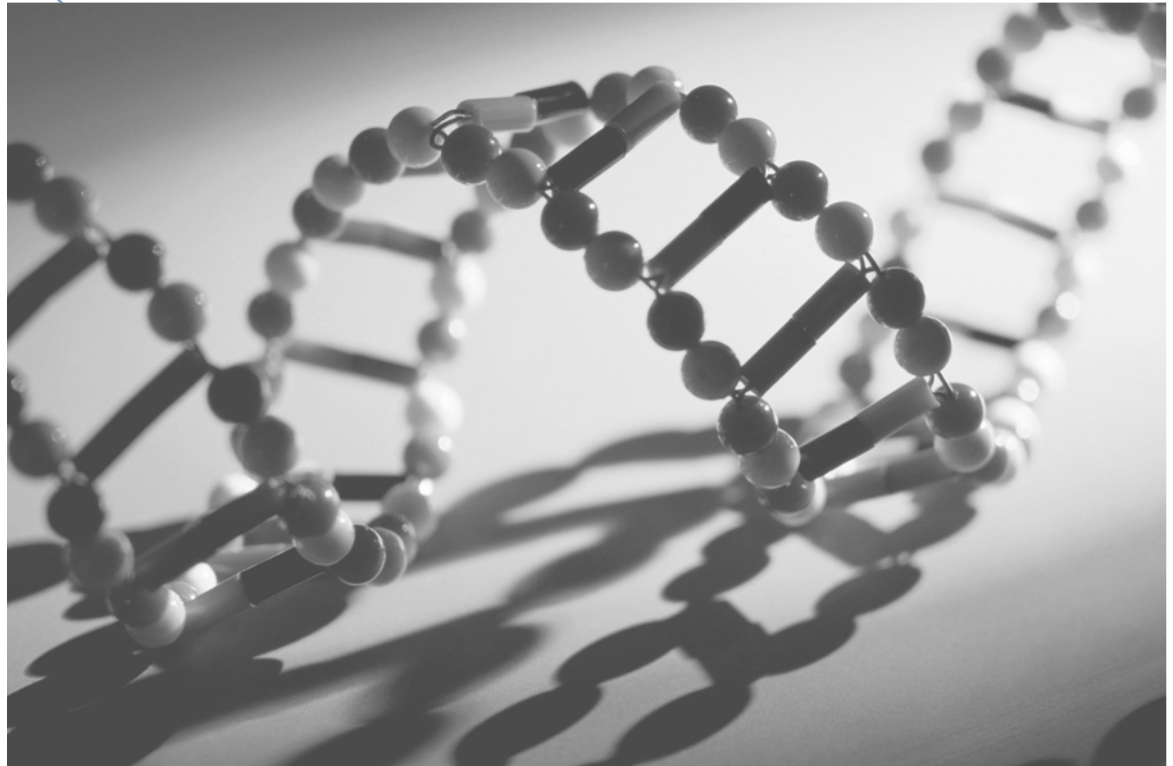


THE MAGAZINE FOR STUDENTS OF STATISTICS

WINTER 2005 • ISSUE # 42

STATS

STATS STATS STATS STATS STATS



MICROARRAY DATA

From a Statistician's Point of View

Stat Bowl 2005

Cluster Sampling: High Quality Information at a Bargain Basement Price

Was Pinkerton Right?

Non Profit Org-
U.S. Postage
PAID
Permit No. 361
Alexandria, VA



Did you recently complete
your statistics degree?

Postgraduate Members
pay only

\$ 40

For the first year after your graduation, you can join the ASA for only \$40. That is over 50% off the regular ASA membership rate!

Postgraduate members receive discounts on all meetings and publications, access to job listings and career advice, as well as networking opportunities to increase your knowledge and start planning for your future in statistics.

JOIN NOW!

To request a membership guide and an application, call 1 (888) 231-3473 or join online at

www.amstat.org/join



The Magazine for Students of Statistics

Winter 2005 • Number 42

Editor

Paul J. Fields
email:
pjfields@stat.byu.edu

Department of Statistics
Brigham Young University
Provo, UT 84602

Editorial Board

Peter Flanagan-Hyde
email:
pflanaga@pcds.org

Mathematics Department
Phoenix Country Day School
Paradise Valley, AZ 85253

Jackie Miller
email:
jbm@stat.ohio-state.edu

Department of Statistics
The Ohio State University
Columbus, OH 43210

Chris Olsen
email:
colsen@cr.k12.ia.us

Department of Mathematics
George Washington High School
Cedar Rapids, IA 53403

Bruce Trumbo
email:
brumbo@bay.csuhayward.edu

Department of Statistics
California State University, Hayward
Hayward, CA 94542

Production

Jennifer Campanile
email:
jennifer@amstat.org

American Statistical Association
1429 Duke Street
Alexandria, VA 22314-3415

Michael Campanile
email:
michaelc@amstat.org

American Statistical Association
1429 Duke Street
Alexandria, VA 22314-3415

Megan Murphy
email:
megan@amstat.org

American Statistical Association
1429 Duke Street
Alexandria, VA 22314-3415

A. Veronica Precup
email:
ronnie@amstat.org

American Statistical Association
1429 Duke Street
Alexandria, VA 22314-3415

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, VA 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 the 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. Material can be sent as a Microsoft Word document or within an email. Accompanying artwork will be accepted in graphics format only (.jpeg, etc.), minimum 300 dpi. No material in WordPerfect will be accepted.

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

Copyright © 2005 American Statistical Association.

Features

4 Microarray Data from a Statistician's Point of View

Johanna Hardin

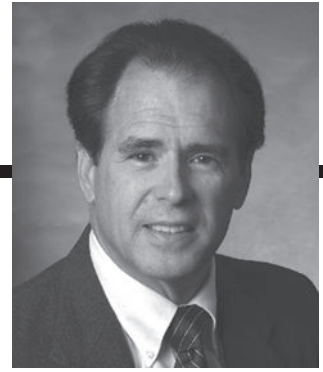


- 14 Play ASA Stat Bowl at JSM 2005 in Minneapolis!
Mark Payton
- 16 Advice from the 2004 Stat Bowl Champion
Jesse Frey
- 17 The First United States Conference on Teaching Statistics
Deb Rumsey

Departments

- 2 **Editor's Column**
- 20 **AP Statistics**
Cluster Sampling: High Quality Information at a Bargain Basement Price
Peter Flanagan-Hyde
- 24 **Statistical μ -sings**
Was Pinkerton Right? A Data Analysis of an Attempt at Espionage
Chris Olsen

EDITOR'S COLUMN



Paul J. Fields

Microarray technology is revolutionizing our understanding of life. Instead of looking at genes one at a time, with this new technology biologists can look at thousands of genes simultaneously to sort out their actions and interactions. In this way they can see a much clearer picture of cell processes and they can see that picture emerge much faster than ever before. By studying microarray data, biomedical researchers can find better ways to diagnose and treat diseases.

In this issue of *STATS*, Johanna Hardin provides us with an overview of this new and exciting field and explains the role of the statistician in facilitating the ability of scientists to see through the haze of uncertainty to learn how genes influence cell function. Pushing forward this new frontier can help biologists to better understand the most fundamental processes of life itself.

Emphasizing the partnership between biologists and statisticians, Professor Hardin explains how the statistician's tool kit of data analysis techniques can be used to learn about what is going on at the gene level inside living cells. In particular, she highlights the use of cluster analysis in examining microarray data. She also points out the challenges in microarray data analysis.

Microarrays produce huge amounts of data—millions of data points with thousands of parameters to be estimated. Careful statistical thinking and analysis are required to find the underlying structure in the

data. The unprecedented amounts of data produced by microarrays raise new challenges for statisticians to be able to perform inference on a scale never before conducted. Her article is a great introduction to the topic of microarray data analysis and a handy reference for further study in this burgeoning new field.

Statistics is the cornerstone of scientific inquiry. It embodies the philosophical basis for research and provides the tools for discovery. Microarray technology is an example of the new generation of scientific approaches we, as statisticians, will be part of in the 21st century. The challenges are historic in scale and the potential benefits to society are immense. What an exciting time to be a statistician!

Stat Bowl is a great opportunity to learn more about statistics and have some fun at the same time during the Joint Statistical Meetings. Mark Payton, Stat Bowl moderator, invites you to come and participate in Minneapolis next August. Jesse Frey, last year's STAT Bowl champion in Toronto, offers helpful advice to prospective participants. Try the sample Stat Bowl questions and see how you do. Remember that there are even some travel funds available for participants, so sign up early.

Peter Flanagan-Hyde is a new member of the *STATS* editorial board. He will be contributing to our AP Statistics column. In this issue he explains cluster sampling and how this technique can help researchers efficiently gather representative data from a population even though the cost of sampling makes taking a simple random sample impractical.

Finally, in this issue's μ -sing, Chris Olsen walks us through a statistical analysis to answer a historical question: Was Civil War military intelligence sleuth Alan Pinkerton as incompetent as historians have said? Read his description of applying statistical estimation to espionage and find out his answer!



A handwritten signature in black ink that reads "Paul J. Fields". The signature is written in a cursive, flowing style.

Paul J. Fields

CALL FOR PAPERS

STATS: The Magazine for Students of Statistics is interested in publishing articles that illustrate the many uses of statistics to enhance our understanding of the world around us. We are looking for engaging topics that inform, enlighten, and motivate readers, such as:

- statistics in everything from sports to medicine to engineering.
- “statistics in the news,” discussing current events that involve statistics and statistical analyses.
- statistics on the Internet, covering new web sites with statistical resources such as data sets, programs, and examples.
- interviews with practicing statisticians working on intriguing and fascinating problems.
- famous statisticians in history and the classic problems they studied.
- the “statistics almanac” that tell us what happened during this month in statistics history.
- how to use particular probability distributions in statistical analyses.
- examining surprising events and asking, “What are the chances?” and then providing the answers.
- “statistical data sleuth” problems.
- reviews of books about statistics that are not textbooks.
- student projects using statistics to answer interesting research questions in creative ways.

So think of some great ideas, and send a description of your concepts for feature articles that you would write to Dr. Paul J. Fields, Editor, pjfields@byu.edu.



Microarray Data from a Statistician's Point of View



Johanna Hardin

Reports in the news often tell about how genes determine the chances of getting a particular disease or how a genetic mutation can increase susceptibility to certain environmental changes. For example, Familial Adenomatous Polyposis (FAP) is a type of colon cancer that affects one in 8,000 people in the United States. FAP is caused by a mutation in the adenomatous polyposis coli (APC) gene. It has been estimated that a person with FAP has over three times the relative risk of dying than a person without FAP (Nugent et al. 1993).

It is well known that the DNA in a cell's nucleus contains the instructions for building proteins. A gene is a segment of DNA that contains the instructions for building a specific protein. If different genes are active, then different proteins will be produced in a cell. Skin cells are different from muscle cells, for example, because different proteins are present in the two different types of cells. When a gene is active in a cell, we say that the gene is "expressed." Information about genetic activity can give insight into biologic processes and cell behavior—both normal and cancerous.

Measuring genetic activity is the role of molecular biologists. Until recently, scientists analyzed gene activity one gene at a time. Now activity can be measured on tens of thousands of genes simultaneously using a new tool known as a DNA microarray (Eisen and Brown, 1999). Interpreting the gene expression data is the role of statisticians. The huge volume of data from microarray analyses brings new statistical challenges and the need for new analytical techniques.

As statisticians, our role in many scientific fields, particularly in the field of molecular biology, is vital and fascinating. Because microarray data analysis is a new and expanding research area, I cannot hope to cover in this article all of the current research associated with

microarray analysis. So my goal is to give an overview of the analysis process and the related statistical issues.

Why Microarrays?

Information (that can be obtained from microarrays) about genes helps us answer a myriad of biological questions:

- What genetic differences are there between healthy people and people with a particular disease?
- Are there genetic subgroups of people with a particular disease who respond positively to a given treatment?
- What kinds of genetic changes happen across time or after frequent doses of a treatment?
- Which genes are co-regulated—have expression levels that increase or decrease concurrently—in a particular biological system?
- What is the likelihood of acquiring a particular disease, given a person's genetic make-up?

What is a Microarray?

DNA microarrays, first introduced commercially in 1996, come in a variety of forms, but they all contain the same basic design. Each microarray consists of thousands of single strands of genetic material tethered to a "chip" the size of a thumbprint. The chips (which are not reusable and should not be confused with computer chips that can store and restore information) are produced at numerous academic and research laboratories, and they are also produced commercially.

Microarray technology uses a fundamental property of DNA called "complementary base pairing." Our DNA gives the blueprint for the functioning of the cell written in sequences of chemical bases: adenine (A), cytosine (C), guanine (G), and thymine (T). These bases bind in a double helix structure to create the DNA molecule (See Figure 1). At each rung along the DNA ladder, A always binds with T, and C always binds with G. Thus A is complementary to T, and C is complementary to G. Each spiral strand is connected to a complementary

Johanna S. Hardin (jo.hardin@pomona.edu) is an Assistant Professor at Pomona College. She received a BA from Pomona College and an MS and PhD from the University of California, Davis. Her current research interests include analysis of microarray data (normalization, distributional qualities, clustering, and outlier detection) and other high dimensional data sets.

strand by the paired bases. A subsequence of the gene characterized by TGAACT on one strand would have ACTTTGA on the complementary strand of the DNA molecule.

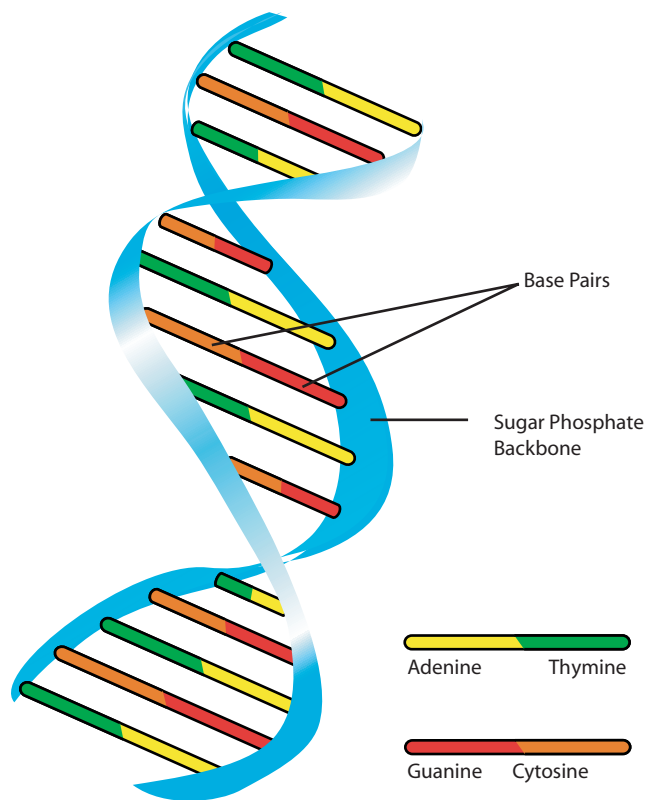


Figure 1. Illustration of the DNA double helix molecule showing the complementary base pairing on the rungs of the DNA ladder.

