STATS

THE MAGAZINE FOR STUDENTS OF STATISTICS : : ISSUE 47

Christine M. Anderson-Cook interviews Sallie Keller-McNulty

Peter Flanagan-Hyde asks Can You See the Trees for the Forest?

WHICH CAME FIRST THE CHICKEN OR THE EGG?

Looking for a JOB?

Your career as a statistician is important to the ASA, and we are here to help you realize your professional goals.

The ASA JobWeb is a targeted job database and résumé-posting service that will help you take advantage of valuable resources and opportunities. Check out the many services available from the ASA JobWeb.

VIEW ALL JOBS…Search by keyword, job category, type of job, job level, state/country location, job posting date, and date range of job posting.

ADVANCED SEARCH…Use multiple search criteria for more targeted results.

MAINTAIN A PERSONAL ACCOUNT…Manage your job search, update your profile, and edit your résumé. (ASA members only)

USE A PERSONAL SEARCH AGENT…Receive email notification when new jobs match your criteria. (ASA members only)

ADVERTISE YOUR RÉSUMÉ…Post a confidential profile so employers can find you. Registered job seekers can submit their résumés to the résumé database in a "public" (full résumé and contact information) or "confidential" (identity and contact information withheld) capacity. A confidential submission means only an employer can contact the applicant using a "blind" email. (ASA members only)

http://jobs.amstat.org

Visit the ASA JobWeb online **TODAY!**

STATS contents

> "If we are going to make a mark, it may as well be a meaningful one. The simplest—and most useful—meaningful mark is a digit."
>
> John W. Tukey, *Statistical Papers in Honor of George W. Snedecor, 1972*

features

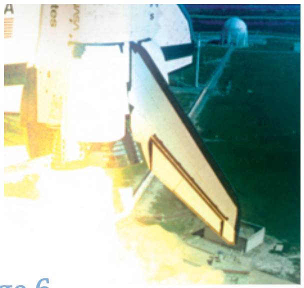
SPECIAL TOPICS

COLUMNS

page 9

puzzles

STATISTICAL SHORTS



page 6

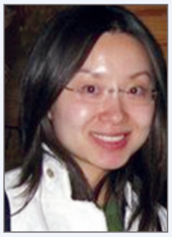
guest writers

An Interview with Sallie Keller-McNulty

Christine Anderson-Cook is a technical staff member of the Statistical Sciences Group at Los Alamos National Laboratory. She is a Fellow of the American Statistical Association and a senior member of the American Society for Quality. She is the current chair of the ASA Section on Quality and Productivity. She graduated with a PhD in statistics from the University of Waterloo and was a faculty member in the Department of Statistics at Virginia Tech. Her research interests include design and analysis of experiments, response surface methodology, graphical methods, and reliability.

Which Came First, the Chicken or the Egg?

Juana Sanchez is a lecturer in the Department of Statistics at the University of California, Los Angeles (UCLA). She has been teaching there since 1997. Her research interests span statistics education, time series analysis, and Bayesian statistics. She particularly enjoys working on research projects with undergraduate students.



Jean Wang works as a statistician at the Mayo Clinic. She received a bachelor's of science in statistics from UCLA in December 2005. The statistics program at UCLA is only two years old, so Wang is one of the first graduates of the program.

EDITOR
Paul J. Fields
Department of Statistics
Brigham Young University, Provo, UT 84602
pjfields@stat.byu.edu

DIRECTOR OF PROGRAMS
Martha Aliaga
American Statistical Association
732 North Washington Street, Alexandria, VA 22314-1943
martha@amstat.org

PRODUCTION/DESIGN
Megan Murphy
Communications Manager
American Statistical Association

Val Snider
Publications Coordinator
American Statistical Association

Colby Johnson • Lidia Vygãzõ
Graphic Designers/Production Coordinators
American Statistical Association

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, 732 North Washington Street, Alexandria, VA 22314-1943** USA; (703) 684-1221; *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 provided to all student members of the ASA at no additional cost. Subscription rates for others: $15 a year for ASA members; $20 a year for nonmembers; $25 for libraries.

