ =11.04, s_{new} =2.52	

Figure 1: Stemplot of Wear-and-Tear Measurements with Old and New Material

$$t = \frac{\bar{x}_{new} - \bar{x}_{old}}{\sqrt{\frac{s^2_{new}}{n_{new}} + \frac{s^2_{old}}{n_{old}}}} = \frac{11.04 - 10.63}{\sqrt{\frac{2.52^2}{10} + \frac{2.45^2}{10}}} = 0.37$$

two-sided p-value = .72

There is no evidence of a difference in average wear-and-tear between these two materials. Part of the reason is what we saw in the stemplots, the distributions are very similar. However, another reason is there is lots of variability in the two distributions, and lots of overlap between the two groups. This makes it difficult to detect any differences there might be in the two groups. We would be better able to compare the two groups if we could account for some of this variability first.

The Complete Story for Soles of their Shoes

However, the researchers chose a much better design for this experiment. They had each of ten boys wear a special pair of shoes with one shoe's sole made of the new material and the other shoe's sole made of the old material for a fixed time period, creating a matched pairs design. They randomly selected which foot

would wear the new material by flipping a coin for each boy. This randomization within each pair is necessary in case one foot tends to be rougher on shoes than the other. Table 1 reports the data.

Notice that some boys (e.g., 1 and 4) tend to have more wear and tear in general. In fact, if we plot these paired observations (Figure 2), we see there is a strong positive association. Boys who are rougher on one type of material tend to be rougher on the other type of material as well.

By having each boy wear both types of materials instead of just one, we can directly compare the wear-and-tear within each boy, instead of comparing the wear-and-tear behavior for boys 1 and 4 to the rest of the boys. This gives us a much more direct comparison of the two materials under more similar conditions (the "roughness" level of each boy). In fact, we can see that most of the boys fall above the "equal wear" 45º line. This gives us some evidence that the wear-and-tear has a tendency to be higher for the new material than for the old.

While pairing the data makes sense, we can no longer treat these as two independent samples, and the above analysis is no longer appropriate. Instead, to compare the wear and tear between the materials, we calculate the difference in wear for each boy (New-Old). The results are in Table 2 and Figure 3.

Again we see that on average the differences are positive. Performing a paired *t*-test gives:

$$t = \frac{\bar{x}_{diff}}{s_{diff}/\sqrt{n}} = \frac{0.41}{0.39/\sqrt{10}} = 3.35$$

two-sided p-value = .0086

We have strong evidence that the mean difference is different from zero. From the plot, we see that the new material is not as long lasting on average as the older material.

Why are these test results so different? Comparing the two test statistic calculations provides insight. Both numerators equal 0.41, but the denominator is much smaller with the paired test because the standard deviation of the *differences* is much smaller than the standard deviation of wear-and-tear measurements. As a consequence, the paired *t*-test is much more powerful in this case. In other words, controlling for that variation among boys helps us to detect a difference between the two materials.

After we explored the "Soles of the Shoes" example in a beginning statistics class at UCLA, Christine Corpus and Fiona Leung shared another data set that illustrates the same principles. These data came from DASL, and were part of a larger study on superstitions.

Another Example Is Friday the 13th Unlucky?

Researchers in England collected data on traffic accident victims who were admitted to the hospital emergency department during a four-year period (Scanlon et al., 1993). They recorded the number of admissions on Friday the 13th and also the previous Friday (the 6th). The data appear in Table 3.

Here, it is less clear that there is an advantage in collecting paired data. However, if we examine another scatterplot (Figure 4), there does appear to be a positive association between accidents on the 13th and on the previous Friday.

The pairing makes sense here because weather conditions and traffic patterns vary at different times of year, and we expect the results during a two week time period to be more similar than the results across different years. The choice of Friday the 6th rather than Thursday the 12th is because of different traffic patterns on different days of the week.

If we think (incorrectly) that these data are independent, we find no real difference in average admissions (t = 1.66, p-value = .64). But when we realize that the data are paired by month and year, then the differences are apparent (paired t = 2.71, p-value=.021). This supports that there are more admissions on average from traffic accidents on Friday the 13th than on Friday the 6th. (Some may also wonder about admissions on other Fridays or on Friday the 20th. Maybe the peak on the 13th is really just a midmonth high and not the result of one unlucky day.) By pairing, we were able to control for the variation in traffic patterns through the year (notice again that the standard deviation of the differences is smaller than the standard deviation for each group), giving us more power in our statistical analysis.

Conclusion

Both examples may help students understand the power of pairing, which is a form of blocking. The design of the study and the question of interest determine how the analysis should proceed. If the data were collected through pairing and not independent samples, then one analyzes differences rather than the raw data. If the variables are strongly associated, the pairing can enable one to detect an effect that could have been missed with independent samples.

Acknowledgements

I am grateful to Stu Hunter who shared the Soles of the Shoes data with readers at the first AP Statistics grading session in 1997. Stu also shared this wonderful quote:

"Statisticians are not number librarians."

References

Table 3: Traffic Accident Victims Admitted to Hospital Emergency Departments

Month, Year	Friday the 6th	Friday the 13th	Difference (13th - 6th)
October 1989	9	13	4
July 1990	6	12	6
September 1991	11	14	3
December 1991	11	10	-1
March 1992	3	4	1
November 1992	5	12	7
Average	7.50	10.83	3.33
Standard deviation	3.33	3.60	3.01

Box, G.E.P., Hunter, W.G., and Hunter, J.S., *Statistics for Experimenters*, John Wiley and Sons, 1978.

Scanlon, T.J., Luben, R.N., Scanlon, F.L., and Singleton, N. (1993), "Is Friday the 13th Bad For Your Health?," *British Medical Journal*, 307, 1584-1586.

Web Resources

The link to the Friday the 13th data is: *http:// lib.stat.cmu.edu/DASL*

The link to College Board web site that contains the 2001 free response questions and their rubrics is: *http://www.collegeboard.org/ap/statistics/frq01/index*

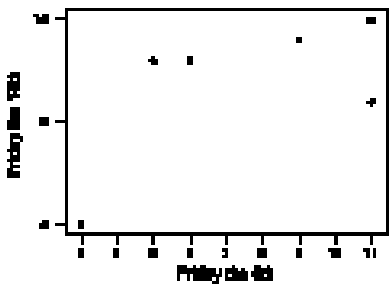


Figure 4.