The DNA code is identical in each cell nucleus through the entire body. However, in order for cells to function appropriately, each different cell type receives a different message from the DNA. A segment of DNA is converted into an intermediary form known as messenger RNA (mRNA) that exits the nucleus and serves as a template for building proteins.

Consequently, we could determine which genes are expressed in a cell by measuring the quantity of mRNA there is in the cell corresponding to that gene. However, free mRNA in a cell is very unstable, so it is treated with an enzyme to convert the mRNA back into DNA. This form of DNA is known as complementary DNA (cDNA).

Through a denaturing process the double-stranded DNA molecules in the sample are unzipped down the middle into two single-stranded molecules. The microarray chip itself also contains single strands of genes that will attract the single gene strands from the sample. The single strands from the sample will bind with the single strands on the microarray chip to reform the DNA double helix.

In a microarray experiment, the test sample is labeled with a dye and a reference sample is labeled with a dye of a different color. The reference sample serves as a control to which the gene expression in the test sample is compared. For instance, if we wanted to determine which genes are expressed in a tumor sample, we could use a tissue sample from a healthy individual as the reference sample. We would then compare the expression level of each gene in the tumor sample to the expression level of each gene in the reference sample. Suppose the tumor sample had been labeled with a red dye and the reference sample had been labeled with a green dye. Then a red spot on the microarray would indicate that the gene corresponding to that spot is expressed at a higher level in the tumor sample than in the reference sample. Similarly, a green spot would indicate that the gene is expressed at a lower level in the tumor sample.

There are several techniques for constructing DNA microarrays (Schena et al. 1995; Velculescu et al. 1995; Lockhart et al. 1996). Though there are slight differences in the microarray technologies, one basic outline of the microarray procedure can be summarized as follows:

1. Label the sample with a fluorescent dye.
2. Isolate the cDNA from the cells of interest, e.g., tumor cells, plasma cells, etc.
3. Denature the sample so that the cDNA are in single strands instead of the double helix form.
4. Place the sample onto the microarray chip and allow the double helix structure to restore itself.
5. Wash the remaining sample off the chip so only the parts of the sample that have bound to the chip remain.
6. Scan the microarray chip with a laser to quantify the fluorescence of each individual gene. The more of the sample that is stuck to the chip, the higher the fluorescence.

A. Malcolm Campbell at Davidson College has put together an animation of the microarray process which can be seen at website: www.bio.davidson.edu/courses/genomics/cgip/chip/html.

In general, the amount of activity of a gene is represented by the number of replicates of that gene in a particular sample of cells. A high fluorescence level indicates that multiple copies of a gene have bound to the chip and that the gene has high activity in the cell. Similarly, a low fluorescence level indicates low activity of the gene in the cell. By quantifying the fluorescence, the gene activity can be compared across different samples, e.g., a group of healthy samples compared to a group of tumor samples.

A sample scan of part of a chip is shown in Figure 2 (see page 6). The image shown in the figure is only part of the chip. Each spot represents a gene and there are thousands of genes on a chip. A red spot indicates that sample 1 (the "red" sample) has high genetic activity for that gene. A green spot indicates that sample 2 (the "green" sample) has high genetic activity for that gene. The yellow spots indicate the genes where the

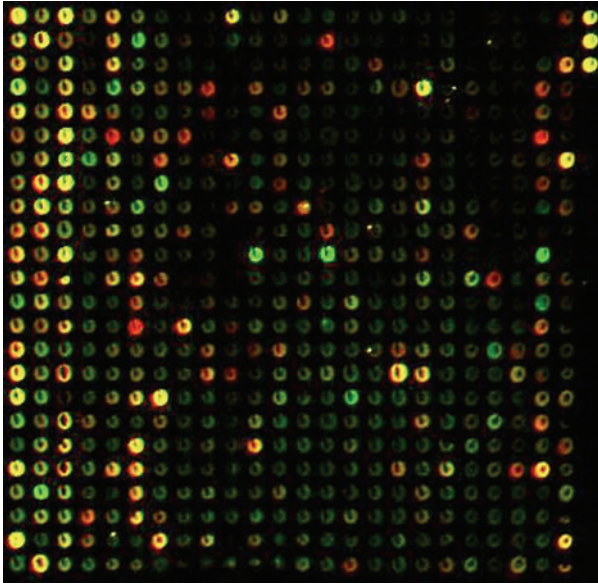


Figure 2. Part of a microarray chip. Each spot represents one gene and the color represents the activity level of the gene in the test sample.

two samples have similar activity, and the black spots indicate where there is no activity.

An example of some microarray data is given in Table 1. The data came from an experiment on aging yeast in Laura Hoopes' lab at Pomona College. The test sample (treatment) contains older yeast cells, while the reference sample (control) contains younger yeast cells. The test sample was dyed red and the reference sample was dyed green. The table only shows the expression level of ten genes as an illustration. In an actual analysis there would be data for many more genes.

From the numerical values we can identify the genes that are highly expressed overall in the experiment and

Gene	Red Intensity	Green Intensity	Ration of Intensities	Log ₂ of Ratio
YBR124W	92	78	1.179	0.238
YBR100W	103	77	1.338	0.420
MRS5	369	357	1.008	0.012
ECM33	3423	2663	1.285	0.362
YBR075W	196	133	1.474	0.559
HSP26	805	175	4.600	2.202
VAP1	158	175	0.903	-0.147
YRO2	118	373	0.316	-1.660
YBR051W	125	135	0.926	-0.111
RPS11B	3855	3739	1.031	0.044

Table 1. Sample microarray data from an experiment on aging yeast cells. Red intensity refers to the test sample and green intensity refers to the reference sample. The ratio of intensities tells us the multiplicative change and the log base-2 ratio gives the difference of the data after a log transformation. Data courtesy of Laura Hoopes of Pomona College.

the genes that are just barely expressed. Note genes RPS11b and YBR124W, for example.

Additionally, by taking the ratio of intensities we can identify the genes that are most highly expressed in the treatment sample relative to the control sample and vice versa. Taking the logarithm of the ratio helps to further distinguish the genes with the highest and lowest relative expression levels. Note genes HSP26 and YRO2, for example.

What is the Statistician's Role?

Although it is preferable for the statistician to have a hand in the experimental design, the statistician often comes into a microarray analysis project once the data have been collected. The statistician's job is to use the numerical fluorescence levels to make claims about the populations of interest. Of course, the methodology will depend on the question at hand. The computations can be broken down into two main parts: data cleaning and data analysis.

Though the microarray construction seems straightforward in theory, in reality there are numerous sources of variation. For example:

- Spots that are not systematically placed on the chip,
- Samples that smear outside of the measurement surface,
- Dyes that fluoresce at different levels (green is "stronger" than red), or
- Arrays with a variable amount of dye.

To address these problems, the data cleaning step involves image processing, normalization, and standardization. Current research on all three cleaning steps is active and growing. In this article I focus on data analysis instead of data cleaning, assuming the data are already "clean." Many software programs designed for microarray analysis give options for cleaning the data.

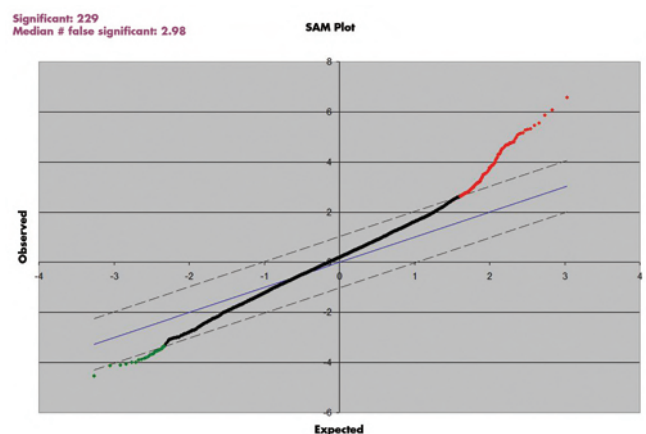


Figure 3. Plot from the SAM analysis for the MM versus MGUS comparison. Each dot represents a particular gene. The x-coordinate is the observed value of the test statistic and the y-coordinate is the expected value of the test statistics under hundreds of permutations. The dotted boundary is the cutoff for significance given a specified false discovery rate.

What is an Example of Microarray Analysis?

To illustrate some of the typical statistical techniques applied to microarray data, let's examine a real data set from a particular type of commercial chip—Affymetrix (version 5). The samples were taken from three populations: a group with multiple myeloma (a blood cancer abbreviated MM), a group with signs of developing MM (abbreviated MGUS for “monoclonal gammopathy of undetermined significance”), and a healthy group.

In this situation, plasma cells from each of the test subjects were isolated and placed on a microarray chip since multiple myeloma is characterized by plasma cells replicating out of control, which in turn causes organ damage. The Affymetrix chip measures 12,625 genes simultaneously. There were 218 MM samples, 21 MGUS samples, and 45 healthy samples. The data were collected at the Donna D. and Donald M. Lambert Laboratory of Myeloma Genetics, University of Arkansas for Medical Sciences by John Shaughnessy, Jr., and his colleagues (Zhan et al. 2002).

What Statistical Techniques Can We Use?

The tools from basic statistics can be used to address many microarray research questions; however, each research hypothesis requires a different statistical tool.

Comparing Two Groups

Probably the most common research question associated with microarray data is the two group comparison: What differences in genetic activity are there between one group of samples and another group of samples? Usually the first group of samples comes from people with a particular disease and the second group comes from healthy people. We'd like to know what type of genetic activity differentiates the two groups.

For example, we might be interested in comparing the MM group with the MGUS group. The *t*-test from basic statistics can be used to test whether the means of the two populations are the same. By applying the *t*-test separately to each of the 12,625 genes on the chip, we can tell which genes have an average gene expression that is different between the two groups. In the multiple myeloma example the *t*-test found 422 gene comparisons with *p*-values less than 0.001. Such a *p*-value indicates that the probability is less than one in a thousand that the difference occurred simply by chance.

When the *t*-test is not appropriate (when the data are not normally distributed, for example), we could use the Wilcoxon rank sum method to test whether the median expression levels are the same in the two populations. In the multiple myeloma example the Wilcoxon rank sum test found 341 gene comparisons with *p*-values less than 0.001. In comparing the MGUS and MM samples, the intersection of the genes with *p*-values less than 0.001 for both the *t*-test and the Wilcoxon rank sum test was a set of 269 genes. One analysis approach is to investigate further only

those genes that are significant using both types of comparisons.

Other methods have been developed to compare two groups in the context of microarray data. Some researchers have used a modification of the *t*-test (Golub et al. 1999). Researchers at Stanford University have developed a software package (SAM—Significant Analysis of Microarrays) to conduct a permutations test to establish cutoffs for the pairwise comparisons (Tusher et al. 2001; Tibshirani et al. 2002).

The SAM technique applied to the MM versus MGUS data identified 229 genes as showing significant activity. The false discovery rate (FDR) was only about three genes based on hundreds of random permutations of the gene values. Figure 3 (see page 6) shows an output plot from the SAM analysis.

For comparison, Table 2 presents the results of the *t*-test, the Wilcoxon test, and the SAM method.

Comparison	Number of Significant Genes ($p < 0.001$)			
	<i>t</i> -Test	Wilcoxon Rank Sum Test	SAM (FDR=3/229)	By Chance
MM versus MGUS	422	341	229	12.63

Table 2. The number of genes judged by each method to have a statistically significant expression level in comparing the MM and MGUS groups. Note that with $\alpha = .001$ there could have been 12.63 genes identified as significant simply by chance even if there was no effect due to the disease.

Comparing Multiple Groups

Since the data actually contain three groups (MM, MGUS, and healthy groups), we could use analysis of variance (ANOVA) to find genes that have an average expression level that is different in at least one group. Just as we used a nonparametric version of the *t*-test (Wilcoxon rank sum test) in the two-group comparison, we could also use a nonparametric version of ANOVA (Kruskal-Wallis test) to analyze non-normal data from more than two groups. Each of these tests produces a *p*-value for the difference across the groups for every single gene. Significant differences can then be identified based on the magnitude of each *p*-value.

Classification

Sometimes the research question has to do with predicting class membership in a set of data for which the classes are already known, that is, classifying a new sample into a known class. In this situation, we could use past data to set up a logistic regression model that can classify a future sample data point. One way to test the accuracy of the model is to classify a subset of points with known class membership that was not used in building the model. These independent data values will give unbiased information about the accuracy of

the classification procedure. When applied to an entire dataset, this procedure is called cross validation. The algorithm for the cross validation procedure is:

1. Partition the data into k groups of the same size.
2. Remove the first group from the data and build the model on the remaining $k-1$ groups.
3. Test the removed group of data using the above model and record the predicted class membership.
4. Repeat steps 2 and 3 for each of the k groups.
5. Compile the false positive and the false negative rates as a measure of model accuracy.

Using logistic regression with expression level as the explanatory variable and the disease groups (MM vs. healthy, MM vs. MGUS, healthy vs. MGUS) as the response variable, we can create models that predict the classification of future observations into dichotomous categories such as sick or healthy. The accuracy of these three separate models was evaluated using cross validation. The results are displayed in Table 3.

	MM	Healthy	MM	MGUS	MGUS	Normal
Percentage Correctly Classified	96.79%	84.44%	93.58%	38.10%	91.11%	71.43%

Table 3. Results from using logistic regression to predict class membership (MM, MGUS, or Healthy). The effectiveness of the model is evaluated using cross validation. Each entry in the table is the percentage of samples correctly classified.

It is apparent from the results that logistic regression using gene expression values can be used to discriminate between the healthy group and the malignant groups, but it is not useful for discriminating between the two malignant groups (MM and MGUS). This lack of discrimination is seen in the comparison of MM versus MGUS where most of the MGUS samples (62%) were incorrectly predicted to be from MM patients.

This and other discrimination methods for microarray analyses have been compared using cross validation prediction error rates (Hardin et al. 2004).