Ideas for feature articles and materials for departments should be sent to Editor Paul J. Fields at the address listed above. Material must be sent as a Microsoft Word document. Accompanying artwork will be accepted in graphics format only (.jpg, 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 the ASA office.

EDITOR'S COLUMN

With this issue, we are taking *STATS* to the next level. This is our first issue in full color, and it is jam-packed with some great articles.

We start with an interview with Sallie Keller-McNulty, past president of the American Statistical Association (ASA). She discusses how she became a statistician and what she sees as some of the exciting opportunities for statisticians today and tomorrow. She also provides sage advice for students on preparing for a career in statistics. Thank you, Sallie. We are honored to have you lead off the new *STATS*, and we appreciate your insights and advice.

Next, in our AP Statistics department, Peter Flanagan-Hyde asks, "Can you see the trees for the forest?" With just a little algebra, he shows us how to calculate the variance without calculating the mean. Is that possible? He clears away the forest so we can see the trees.

As we are considering some of the classic questions, Juana Sanchez and Jean Wang ask the proverbial question: "Which came first, the chicken or the egg?" They show us how to answer that question with statistics using time series analysis. See if you can do the analysis with them.

The *STATS* puzzler always has an eye-opening puzzle for us, and this issue's puzzle is no exception. He asks, "How high can the correlation coefficient go?" Well, it can go as high as +1, right? Check out his puzzle and see what you can discover about correlation. There is always something new to learn in *STATS*.

For fun, we have a different kind of statistical puzzle in this issue. Undoubtedly, you have tried your hand at Sudoku—hasn't everyone? Well, turn to Page 8 and try STAT•DOKU. Have fun, and then see if you can answer the statistical questions at the end. Send us your answers to qualify to win an ASA T-shirt.

Antonio Curtis—a student at California State University, East Bay—is a guest author with Bruce Trumbo for our R U Simulating? column. They are using the nonparametric bootstrap technique to look at shrinking students and poisoned children. Thanks, Antonio, for the neat stuff! Remember that if you send in the correct answers to the R U Simulating? challenges, you could win your choice of ASA T-shirt.

Chris Olsen is μ-sing about the role of statistics in the intriguing field of forensics. It is fascinating how the criminal mind works, but it is even more fascinating how the statistically trained mind works.

Also new in this issue of *STATS* is an additional reading department where you can find the references for each article and other resources to help you as you study statistics.

As you think about the world around you and how statistics can help you better understand your world, write up an article for *STATS* to share your thinking with statistics students everywhere.

Paul J. Fields

An Interview with Sallie Keller-McNulty

by Christine M. Anderson-Cook

At the Joint Research Conference on Statistics in Quality, Industry, and Technology—held in Knoxville, Tennessee, last June—I met with longtime friend and colleague Sallie Keller-McNulty. I asked if she would be willing to share her thoughts and insights about statistics with *STATS*. She enthusiastically agreed, and here is what she said.

SALLIE KELLER-MCNULTY is a Fellow and past president of the American Statistical Association. Her research focuses on computational statistics and visualization, statistical modeling, data access and confidentiality, massive dataset analysis, environmental statistics, and sampling. She is dean of the George R. Brown School of Engineering at Rice University in Houston, Texas. Before joining Rice University in July 2005, she was the group leader of the Statistical Sciences Group at Los Alamos National Laboratory.

CAC: ***Can you tell us a bit about how you decided to become a statistician and what part of your first exposure to statistics made it so appealing?***

SKM: I was finishing up all my coursework for a PhD in mathematics at the University of South Florida, and had finished all my exams, when I took my first statistics course. I fell in love with statistics and decided it was what I wanted to do.

It was actually a design of experiments class, which was theoretical combinatorial design—the very mathematical kind with factors labeled with As and Bs and levels indicated with +1s and –1s. For whatever reason, I just loved it. When I talked about this with my advisor, he said, "If you really want to study statistics, we need to prepare you to go to one of the major statistics programs." So, I decided to do a master's thesis with him.