Clustering

Clustering is a process by which data can be grouped without any preconceived knowledge of the groupings or even of the number of groups. While classification models are referred to as “supervised learning,” clustering is sometimes referred to as “unsupervised learning.” As with most techniques, there are different clustering algorithms, yet many use some type of metric to establish a distance between two samples or two groups of samples. The concept in clustering is that the closer two items are to each other, the more likely they are related and should therefore be grouped together.

Clustering techniques provide a visual representation of patterns in the data. Groupings or clusters can illustrate relationships that may or may not be known by the researcher. For example, a particular clustering result may demonstrate what gene expressions are useful for characterizing genes with known functions. Or, a clustering result may lead to the discovery of groups of genes that have similar expression patterns. Clustering can also be performed on samples instead of genes. When we cluster samples, we look for similar genetic patterns in groups of individuals.

In hierarchical clustering, the first step is to link the two closest samples. Subsequently, that pair is compared to the remaining samples and either another two samples are linked or the first pair (cluster) is linked to a third sample based on which of these choices represents a shorter distance. This process continues until every sample in the data set is linked to another.

Figure 4 (see page 10) shows the sequential linkages of a sample of patients in our Multiple Myeloma example. Each vertical line represents one sample. The samples from healthy people are labeled “X” and the samples from the MGUS patients are labeled “MGUS.” The MM samples are not labeled.

We can see that the MM samples tend to cluster together to the left and the healthy samples tend to cluster to the right, while the MGUS samples are dispersed throughout. This could indicate that maybe some of the MGUS samples will develop into MM while others of them will remain benign.

To illustrate the clustering process, figure 5 (see page 10) is a magnification of the grouping of predominantly healthy patients on the right side of figure 4. For merges of a pair of samples, the value on the y -axis represents the Euclidean distance between the two samples. Where two clusters are merged, the value on the y -axis represents the average of the distances between each of the samples in one cluster and each of the samples in the other cluster. Notice that merges shown at the lower portion of the graph are samples that are the closest to each other (most similar), while merges shown at the upper portion of the graph are samples that are the farthest apart (least similar).

Figures 6 (see page 11) shows clusterings of samples from only MM patients. We notice that there still appear to be some groups of samples even though all of the patients have the disease. This might indicate that some of the samples are genetically related in such a way that those patients would respond similarly to treatment.

Figure 7 (see page 11) shows the results of clustering using a set of 50 completely randomized expression values. Because the data are randomly distributed, we should not expect to see any clustering pattern. Interestingly, however, we can see some possible group, even though there should be no structure to the data. But when we compare figures 6 and 7, the groupings of random values in figure 7 are less distinct than the groups of real values in figure 6. We can also see that the distances between random values in figure 7 are much longer than the distances between real values in figure 6.

Consequently, because the clustering algorithm forces some configuration, we must be careful in deducing that there are significant relationships among the samples. A statistician should use these clustering methods carefully, especially when communicating with nonstatisticians, so as not to overinterpret any apparent structure in the data. Interpreting the groups within a hierarchical cluster is somewhat subjective and does not follow a formal structure of decision-making as in hypothesis testing.

Other classification and clustering techniques commonly used on microarray data include “nearest shrunken centroid” classification (Tibshirani et al. 2002) and “model-based clustering” (Yeung et al. 2001).

Advanced Techniques

Advanced techniques are often applied to microarray data and new methods are constantly being developed to better analyze the data. Some examples of advanced techniques we frequently use include:

- *Time Series Analysis* – With time series analysis we can observe trends over time for organisms like yeast, for example, that change rapidly (Zhao et al. 2001).

- *Partial Least Squares and Principal Component Analysis*—Both of these methods allow the analyst to reduce the dimensions of the data in a meaningful way. Since many data sets have hundreds of samples with thousands of dimensions, it is important to reduce the dimensions in a way that captures the signal while discounting the noise (Nguyen and Rocke 2002; Yeung and Ruzzo 2001; Bair and Tibshirani 2004).

- *Discriminant Analysis*—This is a way of partitioning the data and can be used for classification problems (Dudoit et al. 2002).

- *Survival Analysis*—This technique is used to evaluate data with censored endpoints that are common in medical studies. “Censoring” occurs when a patient dies or for some other reason does not complete the study. The Cox proportional hazards model—the standard survival model—is not equipped to handle thousands of explanatory variables and so variable reduction techniques must be used to fit survival analysis models (Pauker et al. 2004; Bair and Tibshirani 2004).

What Statistical Issues are Specific to Microarray Analyses?

Many of the techniques used to analyze microarray data are straightforward applications of well-known methodology and some of the established procedures can be modified to handle large data sets. However, some issues cannot be dealt with using standard statistical approaches and research is needed into new techniques to address specific problems.

One problem with microarray data is that the number of genes is almost always bigger than the sample size. This type of sparse data makes inverting covariance matrices impossible, which in turn forces

us to pare down the number of variables for methods like regression analysis that use inverted covariance matrices to calculate least squares estimates. Some data reduction techniques have been developed, but there is more work to be done to develop new methods for ascertaining what set of variables would be the most informative.

Because we often are interested in understanding particular genes, we use gene-by-gene techniques like *t*-tests, ANOVA, or regression analysis. Each time a gene is judged to be significant according to one of these tests, there is the risk of producing a Type I error. If we were to set our significance level to $\alpha = 0.05$ and run *t*-tests on 10,000 genes, we would expect 500 genes to test as significant, even if there is no signal in the data. The problem of controlling for this type of error in general has been studied widely (Benjamini and Hochberg 1995) and is now being researched in the specific context of microarray data (Storey, 2002).

Another problem is that microarray data do not conform to the usual assumptions of many standard statistical tests. The data themselves are in units of fluorescence and are often highly skewed right and can even be negative if a “background adjustment” is needed when the background fluorescence is brighter than the foreground fluorescence. Often log transformations (with some ad hoc adjustment for the negative values) give data that are moderately symmetric. However, log-transformed microarray data may still have highly unequal variances for which many techniques (like ANOVA) are not robust. Transformations and normalizations for microarray data are being researched so that the results from standard statistical analyses, based on the usual requirements, are reliable (Durbin et al. 2002).

What Software is Available for Microarray Analyses?

New software is constantly being developed to perform analyses specifically for microarray data. Because the technology is relatively new, much of the software is being developed in academia and is freely available. Below are summaries of a few of the most commonly used software programs. The synopses are based on my experience and not meant as endorsements or condemnations of any of the software.

- **Bioconductor:** This is a free program that runs in R. It is designed for statisticians who are researching new techniques on microarray data. It is flexible, though it does require basic programming knowledge of R or S-Plus. Bioconductor also has multiple graphs and features designed specifically for extracting information from microarray data.

www.bioconductor.org

- **SAM & PAM:** Significance Analysis of Microarrays (SAM) and Prediction Analysis of Microarrays (PAM) are free software programs that add-in to Microsoft Excel or R. SAM produces pictures and lists of genes that are

Continued on page 12

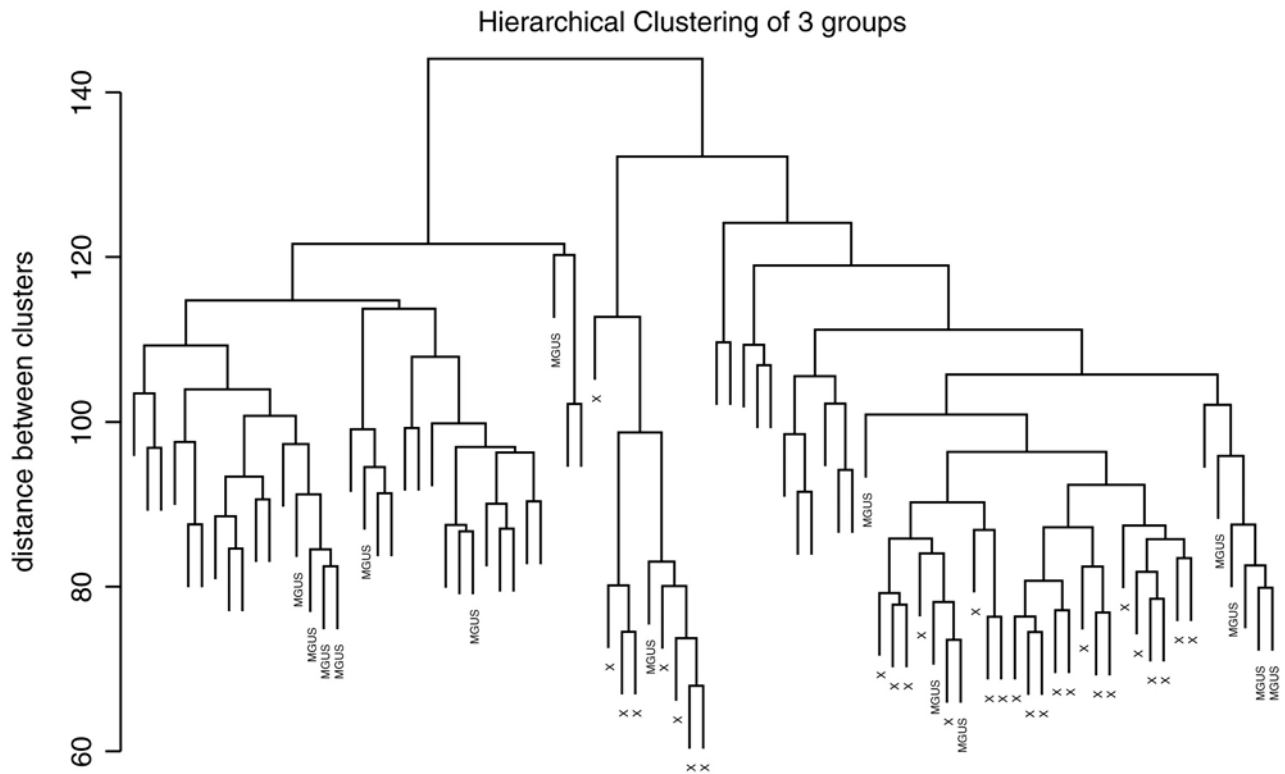


Figure 4. Hierarchical clustering of 85 randomly selected MM, MGUS, and Healthy Samples.

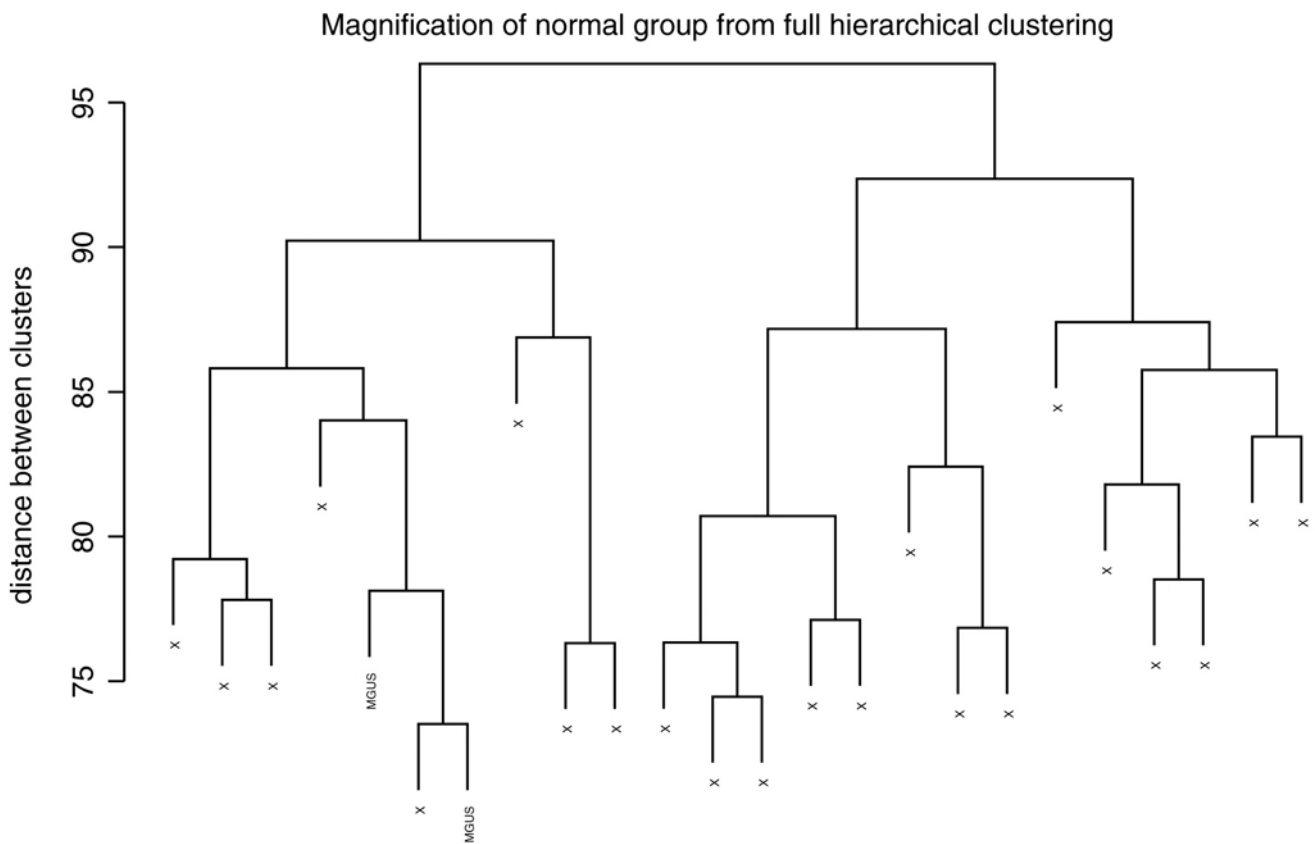


Figure 5. Hierarchical clustering analysis of 24 samples from healthy patients.

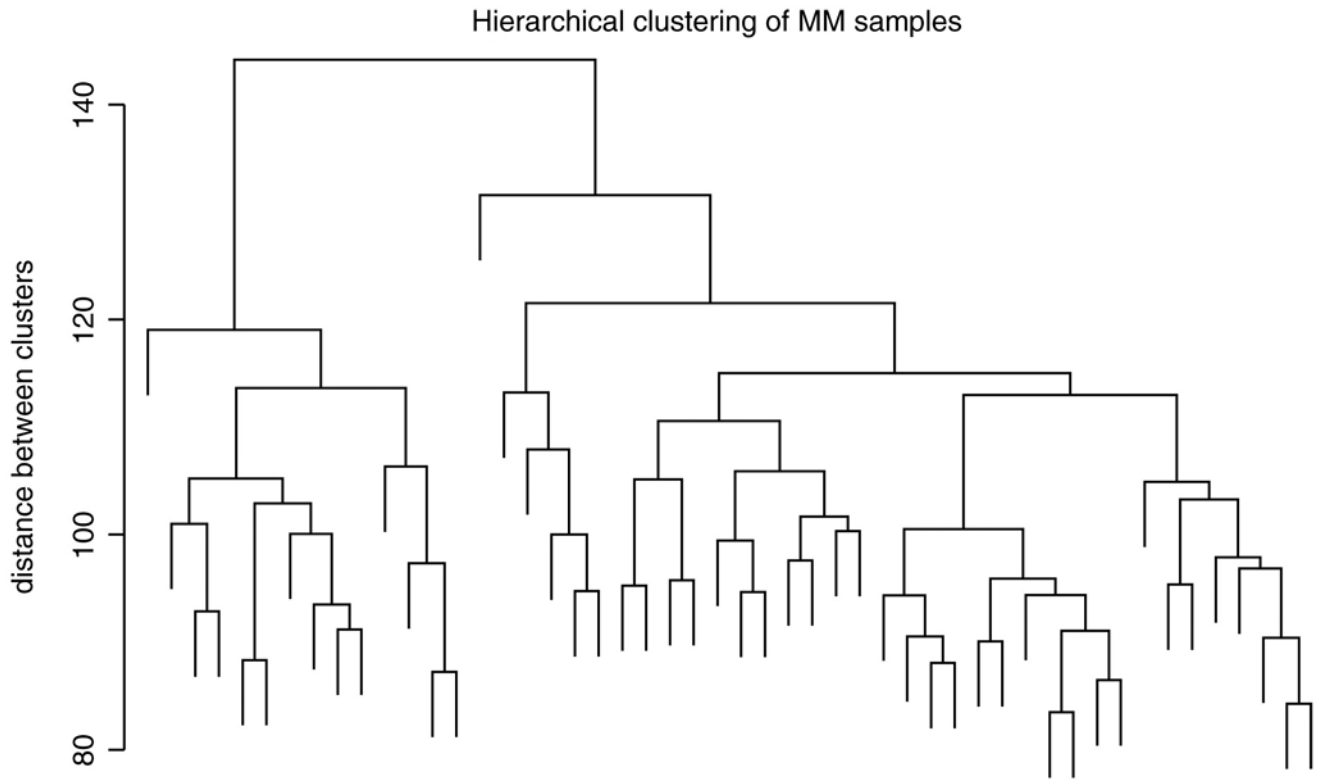


Figure 6. Hierarchical clustering analysis of 50 samples from MM patients.

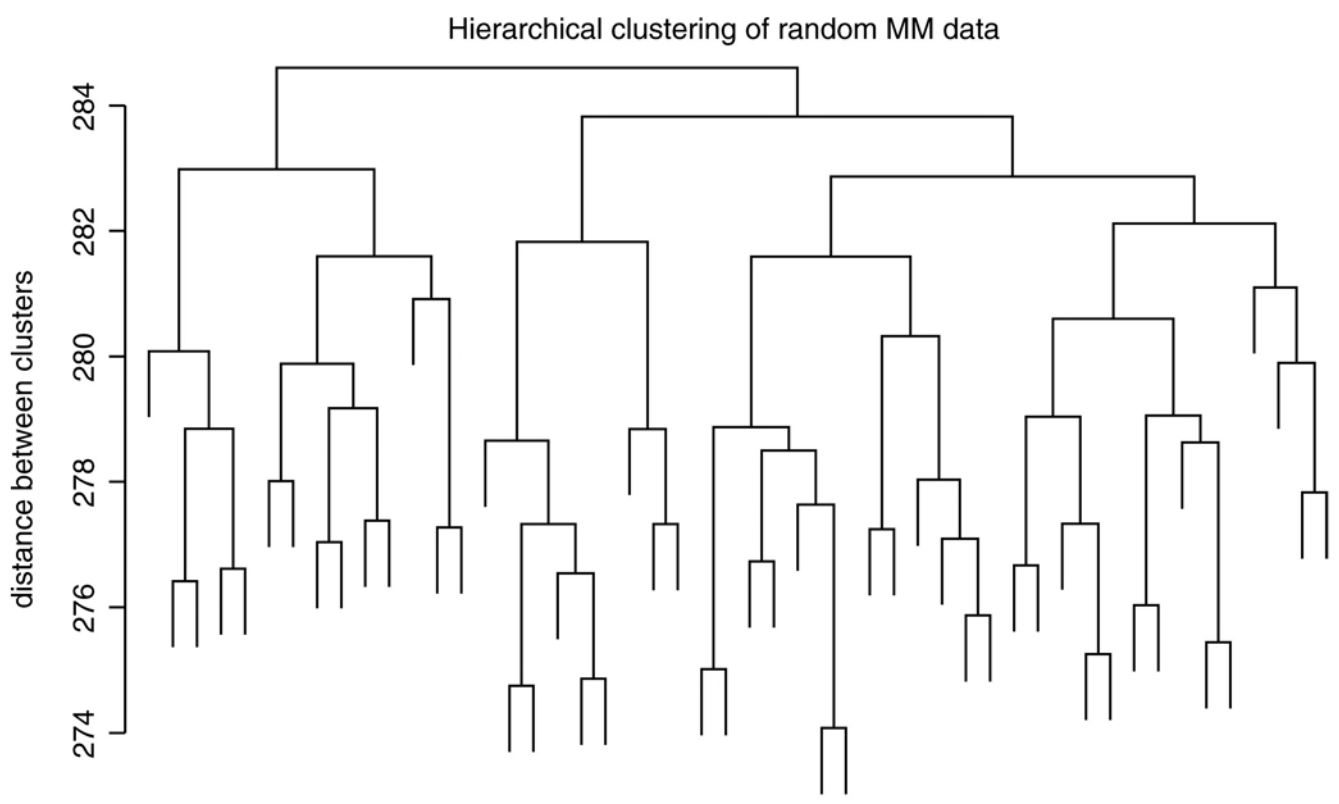


Figure 7. Hierarchical clustering analysis of 50 random expression values.



significant across groups, while controlling for the false discovery rate. It also correlates gene expression to clinical parameters. PAM performs classification of microarray data using nearest shrunken centroid methods.

www-stat.stanford.edu/~tibs

- **Cluster & Treeview:** These programs perform hierarchical clustering across both samples and genes. The results are displayed in tree-based images with label information and colors representing expression levels. Cluster and Treeview are both freely available software programs.

<http://rana.lbl.gov/EisenSoftware.htm>

- **BRB ArrayTools:** This software is designed as a free add-in to Microsoft Excel for visualization and statistical analysis of microarray data. It contains various methods including class comparison, class prediction, and permutation tests for significance levels.

<http://linus.nci.nih.gov/BRB-ArrayTools.html>

- **GeneSpring:** This package is widely used by biologists and geneticists. It is user friendly and has many good statistical techniques, including adjustments

for multiple comparisons. However, it is not free and not as flexible for statistical research as other programs.

www.sigenetics.com/GeneSpring/GeneSpring.html

Summary

For biologists, microarray technology has opened new avenues to access a new world of knowledge quickly and inexpensively. Never before has it been possible to study so many genes simultaneously on so many samples. However, any technology is limited by its ability to extract information.

As statisticians, it is our role to ensure that the information obtained from microarray experiments is valid and interpreted appropriately. Many of the statistical concepts from the last century are applicable to microarray analysis, but we must also open our minds to new techniques and methodologies that will be better suited for this new generation of data. In this century, our contribution to science will be to develop the analytical tools that can handle future generations of data yet to come.

References

- Bair, E., and Tibshirani, R. 2004. Semi-supervised methods to predict patient survival from gene expression data, *PLOS Biology*, 2:511-522.
- Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society*, B, 57:289-300.
- Dudoit, S., Fridlyand, J., and Speed, T. 2002. Comparison of discrimination methods for the classification of tumors using gene expression data, *Journal of the American Statistical Association*, 97:77-87.
- Durbin, B., Hardin, J., Hawkins, D., and Rocke, D. 2002. A variance-stabilizing transformation for gene-expression microarray data, *Bioinformatics*, 18:105S-110S.
- Eisen, M., and Brown, P. 1999. DNA arrays for analysis of gene expression, *Methods in Enzymology*, 303:179-205.
- Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C., and Lander, E. 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, 286:531-537.
- Hardin, J., Waddell, M., Page, C., Zhan, F., Barlogie, B., Shaughnessy, Jr., J., and Crowley, J. 2004. Evaluation of multiple models to distinguish closely related forms of disease using DNA microarray data: an application to multiple myeloma, *Statistical Applications in Genetics and Molecular Biology*, 3, article 10.
- Lockhart, D., Dong, H., Bryne, M., Follettie, M., Gallo, M., Chee, M., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays, *Nature Biotechnology*, 14:1675-1680.

- Nguyen, D., and Rocke, D. 2002. Tumor classification by partial least squares using microarray gene expression data, *Bioinformatics*, 18:39–50.
- Nugent, K., Spigelman, A., and Phillips, R. 1993. Life expectancy after colectomy and ileorectal anastomosis for familial adenomatous polyposis, *Diseases of the Colon and Rectum*, 36:1059–1062.
- Pauler, D., Hardin, J., Faulkner, J., LeBlanc, M., and Crowley, J. 2004. Survival analysis with gene expression, Balakrishnan, N., and Rao, C., editors, *Handbook of Statistics 23: Advances in Survival Analysis*. Elsevier.
- Schena, M., Shalon, D., Davis, R., and Brown, P. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*, 270:467–470.
- Storey, J. 2002. A direct approach to false discovery rates, *Journal of the Royal Statistical Society*, B, 64:479–498.
- Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression, *PNAS*, 99:6567–6572.
- Tusher, V., Tibshirani, R., and Chu, G. 2001. Significance analysis of microarrays applied to the ionizing radiation response, *PNAS*, 98:5116–5121.
- Velculescu, V., Zhang, L., Vogelstein, B., and Kinzler, K. 1995. Serial analysis of gene expression, *Science*, 270:484–487.
- Yeung, K., Fraley, C., Murua, A., Raftery, A., and Ruzzo, W. 2001. Model-based clustering and data transformations for gene expression data, *Bioinformatics*, 17:977–987.
- Yeung, K., and Ruzzo, W. 2001. Principal component analysis for clustering gene expression data, *Bioinformatics*, 17:763–774.
- Zhan, F., Hardin, J., Kordsmeier, B., Bumm, K., Zheng, M., Tian, E., Sanderson, R., Yang, Y., Wilson, C., Zangari, M., Anaissie, E., Morris, C., Muwalla, F., vanRhee, F., Fassas, A., Crowley, J., Tricot, G., Barlogie, B., and Shaughnessy, Jr., J. 2002. Global gene expression profiling of multiple myeloma, monoclonal gammopathy of undetermined significance and normal bone marrow plasma cells, *Blood*, pages 1745–1757.
- Zhao, L., Prentice, R., and Breeden, L. 2001. Statistical modeling of large microarray data sets to identify stimulus-response profiles, *PNAS*, 98:5631–5636. ■



FREE

ONLINE **CIS** ACCESS AVAILABLE FOR ASA MEMBERS!

ASA Members can now enjoy free online access to the *Current Index to Statistics (CIS)*. To activate your *CIS* access, log in to ASA Members Only (www.amstat.org/membersonly) and select the *CIS* Web Access tab at the top of the page for instructions.

STATS

Calling All Students!

Play ASA Stat Bowl at JSM 2005 in Minneapolis

Will the Yellow Jackets of Georgia Tech Fly Back to Repeat as Team Champions?



Mark Payton

Are you a student of statistics who would like to experience the Joint Statistical Meetings (JSM) in Minneapolis in August, 2005, but you're too short of cash to get there? We have a solution to your financial troubles. The ASA Stat Bowl will be back at JSM in 2005 and you can earn up to \$500 to cover your travel expenses if you participate. The \$500 is given to all participants regardless of their performance in the Bowl.

Stat Bowl is an individual competition and teams from individual institutions are not needed to play. Team points are kept and a team championship is awarded, but having a team is not a requirement. The event at JSM 2004 in Toronto was a big success, and we're sure it will be just as much fun this year in Minneapolis. A new twist to this year's competition is the addition of prizes above the travel reimbursement. Lynn Eberly, University of Minnesota Biostatistics and ASA Twin Cities Chapter Representative, is working on getting sponsors for additional cash prizes to be given to the top individual contestants and to the members of the winning team.

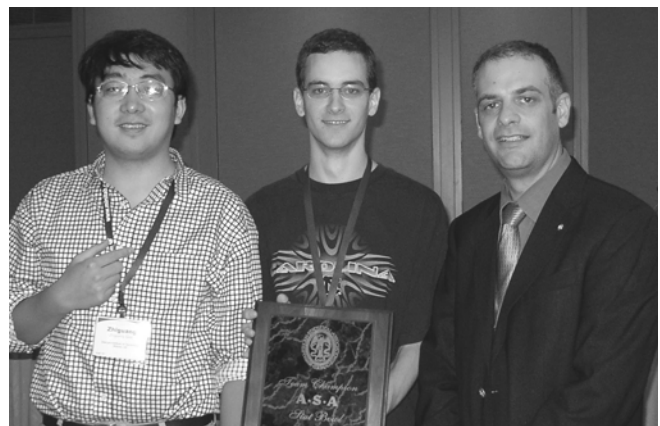
The contest in Toronto was quite entertaining. Jesse Frey from The Ohio State University was the individual champion. He displayed a broad range of knowledge of different statistical topics and history. Andrew Smith from Georgia Tech University also impressed the audience and was awarded the runner-up trophy. Propelled by Andrew's performance, the team from Georgia Tech won the team championship. This is significant in that this was the first time in several years that the team champ wasn't an entry from either the University of Florida or the University of Iowa.

Students will be accepted into the 2005 tournament on a first-come, first-in basis. Notification of a willingness to participate will serve as entry. Inquiries about the Bowl or requests to be registered as a contestant can be made to Mark Payton, Oklahoma State University, mpayton@okstate.edu. A maximum of only 16 players

will be allowed in the contest, so sign up now. In the event that the field of contestants fills to capacity, each university will be restricted to two players to assure diversity. A waiting list will be established to fill unexpected vacancies should they occur at game time.



Mark Payton with 2004 individual champion Jesse Frey of The Ohio State University.



Mark Payton with 2004 team champions from Georgia Tech. From left, Zhiguang Qian and individual runner-up Andrew Smith. Not pictured: Abhyuday Mandal.