The idea was for me to learn a lot more about statistics. I did a thesis on robust permutation tests. (See the Statistical Snapshot on Page 6.) I learned a lot about statistical testing and even more about computational statistics. We ended up doing a nice piece of work that eventually got published. He then recommended that I apply to some of the leading statistics programs. I was accepted at Iowa State and decided to join their program. I ended up working in the area of statistical computing.

CAC: ***You spent a number of years working at Los Alamos National Laboratory (LANL), where I had the privilege of working with you. Something you told me during my job interview was how important the work done there is to the nation. Do you feel this is something more statisticians should strive for—doing important work?***

SKM: I think that statistical science, as a discipline, has everything to do with data. We are driven by and grounded in data. By data, I am really talking about all sorts of information—not just numbers as we sometimes see it as students, but all kinds of soft and hard information that needs to come together to solve problems. All of that is data. Every global challenge we are confronted with today, at its root, is related to data, to the integration of different sources of information, and to guiding policy based on data.

We are the science of "empirical studies." It is not necessarily the data of a single experiment for some little phenomena that should be our focus, but it is about looking at all kinds of information swirling around us. Pick up the newspaper, look on the internet, look at the satellites. All of these make up the data of our modern world. How does statistics help this come together, and how can it help with today's problems? That is the domain of the statistician. And so, if you start to think about that, it means we need to be more visible, we need to be more engaged, and we need to be more in the leadership of finding solutions for today's grand challenges.

The biologist knows about biology. The statistician talks to the biologist and tries to understand his or her data and help define the empirical methods that will be the key to the biologist's discoveries. This is not to say that we should not know something about a lot of other areas of science, technology, engineering, social sciences, humanities, and the arts. It is obviously necessary for us to know something about these areas, but at the end of the day, our science is about data, and that is what we need to really know.

I learned from my time at LANL that if we are not engaged with some of the important global challenges in the world today—everything from world hunger to literacy to security—then we will be disappointing society. We have something to offer. We need to acknowledge this and join in the ownership of those problems.

> "We are the quintessential interdisciplinary scientists."

CAC: ***Not many statisticians at universities end up working in engineering departments, yet you are the dean of engineering at Rice University. How do you feel your experience with statistics helped prepare you for this role?***

SKM: When I was initially considered by Rice University for the position of dean of engineering, I thought it was rather curious. After being hired, the early publicity revolved around the fact that a woman had been selected to be the dean of engineering, which made Rice the only university with women as dean in both science and engineering. The dean of sciences, Kathleen Matthews, is a leading biochemist. After a while, I realized this publicity was really a smoke screen for the fact that Rice University had hired a statistician to be the dean of engineering.

I should point out that Rice has structured the engineering college to include statistics, computer science, and computational and applied mathematics. As I began to meet my faculty in all the various areas—such as civil engineering, electrical engineering, chemical engineering, bioengineering, environmental engineering, computer engineering, and mechanical engineering and materials science—I discovered huge connections between my background and my new colleagues'. We shared many things that we have been talking about for years in statistics, including being grounded in significant real-world problems, trying to drive toward solutions that can be implemented, being systems thinkers, and being willing to jump into the middle of a complex problem that needs a solution.

We statisticians talk about the scientific method and are willing to keep cycling through the method, iterating toward ever-improving solutions. Engineers are willing to do the same thing. I believe we have a major connection in the way we think and the way we approach the world around us. Since I discovered that, I have felt very much at home in the school of engineering, and I think I bring a sense of new curiosity to the school. I like to tell people, particularly junior colleagues, that the wonderful thing about being a statistician is that no one expects you to know volumes about the other areas of science and engineering, even though we often have some depth of knowledge. Colleagues are quite content to provide the details of the application on which you are collaborating. Quite often, a new set of eyes can challenge the basic assumptions of the problem and raise some very important issues.

The fact that we have been trained to ask questions, to probe, to be inquisitive, to be critical thinkers, and to want to document assumptions is a very strong positive and one of the key aspects that we bring to collaborative research. That's what we do, and as I have come to understand, that is what the engineers do as well.