(Photos courtesy of Landon Segó, Virginia Technical Institute)

Mark Payton (mpayton@okstate.edu) is a Professor in the Department of Statistics at Oklahoma State University and he has been involved with some incarnation of ASA Stat Bowl for 10 years.







The contest will be held on the Tuesday of JSM in two sessions. Session One, starting at 10:30 a.m., will consist of four games, each with four contestants. The winners of these four games plus two at-large contestants will advance to Session Two. Players who score the most points in Session One without winning their game will be the at-large winners. Session Two, starting at 2:00 p.m., will consist of the six advancing players from Session One playing two games each with three players. Then the two winners will meet head-to-head in a championship game.

Many different rounds make up a Stat Bowl contest. There's the Toss-Up/Bonus round (which consists of bonus questions awarded to players that buzz in and correctly answer a toss-up question), Lightning Round (quick questions that have no bonus questions attached), and a Category Round (players pick a category and receive questions and bonus questions). In years past we've had final questions in each of the games for which the contestants wrote their answer, but we've decided not to include the final question in this year's games.

Most questions focus on the ASA organization and on statistical history and methodology. A statistical history book is a good place to start if you wish to brush up. However, this event is meant to be fun and so a lot of studying beforehand is not what Stat Bowl is all about!

So what are you waiting for? If you are interested in playing, please contact us ASAP before the player positions are filled. We encourage all players to register before July 1, 2005, but players will be allowed in up to game time provided space is available. Hope to see you in Minneapolis! ■

Here are some sample questions:

-  If five cards are drawn at random from a standard deck of 52, what is the probability that the last card drawn is a diamond?
-  Which Sesame Street character is undoubtedly a closet statistician?
-  What major statistician made Rothamstead famous?
-  When the sampling distribution of the estimator is insensitive to changes in the distribution of the population, we say the estimator is ... what?
-  What is the value of the third moment of a standard normal distribution?
-  Give the name of the international society devoted to the mathematical and statistical aspects of biology.

Answers on page 23

Open the door to professional advancement and greater earning potential with SAS!

SAS solutions are used at more than 40,000 sites—including 90% of the Fortune 500. Individuals with SAS skills have an excellent credential to take into today's tough job market!

"SAS Learning Edition definitely has exceeded my expectations. Having come from using SAS in the corporate environment, I know that these skills will equip our graduates to compete for the best jobs in the marketplace."

Matt North
Lecturer, Information Technology Leadership
Washington & Jefferson College

SAS® Learning Edition 2.0

Begin your SAS journey with this personal learning version of the world's leading business intelligence and analytical software. Use the SAS Enterprise Guide® point-and-click interface or write and modify SAS code with the SAS Program Editor. SAS Learning Edition contains a roster of Windows-based products from the SAS Intelligence Architecture platform—all on a low-cost, single CD-ROM. support.sas.com/le

SAS® Self-Paced e-Learning

Advance further in your career and tap into the power of SAS with Self-Paced e-Learning. This highly interactive Web-based training includes questions, quizzes, and guided practices that enable you to learn in an actual SAS environment. License training at the lesson, course, or library level—all at an affordable price. support.sas.com/selfpaced

The Power to Know.



Advice from the 2004 Stat Bowl Champion



Jesse Frey

If you like statistics, enjoy competitions, and wouldn't mind having the ASA cover part of the cost of your trip to the Joint Statistical Meetings, you should consider participating in the Stat Bowl this summer in Minneapolis. The competition is exciting and you'll enjoy the statistical puns offered up by moderator Mark Payton. You may even wind up taking home a plaque as a winner.

I signed up for the 2004 Stat Bowl after reading a notice in *Amstat News*, and I enjoyed every minute of it. While you might imagine that participating in a statistics quiz bowl competition would be a serious and intimidating undertaking, the Stat Bowl is actually best characterized as fun. Payton and his assistants make a steady stream of jokes, often at each other's expense, and even audience members join in at times.

The questions in the Stat Bowl cover all aspects of statistics. As a participant, you'll find yourself solving simple problems in probability or mathematical statistics, answering questions about the lives of well-known statisticians, or identifying the journals where landmark papers first appeared. You might also find yourself trying to identify famous statisticians based on their initials and dates of birth, classifying statisticians according to their countries of origin, or guessing the full names of statistical organizations based on their acronyms. If you think that you might have trouble with some of these questions, rest assured that you're not alone!

A special point of emphasis in the Stat Bowl is the ASA and its organization. There are questions about Chapters, Sections, ASA Presidents past and present, and ASA history. After one lucky guess in last year's Stat Bowl, I for one will never forget that the Chihuahuan Desert Chapter is based in New Mexico!

Jesse Frey (frey@stat.ohio-state.edu) is a fourth-year graduate student in statistics at The Ohio State University. When he isn't working on his PhD dissertation in order statistics, he enjoys playing tennis, running sports prediction contests with statistical twists, and reading classical Chinese novels.

Many of the questions that you'll run into at the Stat Bowl don't lend themselves to being learned at the last minute. For example, spending a few minutes or even a few hours with a textbook is unlikely to help you solve many additional problems from mathematical statistics. On the other hand, memorizing the names of the editors for all the ASA journals and magazines might pay immediate dividends. Interestingly enough, the ASA provides that sort of information right in the conference program!

Should you wish to spend time preparing for the Stat Bowl, one good place to visit is the ASA web site. In addition to being the authoritative source on the ASA itself, it offers a variety of information on the



Jesse Frey in action at Stat Bowl 2004 in Toronto.

history of statistics. A good preparation strategy might also involve spending some time reading a book or two about the history of statistics. The pair of histories authored by Stephen Stigler are classics, and readable biographies of Fisher, Neyman, and others can be found in your university library.

Should you decide to participate in the Stat Bowl, you'll want to familiarize yourself with the rules ahead of time. Don't spend too much time worrying about the penalty for buzzing in early, be prepared for every match to come down to the final question, and save a space in your suitcase for the champion's plaque! ■

The First United States Conference on Teaching Statistics Aims to Build Connections!



Deb Rumsey

Would you like the opportunity to meet and share ideas with other students involved in teaching introductory statistics? Do you want to find/exchange interesting data sets, examples, or in-class activities that get you and your students more involved? Want to get credit on your resume for presenting at a national level conference? Then plan to join us for the First United States Conference on Teaching Statistics (USCOTS) at Ohio State University, May 19-21, 2005.

The goals of USCOTS are to hold a national conference that focuses on undergraduate level statistics education (including AP Statistics); to share ideas, methods, and research results regarding what teachers want to know about teaching statistics; to facilitate teachers incorporating new ideas, methods, and resources into their existing courses and programs; and to promote connections between all teachers of undergraduate level statistics. Graduate students are encouraged to attend and participate!

In addition to all the resources you'll get and the new connections you will make at USCOTS, we want to know what's unique about your statistics teaching experience. You can contribute to USCOTS by participating in one of our Spotlight Sessions.

What is a Spotlight Session? A spotlight session is a "booths, posters and beyond" session that provides a forum for conference participants to display, demonstrate, test drive, and discuss their favorite examples, activities, exercises, methods, labs, to share their experiences and thoughts on statistics teaching and learning, and get others engaged in idea exchange and discussion.

Who can participate in a Spotlight Session? Anyone who is going to the conference is encouraged to contribute to a Spotlight Session. You will be acknowledged in the conference program, and information from your session will be included in the *USCOTS Resource Notebook*, to be given to all conference participants.

Deb Rumsey (rumsey@stat.ohio-state.edu) is a Statistics Education Specialist in the Department of Statistics at The Ohio State University. She is the program chair of USCOTS, and hopes to see many students and instructors in Columbus in May!

We will also be having a "People's Choice Award" for the best contribution to each Spotlight Session.

What are my choices of topics for a Spotlight Session? USCOTS will feature 3 different spotlight sessions:

- Spotlight on Curriculum: "What's on Your Statistics Syllabus, What's Not, and Why?" Choose one or more topics on the introductory or second course syllabus, or discuss a whole course in general.
- Spotlight on Pedagogy: "What's Your Approach to Teaching Statistics, and Why?" For example, how do you get students involved? What resources do you use beyond your textbook? How do you handle the central limit theorem?
- Spotlight on Research: "Share Your Research on Teaching and Learning Statistics." What research have you applied or carried out in your classroom to assess the teaching and learning process to make improvements in your class?

What are some examples of things I could contribute to a Spotlight Session? The idea is to build connections with other teachers, to share ideas, get people involved, and to have fun. Here are some ideas to get you started:



- Tell us about your favorite in-class examples and your students' reaction to them.
- Bring in examples of projects that your students have worked on and how you evaluated them. Have conference participants play games that you use to teach probability.
- Send us on a statistics scavenger hunt.
- Tell us about your best day of teaching statistics, and what made it that way.
- Test out a new activity for teaching sampling distributions that you are considering.

- Show a video of your class involved in an in-class activity that involved teamwork.
- Share your best statistics cartoons, sayings, top 10 lists, etc that you have shared with your class.

To register for USCOTS, visit the USCOTS website at www.causeweb.org/uscots. On the registration form you can sign up for a spotlight session. For more information, contact Deb Rumsey, USCOTS Organizer and Program Chair: rumsey@stat.ohio-state.edu.

USCOTS is hosted by CAUSE, the Consortium for the Advancement of Undergraduate Statistics Education (www.causeweb.org). ■

Special Features of USCOTS

- Plenary Sessions on Curriculum, Pedagogy, Resources, and Research by National Leaders in Undergraduate and AP Statistics Education:



George Cobb
Mt. Holyoke College



Cliff Konold
University of Massachusetts



Robin Lock
St. Lawrence University



Roxy Peck
California Polytechnic
State University, San Luis
Obispo

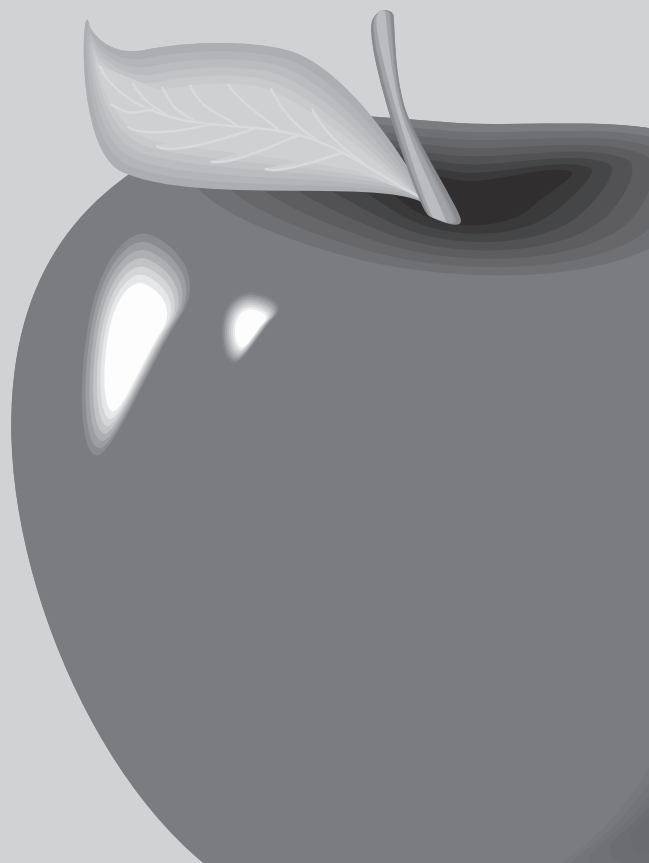


Dick Scheaffer
University of Florida



Ann Watkins
California State University,
Northridge

- Spotlight Sessions exchanging your ideas on teaching and learning statistics.
- Hands-on breakout sessions by leaders in statistics education from all types of institutions and disciplines to help you incorporate new ideas into your courses.
- Opportunities to meet other statistics teachers from a wide range of institutions and disciplines.
- A wonderful Statistics Teaching Resource Packet to take with you.
- A fun, active atmosphere where everyone can be involved!



Sponsor your students' ASA MEMBERSHIP!

For only **\$10** each of your students gets one year of ASA membership!

Students can become members of the American Statistical Association at the special rate of **\$10 for one year** or **\$20 for two years** of membership and only \$25 per year thereafter.

New Students Receive: A subscription to *Amstat News* and *STATS: The Magazine for Students of Statistics*; discounts on all ASA publications, journals, meetings, and products; access to job listings, career advice, and networking opportunities to increase their knowledge and start planning for their future in statistics

SPONSOR INFORMATION

Member ID _____

Organization/Department _____

First Name _____ Last Name _____

Address _____

City _____ State/Province _____

Zip/Postal Code _____ Country _____

Phone _____ Email _____

STUDENT MEMBERS (Attach additional pages if necessary or email student contact information to asainfo@amstat.org)

Name _____

Name _____

Address _____

Address _____

City _____

City _____

State/Province _____

State/Province _____

Country _____ Zip/Postal Code _____

Country _____ Zip/Postal Code _____

Phone _____ Email _____

Phone _____ Email _____

Name _____

Name _____

Address _____

Address _____

City _____

City _____

State/Province _____

State/Province _____

Country _____ Zip/Postal Code _____

Country _____ Zip/Postal Code _____

Phone _____ Email _____

Phone _____ Email _____

PAYMENT INFORMATION: **Total:** \$10/1 yr. x _____ (# of students) and/or \$20/2 yrs. x _____ (# of students) = \$ _____

Check/money order (payable to American Statistical Association in U.S. dollars drawn on U.S. bank)

Credit Card VISA Mastercard American Express

Name on Card _____

Card Number _____ CVS # (3 digit # on back of card) _____ Exp. Date _____/_____/_____

Signature of Cardholder _____

STATS

MAIL: American Statistical Association, Dept. 79081, Baltimore, MD 21279-0081

FAX: (703) 684-2037 **CALL:** 1 (888) 231-3473

Cluster Sampling:

High Quality Information at a Bargain Basement Price



Peter Flanagan-Hyde

Students readily come to understand that random sampling is the “gold standard” of sampling methods. The foundation of all sampling is the idea of a simple random sample (SRS), in which each group of size n in the population has the same chance of being chosen to be the sample. An SRS will provide an unbiased estimate of a population parameter and has predictable sampling variability. This allows the results of a single sample, through a confidence interval, to provide a range of plausible values for a population parameter. So, why would anyone do sampling any differently? What’s the motivation for stratified sampling, cluster sampling, and more complicated multistage samples?

In sampling, there are always two competing interests. The first of these is the obvious goal to get as much information about a population as possible so you can make accurate estimates of the characteristics of the population. The second interest in sampling, often more hidden, is to gather this information at a low cost, expressed either in financial terms or in terms of the effort and time required to complete your sampling. The cost of sampling is often given less consideration in introductory statistics courses. Perhaps this is done to allow students to focus on the very important concepts of bias, sampling variability, and precision. However, to understand why a researcher would adopt a sampling strategy other than an SRS, the cost of carrying out the sample must be considered.

Here’s a setting that can illustrate these ideas. Suppose 1,800 first-year students at a two-year college are enrolled in math classes. The school offers a

Peter Flanagan-Hyde (pflanagan@pcds.org) has been a math teacher for 27 years, the most recent 15 in Phoenix, Arizona. With a BA from Williams College and an MA from Teachers College, Columbia University, he has pursued a variety of professional interests, including geometry, calculus, physics, and the use of technology in education. Peter has taught AP Statistics since its inception in the 1996-97 school year. He has conducted numerous workshops and summer institutes in statistics, presented at a variety of conferences, and authored several sets of student activities.

wide range of courses from remedial algebra through trigonometry, calculus, and differential equations. The student newspaper is doing a story on the curriculum choices of the student body and would like to know the average SAT math score of first-year students. Despite the journalists’ requests, the administration won’t release this information, so they decide to sample the student body to answer the question. They consider taking a sample of 120 students of whom they will ask, confidentially, their SAT math score.

They could take a simple random sample if they had a listing of all first-year students, but even armed with this list an SRS would be difficult to carry out. The student body is very diverse, spanning a wide range of ages and weekly schedules. It would be difficult and time-consuming to actually make contact with each of the 120 students selected in the sample, since they live all over the large city in which the school is located. In cases like this, it is either logistically impossible to take an SRS from the population or the cost of doing so is prohibitive.

At the opening of school orientation, however, students are randomly assigned to 60 orientation groups of about 30 students each. Wouldn’t it be easier to just ask the students in some of these orientation groups for their SAT scores? The students will all be gathered together for the orientation at a specific time and place, so including all 30 of the students in a given group would be easy enough. This will also happen in the math classes to which they are assigned, so another sampling strategy would be to go to the first meeting of each of several math classes to survey the enrolled students. This is the essence and rationale of cluster sampling: Rely on natural groupings of the members of a population to increase the efficiency with which information is gathered.

But, you might ask, isn’t this just a “convenience” sample? It’s important to distinguish between a convenience sample and the convenience of cluster sampling. In a convenience sample, there is no justification for assuming that a sample is representative of the population. However, in cluster sampling we can conveniently take a sample that can still be representative.

To have confidence in the reliability of cluster sampling, two important issues must be resolved. First, is the estimate unbiased? Second, is the sampling variability predictable? We'll use a simulation to gain some insight into these issues and to illustrate the theoretical basis for cluster sampling.

Let's go back to our hypothetical college, for which I've made up some hypothetical, but realistic, data about the students' SAT math scores, orientation groups, and math courses.

Table 1 shows these values with the students' ID numbers, SAT math scores, assigned orientation groups, and assigned math class course numbers.

Student ID	SAT Math	Orientation Group	Math Classes
1401	560	47	41
661	420	23	5
1419	710	48	57
264	450	9	6
1638	510	55	43
235	530	8	29
758	420	26	3
1332	560	45	34
340	500	12	6
⋮	⋮	⋮	⋮

Table 1. Simulated Math SAT Score Data.

The orientation groups are based on the randomly assigned ID numbers, whereas the math class groupings are based on scores on a placement test that has shown moderate correlation with SAT scores in the past. As is true at many colleges, higher course numbers correspond to the more advanced electives in math.

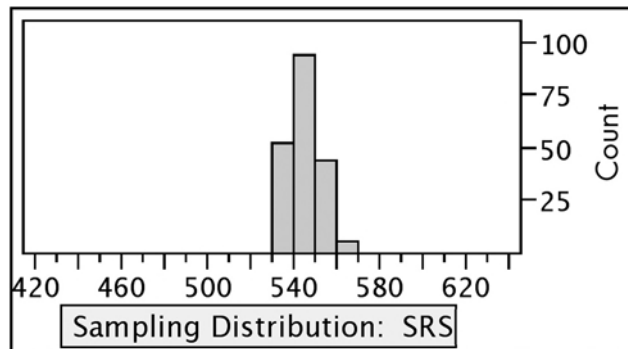
Which grouping—orientation groups or math classes—would be the best to use as clusters and why?

Our standard for evaluating the reliability of the cluster sampling approach will be a simple random sample of the students. Since this is theoretical data, it's easy enough to take an SRS and to repeat the process many times, so that we can have a sense of the sampling distribution for an SRS.

Figure 1 is a histogram and a summary for 200 simple random samples each composed of 120 students.

Let's compare this with the results of sampling with a cluster approach. First, we need to use comparably sized samples. Since the orientation groups are each 30 students, choosing four of the groups, and sampling all students in the group, will make samples of 120 students. But how do we choose the four groups? This is a critical question and the answer is to choose them randomly.

All sampling, if it is to be without bias, should include random selection. Here, the only difference is that we are randomly selecting among the groups rather than among the individuals. If the cluster groups are the same size, each individual's chance of being

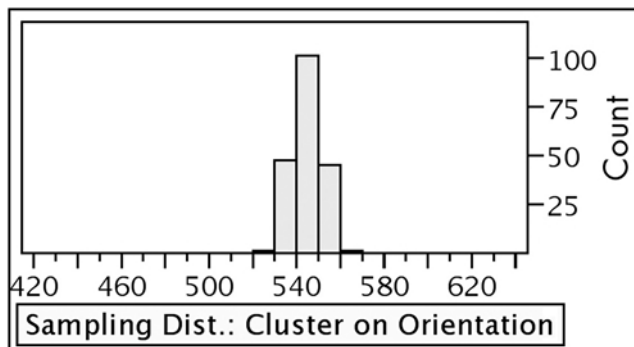


Mean 545.5
Std Dev 7.35
n 120

Figure 1. Two Hundred Simple Random Samples of 120 Students.

included in the sample is the same as for an SRS (sample size/population size, n/N), so, in the long run, each student makes the same contribution to the mean of the sampling distribution as they would in an SRS. This is what makes the method free from bias. If the cluster groups are different in size, though, each individual's chance of being included is not the same, so there is some potential that the estimates made with cluster sampling will not be unbiased. In our simulation, we've idealized this by making the groups all the same size. This lets us focus on the variability in cluster sampling.

Figure 2 shows the results of 200 samples done by randomly choosing four groups from the 60 orientation groups.



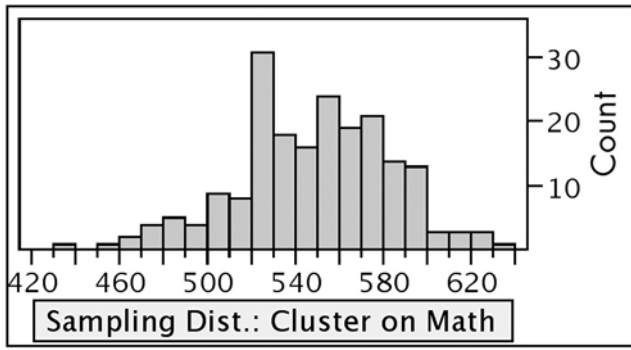
Mean 544.7
Std Dev 7.08
n 120

Figure 2. Two Hundred Samples Using Orientation Groups as Clusters.

The histogram of the sampling distribution is very similar to that of the SRS and the numerical summary indicates that the mean and standard deviations are very close also.

Does cluster sampling always work so neatly? Let's try choosing 200 cluster samples again, but this time

we'll use the math classes as our clusters. Figure 3 shows those results.



Mean	548.5
Std Dev	35.42
n	120

Figure 3. Two Hundred Samples Using Math Class as Clusters.

Wow! That's different! Even noting that the vertical axis has a different scale, the sample-to-sample variability is obviously much larger in this case. Although the mean of this sampling distribution is quite close to the mean of the previous simulation (there isn't any bias), the standard deviation is five times that of the orientation group sample.

Why is this result so different from the previous cluster sampling? The reason that the first method (using the orientation groups) worked so well is that the randomly assigned orientation groups can be thought of as reasonably representative of the entire student population in terms of the variable of interest—SAT math scores. Each orientation group is just as likely to have a very high-scoring student as to have a student who struggles in math. So, on average, each different set of four groups will tend to give a result close to the population average. Figure 4 illustrates the distribution in each of the 60 different orientation groups with the mean marked with a diamond.

This is not the case with the groupings in the math classes. In Figure 5 the composition of the groups is

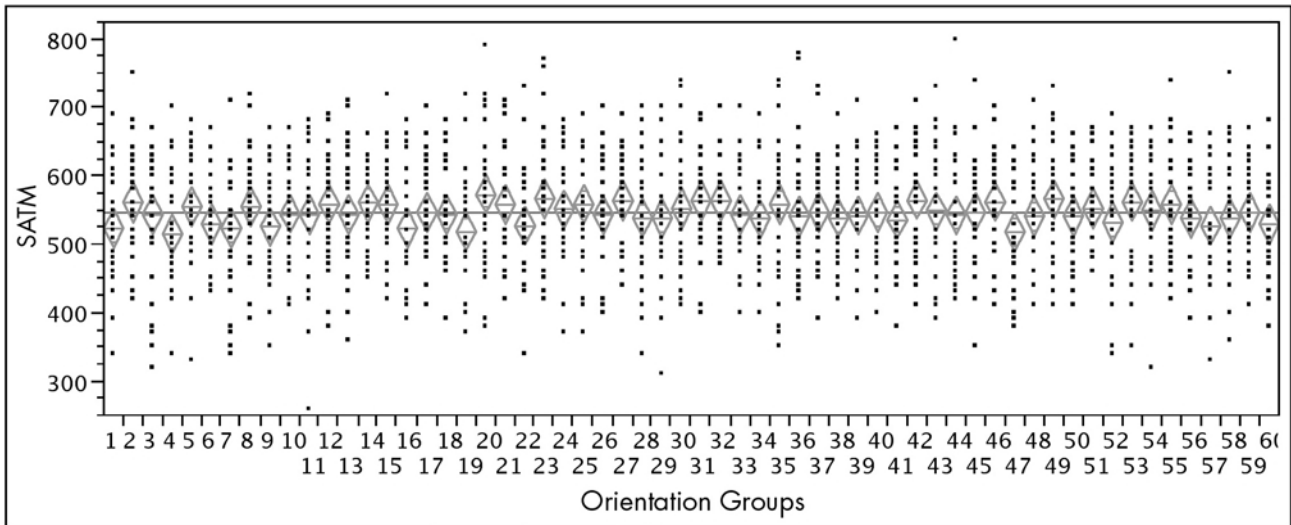


Figure 4. Math SAT Scores by Orientation Group.

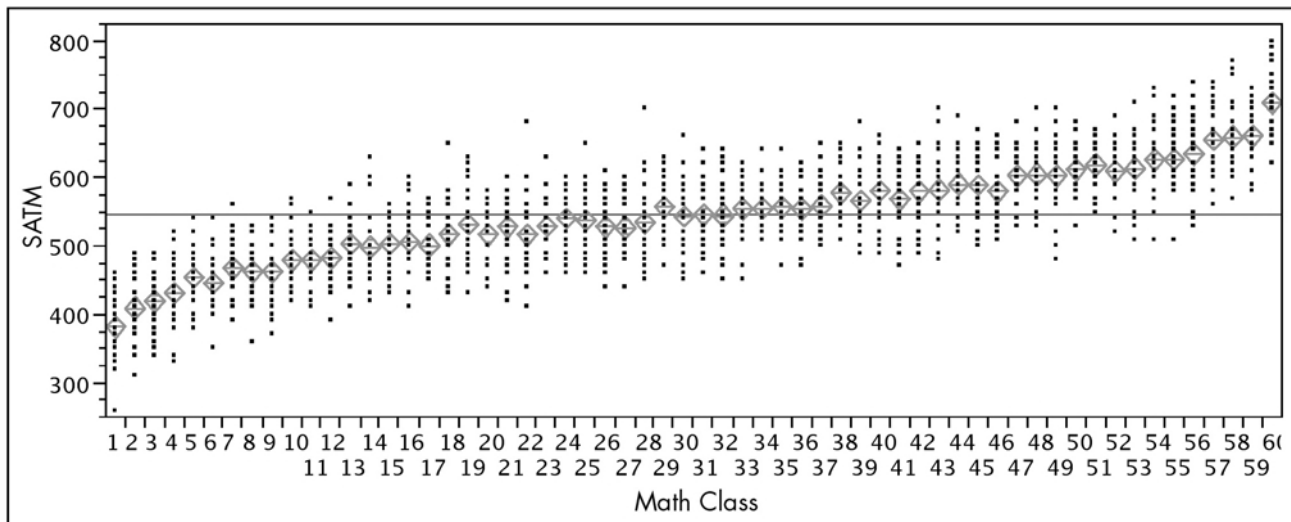


Figure 5. Math SAT Scores by Math Class.

not the same and each group doesn't typically have the complete range of math SAT scores. If we happen to choose several of the more basic math classes in our set of four, the mean SAT math score will be much lower than the population mean. Conversely, if we happen to get a couple of calculus classes and the differential equations class in one group of four, the mean of this sample would be much higher. In Figure 4 you can see that, although the spread of scores in each math class tends to be narrower than in the orientation groups, the mean SAT scores for the math classes varies widely. While the sample means from the orientation groups range from 500 to 600, the sample means from the math classes range from less than 400 to over 700.

Since the students were not randomly assigned to the math classes, the result is greater variability in the sampling distribution that used the math classes as clusters. A theoretical justification for this, complete with estimates of sampling variability, can be found in textbooks on sampling methodology (Scheaffer et al. 1996, for example), but the essence of the argument is illustrated in these graphs.

So, cluster sampling can be an efficient means of gathering high-quality data from a population. It can provide a wealth of information at low cost, providing the clusters are chosen so that the clusters are nearly identical, in terms of both center and variability, for the variable that is being measured. If your population includes natural groups that are reasonably representative

of the population, then cluster sampling is an alternative to consider.

Reference:

Scheaffer, Richard L., William Mendenhall III, R. Lyman Ott. 1996. *Elementary Survey Sampling*, Fifth Edition, California: Duxbury Press: Balmont. ■

Answers to Sample Questions:

- $1/4$
- The Count
- R.A. Fisher. Note that "Fisher" is a very common answer to many of these questions!
- Robust
- 0
- Biometric Society



MEMBER SERVICES

Do you have an address change, membership question, claim, or general inquiry? Please call the ASA's Member Service Toll-free Direct Line 1 (888) 231-3473 for all of your ASA needs. If you prefer, email Member Service at

asainfo@amstat.org

or fax (703) 684-2037.

1 (888) 231-3473

Member Service • Toll-free • Direct Line



STATS

Was Pinkerton Right?

A Data Analysis of an Attempt at Espionage



Chris Olsen

I am heartened by the fact that what I tell my students is sometimes, by a quirk of fate, actually true. For example, as a statistics teacher, I try to inculcate an attitude that statistics can illuminate and help in the solution of problems over a wide variety of fields of endeavor. I further suggest that even a little bit of statistics will go a long way. I must assure the reader that I actually do believe this. (I am not yet totally convinced about the practicality of finding that point in time when two speeding trains are going to be 30 miles apart.) One reason I push this attitude about statistics is that it actually did throw light on an interesting problem I found a few years ago when I was reading Bruce Catton's *Terrible Swift Sword*, a centennial history of the Civil War. In order to set up the problem that my elementary statistical knowledge helped solve, we will need to review a little history about the War Between the States, specifically the Battle of Bull Run fought near Manassas Junction, Virginia, on July 21, 1861.

The battle ended in a victory for the Confederate forces, causing great consternation in Washington. The public fingers of blame soon pointed to people at the top, and President Lincoln deemed it necessary to replace the commanding general at Bull Run with General George McClellan. Lincoln felt McClellan was just the man to transform the Union's post-Bull Run rabble of men into a fighting force. Unfortunately for McClellan, his West Point training had not prepared him for a problem that would lead to his undoing: espionage.

General McClellan contacted Allan Pinkerton, the famous detective, asked him to come to Washington, and gave him two tasks: Find the nests of Confederate spies in Washington and root them out, and gather information about the number and location of Confederate troops in Virginia.

Pinkerton has generally been given some credit by Civil War historians for his anti-spy work, but judgments

about his efforts in the estimation game have been significantly less than laudatory.

Pinkerton's estimates of Confederate troops have largely determined his reputation as a failure at espionage because of the role his estimates seemed to have played in McClellan's decision-making. By fall 1861 members of the Lincoln administration (and the press!) felt that McClellan should be doing something that looked like marching to Richmond, engaging the Rebels, and putting an end to the insurrection. Conventional military wisdom, however, decreed that "enough" men to successfully advance on entrenched forces was three to four times the number of enemy troops. Although McClellan did not know the Confederate troop strength, he suspected the army across the Potomac possessed lots of bullets with his soldiers' names on them, and he was not about to go to battle until he was sure he had enough men.

In point of historical fact, no one knew the correctness (or lack thereof) of McClellan's concerns better than the Confederate generals, whose memoirs began to appear in force in the 1880s along with an encyclopedic collection of war documents known as the *Official Records*. By 1885, legions of authors were busily searching all these documents, and by the 100th anniversary of the onset of the Civil War the general consensus about McClellan was that he had been misled by Pinkerton, who should have stuck to railroad security. The only real historical argument centered on whether McClellan was culpable for believing Pinkerton.

Now, fast forward to the present. Catton writes (p. 271), "[Pinkerton] gets credit, nowadays, for having been worse than he was ... [but his estimates of troop strength] got worse instead of better as time went on...." Hmmm, I wondered, how good or bad was Pinkerton at estimation? This sounded like a problem that could be addressed statistically!

My nearby university library was not helpful in providing answers and the CIA was polite, but firm, in their refusal to let me read their material. The Library of Congress, the National Archives, and the archives of the former Confederate States were the primary sources of information leading me to my conclusions below.

Chris Olsen (colsen@cr.k12.ia.us) teaches mathematics and statistics at George Washington High School in Cedar Rapids, Iowa. He has been teaching statistics in high school for 25 years and has taught AP Statistics since its inception.

Pinkerton's first known report, dated October 4, 1861, and surviving in the McClellan papers at the Library of Congress, contains calculations with a well-defined estimation method in what we today would call a spreadsheet format. He assumed that Confederate infantry regiments consisted of 700 men. Regimental sizes for Maryland and cavalry regiments were assumed to be 600. Basically, Pinkerton identified the numbers of regiments then in Virginia from the several Confederate states and multiplied by the appropriate factors, either 600 or 700. He then subtracted 15% for sickness, giving his final estimate. (It appears that Pinkerton could have used a math teacher for that 15% part. The results of the calculations indicate a subtraction not of 15%, but of 1/15th.) Now we come to the question statistics can help answer: How good was Pinkerton's method of October 4?

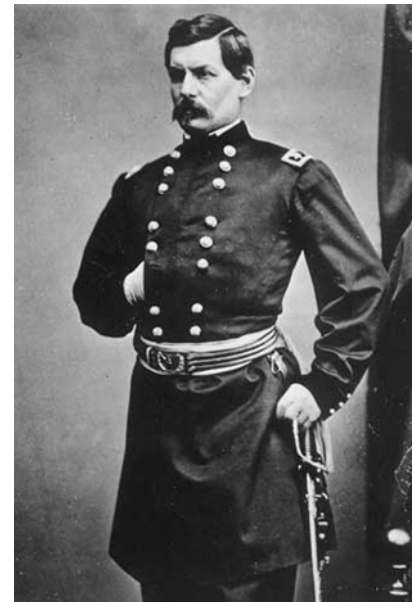
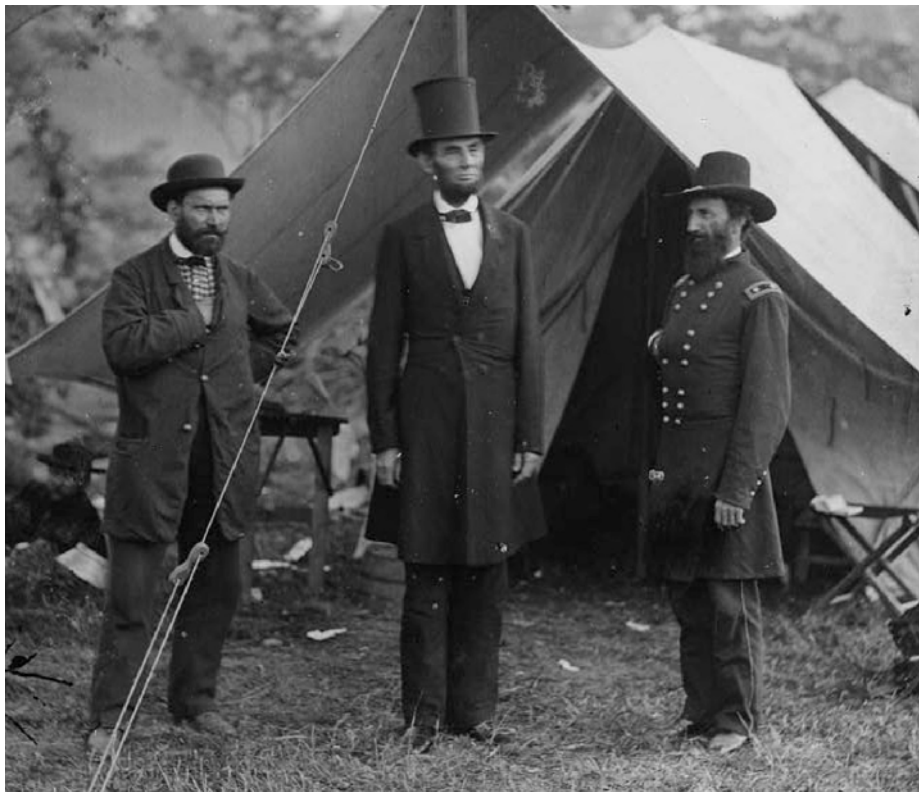
In order to evaluate Pinkerton's method we need to check his assumptions with the actual state of affairs. Data on the numbers of men, as well as numbers of individuals sick, were recorded in bimonthly Confederate company muster rolls during the war. Although somewhat incomplete, these muster rolls were produced for each company of soldiers throughout the war and now reside in the National Archives. In the Confederate army at the beginning of the war, the regiments were almost invariably made up of 10 companies, so the leap from company sizes to regimental sizes is a shift of a decimal point. The data presented below are from an examination of these company muster rolls for the months of August and October of 1861. These rolls were taken at the end of the respective

months, and it seems reasonable to regard Pinkerton's report of October 4 as roughly halfway between these two snapshots of the Confederate regiments. Figure 1 (see page 26) is our first opportunity for a little data analysis to help us.

The distributions of the company sizes reveal approximately normal distributions with a mean company size of 85.05 at the end of August and 82.81 at the end of October. Using 84 as an interpolated value, 84 men per company produces 840 per regiment. Pinkerton's estimate of the regimental size was thus approximately 16.7% low (700 compared to 840). The distributions of the percentage sick are skewed right with mean values of 19.53% and 17.64% for the end of August and October, respectively. Interpolation gives approximately an 18.6% actual sick rate compared to Pinkerton's estimate of 15%. Translated to the regimental sizes, Pinkerton was estimating approximately 85% of 700 men or 595 effective troops per regiment, while the true number appears to have been 81.4% of 840, or 684 men. Thus, Pinkerton's estimate of the number of effective troops in a regiment appears to have been 13% below the true number (see figure 2 on page 26).

To ascertain which Confederate regiments were in Virginia and when, I used the Official Records, regimental histories, individual war memoirs, and other writings I could find in the archives of the Confederate States. One simplifying historical fact, as it turned out, was that once a regiment arrived in Virginia from one of the other Confederate states, it almost invariably did

Continued on page 27



From left, Allan Pinkerton, President Abraham Lincoln, and unidentified Union general. Above, General George B. McClellan, Union Army.

Photo Credits: Library of Congress.

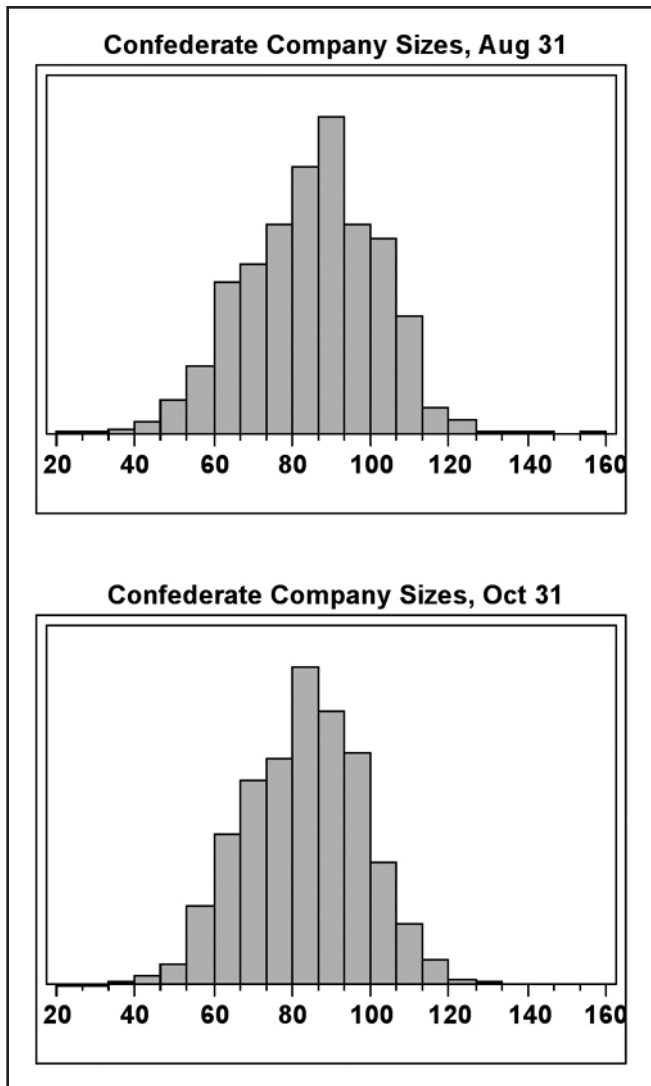


Figure 1. Distributions of Company Sizes (Troops per Company) in the Confederate Army during August and October 1861.

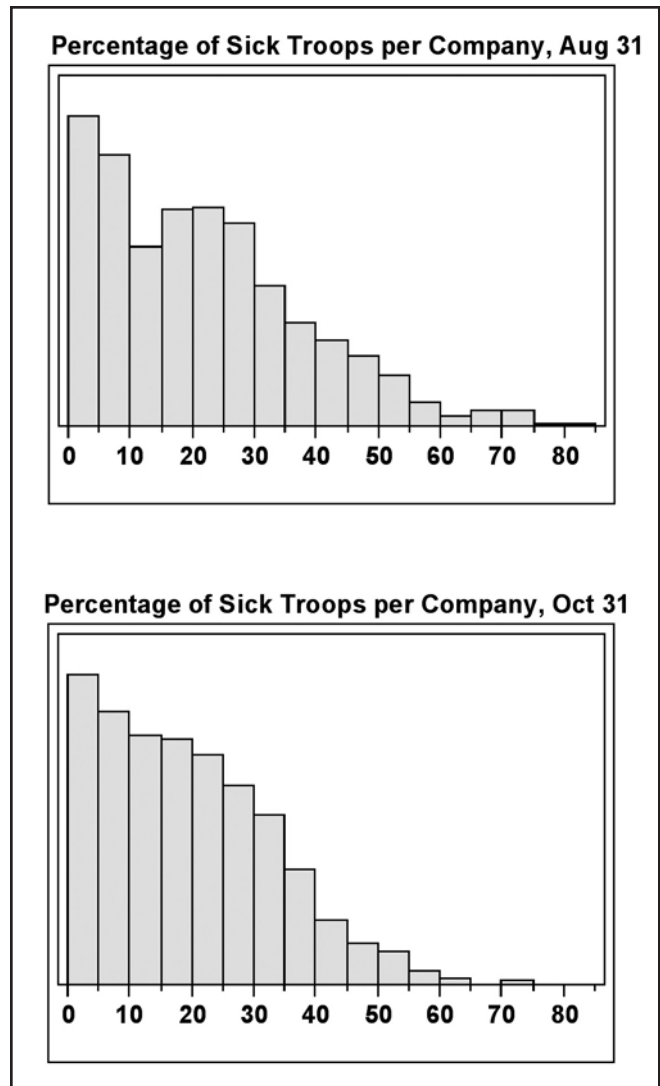


Figure 2. Distributions of Percentage of Sick Troops per Company in the Confederate Army during August and October 1861.

not leave during the time frame we are considering. One Virginia regiment seems to have disappeared into North Carolina and three Tennessee regiments seem to have gone back and forth across the Tennessee-Virginia border a few times, but everyone else stayed put once they arrived in Virginia. The wisdom at the time was that the war would be fought on a Washington-Richmond axis, so the early regiments were mustered into Confederate service after training in Richmond.

For purposes of comparison with Pinkerton's report, whole regiments need to be identified as in Virginia or not. In our calculations, we could assume a regiment to be in Virginia if any company of the regiment had arrived. Is this a reasonable assumption? Perhaps there might be only parts of regiments in Virginia, thus introducing error into the estimation process.

Here again, a little data analysis comes to our aid. It turns out that as companies arrived at Richmond in June and July of 1861, their arrival dates by company

were recorded (Luhn 1992). Even though the data are somewhat incomplete, we can still get a sense of whether the regiments traveled together. The arrival "gaps," i.e., the number of days between the arrivals in Richmond of the first and last companies of a regiment, reveal that in 105 out of 110 cases these gaps were less than seven days, and in 78 out of 110 cases all companies arrived on the same day (see figure 3 on page 27). So, it seems fair to consider a regiment as "in" Virginia as soon as any of its companies have arrived.

We now turn to the question of the numbers of regiments from the several states. After having located the regiments with 20/20 140-year-old hindsight, how do the data compare with Pinkerton's figures?

Table 1 contains the actual vs. Pinkerton's estimates of the number of regiments in Virginia. It seems apparent that for the most part, Pinkerton had a reasonably clear idea about which states were supplying the most and the least regiments, but he overestimated the total

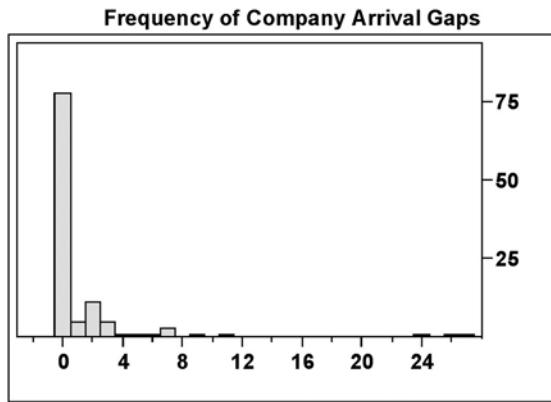


Figure 3. Distribution of the Number of Days Between the Arrival in Virginia of the First and Last Companies of Confederate Regiments.

State	Actual Number	Pinkerton's Number
Alabama	11	18
Arkansas	2	2
Florida	1	2
Georgia	22	18
Kentucky	1	4
Louisiana	11	18
Maryland	1	4
Mississippi	10	12
North Carolina	15	14
South Carolina	9	18
Tennessee	7	8
Texas	3	3
Virginia	50	45

Table 1. Confederate Regiments in Virginia.

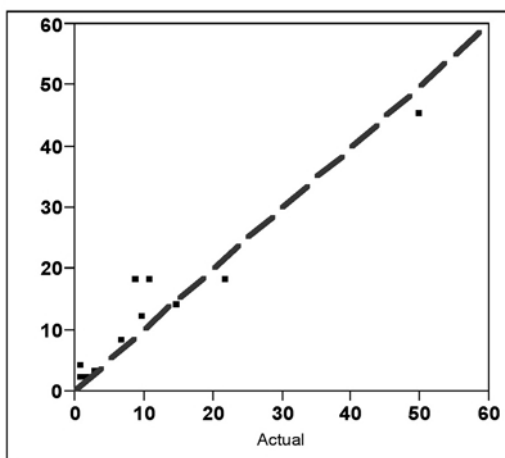


Figure 4. Estimated versus Actual Numbers of Confederate Regiments in Virginia in Late September 1861.

number of regiments by about 16.1%. Figure 4 reveals that much of his error seems to be accounted for by three states estimated as having 18 regiments in Virginia: Alabama, Louisiana, and South Carolina. The dashed line indicates equality between estimate and actual. He seemed to have a tendency to estimate on the high side. Let's look at Table 2.

	Actual	Estimated	% Error
Troops per Regiment	840	700	-16.7%
Proportion of Healthy Troops	0.814	0.850	4.4%
Number of Regiments	143	166	16.1%

Table 2. Estimation Errors.

We are now in a position to evaluate Pinkerton's method of estimating the number of troops in Virginia. Remember, the question is not did Pinkerton's estimation method in his report on October 4 work, but should it have worked as designed? We can perform some elementary calculations to arrive at an overall estimate of Pinkerton's methodological error. Recall that Pinkerton's basic methodology was to multiply the number of troops per regiment times the percentage of healthy troops times the number of regiments to get a total number. How well would multiplying these numbers work?

A number of troops per regiment 16.7% too low multiplied by a proportion of healthy troops 4.4% too high multiplied by a number of regiments 16.1% too high gives an estimate that is astonishingly 1.0097 times actual. In effect, Pinkerton's errors cancel out.

If we remember that the expected value of the product of independent variables is the product of their expectations

$$E[X \cdot Y] = E[X] \cdot E[Y] \text{ if } X \text{ and } Y \text{ are independent,}$$

then we realize that this multiplication is technically correct, if the regimental size and the proportion of healthy troops are independent. The actual correlation between these two values over all companies is estimated from surviving muster rolls to be 0.03, which supports the assumption of independence.

So, if one may be permitted some dramatic license: Pinkerton's method was right: his estimate could have been within 1% of actual.

Circumspection does require that we be somewhat cautionary in our praise. First of all, Pinkerton or his clerk was rather sloppy in subtracting 1/15th rather than 15% in estimating the proportion of healthy troops. Second, although in his methodology Pinkerton treats his estimate of 700 per regiment as an "average," in

the text of his actual report he seems to assert that 700 was a “maximum” number, suggesting there was some upward bias built into his estimates. However, to give a “maximum” estimate he would have needed a way to estimate the variances of each component in his method as well as their means.

So, an analysis of the data using elementary methods has helped us reevaluate the Civil War contribution of Alan Pinkerton. Expert opinion (Fishel 1996) holds that after Pinkerton’s report of October 4, McClellan importuned Pinkerton in favor of some method that would give higher numbers. If that is true, the blame for the steady decrease in accuracy of Pinkerton’s estimates while serving under McClellan would seem to rest with the General himself.

One thing seems very clear: Pinkerton’s reputation as an incompetent agent of espionage is seriously undeserved. And perhaps we can conclude that in statistical estimation, even a simple yet rational model can give surprisingly good estimates: just do the math right!

Bibliography

- Catton, B. 1963. *Terrible Swift Sword*. Garden City, New Jersey: Doubleday.
- Fishel, E. 1996. Pinkerton and McClellan: who deceived whom? *Civil War History*, Vol XXXIV, No 2, 1988 reprinted in Fishel, E. *The Secret War for the Union: The Untold Story of Military Intelligence in the Civil War*. Boston: Houghton-Mifflin.
- Government Printing Office. 1880-1901, *The War of the Rebellion: A Compilation of the Official Records of the Union and Confederate Armies*, 128 vols.
- Luhn, E. R. Appendix V, Company Arrival Roster *Luhn’s Edition, CS Army Special Orders 1861*. Newville, PA. Appendix V, Company Arrival Roster: Civil War Source Book Publishers.
- Pinkerton, A. In McClellan papers, #6013 - #6016, Library of Congress. To preserve his identity, his correspondence to McClellan is signed, “E. J. Allen.” ■



Are You a Student Majoring in Statistics?

First-time Student
Members Pay

\$10

/year

Become a student member of the American Statistical Association! For a special rate of \$10 for each of your first two years and only \$25 per year thereafter, you can join the premier statistical organization in the United States. With your membership you will receive member discounts on all meetings and publications, as well as access to job listings and career advice. You will also enjoy networking opportunities to increase your knowledge and start planning for your future in statistics.

Join NOW!

To request a membership guide and an application, call 1 (888) 231-3473 or join online now at www.amstat.org/join.html.



STATS

Circle the desired size & color for each selection (please indicate alternate color selection in case your first choice is out of stock). Allow 6-8 weeks for delivery of your items. Please call customer service at 1(888) 231-3473 for any questions.



Nephew Alex, and nieces Emily and Hannah of Carolyn Kesner, ASA Development and Grants Manager

Adult Shirts	Sizes Available	Price	Quantity	Total
Fleece Pullover with ASA Logo (SHIRT-FLEECE) Navy	M L XL	\$45.00	X _____ = _____	
Denim Shirt with embroidered ASA Logo (SHIRT-DENIM) Denim	S M L XL XXL	\$35.00	X _____ = _____	
Polo Shirt with "ASA" (SHIRT-POLO2) White	S M L XL XXL	\$35.00	X _____ = _____	
"Got Data?" (SHIRT-DATA) Black White	M L XL XXL	\$20.00	X _____ = _____	
"Top 10 Reasons to be a Statistician" (SHIRT-TOP10) Navy White	M L XL XXL Alternate Color(s): _____	\$20.00	X _____ = _____	
"I'm Statistically Significant" (SHIRT-STAT-A) Navy Dark Red Steel Grey	M L XL XXL Alternate Color(s): _____	\$20.00	X _____ = _____	
"The Evolution of Statistics" (SHIRT-EVOLVE) White	M L XL XXL	\$20.00	X _____ = _____	
"What Part of Normal" (SHIRT-WHAT) Dark Red Steel Grey	M L XL XXL Alternate Color(s): _____	\$20.00	X _____ = _____	
"Approximately Normal" (SHIRT-APPROX) Denim Blue Yellow Steel Grey	M L XL XXL Alternate Color(s): _____	\$20.00	X _____ = _____	
"In God We Trust...All Others Bring Data" (SHIRT-INGOD) Denim Blue Black Hot Pink	M L XL XXL Alternate Color(s): _____	\$20.00	X _____ = _____	
"Absence of Evidence" (SHIRT-ABSEN) White	M L XL XXL	\$20.00	X _____ = _____	

Youth/Child T-Shirts

	Sizes Available	Price	Quantity	Total
"Future Statistician" (SHIRT-FUTURE) Navy Red White	Alternate Color(s): _____ Toddler: 2T 3T 4T Youth: S M L	\$10.00	X _____ = _____	
"I'm Statistically Significant" (SHIRT-STAT-C) Toddler: Lt Blue Lt Pink Lt Yellow Steel Grey Youth: Red Forest Steel Grey	Alternate Color(s): _____ Toddler: 2T 3T 4T Youth: S M L	\$10.00	X _____ = _____	
"Dependent Variable" (SHIRT-DEPEND) Toddler: Lt Blue Lt Pink Lt Yellow Steel Grey Youth: Red Forest Steel Grey	Alternate Color(s): _____ Toddler: 2T 3T 4T Youth: S M L	\$10.00	X _____ = _____	

Logo Items

Denim Baseball Cap w/ Khaki Brim & ASA logo (HAT-DENIM)	\$15.00	X _____ = _____
Stainless Steel Travel Mug w/ASA Logo (STEELMUG)	\$10.00	X _____ = _____
Silvertone Chrome Keychain w/ASA Logo (SKEYCHAIN)	\$8.00	X _____ = _____
Static Cling Car Window Decal w/ASA Logo (CARDECAL)	\$1.00	X _____ = _____

SPECIAL CLEARANCE ITEMS

Off-white Polo Shirt with embroidered ASA Logo (SHIRT-POLO)	S M	\$25.00	X _____ = _____
2003 JSM San Francisco (SHIRT-JSM03)	M L XL XXL	\$10.00	X _____ = _____
2004 JSM Toronto (SHIRT-JSM04)	M L XL XXL	\$10.00	X _____ = _____

Orders must be prepaid. Send order form and payment to:

ASA Souvenirs

American Statistical Association
1429 Duke Street, Alexandria, VA 22314-3415 USA
or fax to: (703) 684-2037. For more information, call 1 (888) 231-3473

Please make remittance payable in U.S. currency drawn on a U.S. bank.

Subtotal	\$ _____
VA residents add 5%	\$ _____
Postage & Handling	\$ _____
(See postage chart below)	
TOTAL	\$ _____

PAYMENT INFORMATION

Check or money order enclosed for \$ _____, made payable to ASA.

VISA MasterCard American Express for \$ _____

Card Number: _____ CVS# (3-digit number on back of card) _____ Expiration Date: ____/____

Name of Cardholder: _____

Cardholder's Signature: _____

ASA ID#: _____ Telephone # _____

Email: _____ Fax # _____

Ship to: Name _____

Address (No PO Boxes) _____

City: _____ State/ZIP or Postal Code/Country _____

POSTAGE & HANDLING CHART

For U.S. Orders Add:

Up to \$10	\$3.00
\$10.01-\$25	\$5.00
\$25.01-\$50	\$8.00
\$50.01-\$100	\$11.00
Over \$100	\$15.00
CANADA	\$20.00

INTERNATIONAL & EXPRESS SHIPPING
Is available for the actual Shipping cost plus a \$3 handling charge. Please call Customer Service at 1(888) 231-3473 to receive an estimate for your items.

The MOST Cited Math Sciences Journal in the WORLD

It's true!

JASA is the most cited mathematical sciences journal in the WORLD for the period 1991–2001. *

JASA had 50% more citations than any other journal.

Subscribe to *JASA*
and receive four issues per year
at the member rate of \$45 a year
(includes free online access)

**Subscribe
TODAY!**

* Source: ISI Essential Science Indicators, Science Watch, Vital Statistics on the Numbers Game, May/June 2002.

To subscribe to *JASA* today, return this advertisement with payment.

- YES!** I would like to subscribe to *JASA* for one year (four issues per year) at the member rate of \$45/year.
My member ID Number is _____
- I am not a member. Enclosed is my payment for \$480.
 I would like to join ASA and subscribe to *JASA* for \$130/year.

Name _____

Affiliation _____

Address _____

City _____ Country _____ State/Province _____ Zip/Postal Code _____

- Check/money order enclosed. (Make payable to the ASA in U.S. funds drawn on a U.S. bank.)
 MasterCard VISA American Express

Name on card _____

Card No. _____ CVS # (3 digit # on reverse of card) _____ Expiration Date _____

Signature _____

ONLINE : www.amstat.org/publications/index.html

FAX orders: (703) 684-2037 (please include credit card information)

MAIL order form with payment to: American Statistical Association, Dept. 79081, Baltimore, MD 21279-0081

CALL Customer Service at 1-888-231-3473