CAC: ***I know you feel that this is a very exciting time to be a statistician. What opportunities do you see coming, and what do you think statisticians can do to help increase their impact?***

SKM: We have long been told as part of our statistical training that we, as statisticians, are the guardians of the scientific method. It provides a means to develop and test hypotheses, to design studies and perform analyses, and to make inferences about what our experimental data are telling us. This is an iterative process for thinking about problems and cataloguing what we know. But today, the scientific landscape has really changed. Today, we have a lot of information available to us, and we really need to incorporate all sorts of things into solving the complex problems confronting us.

Interdisciplinary science is the future. Interdisciplinary teams, comprised of scientists and engineers from a broad spectrum of disciplines, will tackle the significant problems of our generation. There is no longer a call for the Renaissance man or woman, but rather a focus on the Renaissance team. Innovation will come in the spaces between disciplines. Everything is coming together to help us solve some great challenges, whether it is nanotechnology or the spread of a pandemic or going to Mars. All will require a new kind of interdisciplinary thinking.

I will argue that it is our field—statistics—that understands the value of information and will find creative ways of combining information. We should be the ones emerging as the leaders in this new era of interdisciplinary science. We are the quintessential interdisciplinary scientists, and we need to step up to the plate and lead the integration of science, not simply 'guard' the scientific method.

CAC: ***As someone who worked with you at LANL, I know you are a very inspiring leader. What have you done over the years to work on developing your leadership skills?***

SKM: I have a philosophy about that, which is not necessarily consistent with what others believe. In my mind, there is a real distinction between leadership and management. I really believe people can learn management skills. But leadership is a little different. Leadership is something you either have or do not have, and becoming a leader seems to just happen when you are not looking.

Many people have said to me that I am a great manager, but the truth is, I would be happy to hand that off to someone else. Other people have said to me that I am an inspiring leader, yet I do not necessarily see myself in that same light. I do feel that once you realize you have leadership capacity and realize you are seeing things differently than other people, it is important to work on developing those skills further to broaden the arenas in which you can apply your leadership and develop broader vision.

I have had a wonderfully devoted, eclectic set of mentors throughout my career—people who have both encouraged and inspired me. Over the years, I have taken the time to watch, to listen, and to wonder how in the world they were able to put complex ideas and issues together. Then, I began to realize that my own mind was similarly making some novel connections. Going down this path, I learned how to focus my listening and recognize the value of these insights. The next phases of my leadership development involved continuing to listen to other people, helping them see potential new paths, and synthesizing detailed information to create a larger vision.

One of my managers at LANL was a really interesting person who influenced my leadership development. I learned from him that leaders must have "no fear." They need to surround themselves with the smartest people they can—even people smarter than themselves—and to really listen to what they are saying. If you do this, magic will happen. However, I have observed this is a very hard thing for many people to do.

CAC: ***What skills do you think are helpful for students of statistics to work on or study during their university years?***

SKM: Communication! Every form is important—writing, speaking, and presenting. Being a good team member and a good team leader require communication skills. It is important to be able to communicate complex ideas at a very high general level to capture the most important aspects and communicate them in a compelling way. At the same time, it is important to be able to talk clearly about details with our statistics, science, and engineering colleagues. It is important to have the patience to learn, develop, and grow in all these areas. It is not an easy process, but good communication skills will help your ideas be heard, and they also will clarify your own thinking about important problems.

Fundamentally, statistics is not about data analysis; it's about understanding and supporting decisionmaking. This should not be confused with decision analysis, which is another technical mathematical concept. The decisionmaking I am talking about involves shaping policy and considering the politics of the decision and the personalities involved. To be effective in that arena, it is important to be a perceptive listener and to have a willingness to throw yourself into the middle of a debate and make a strong argument based on a clear interpretation of the information available. Sometimes, it involves mediating; sometimes, it is about being very firm and direct; sometimes, it is figuring out how to be a good team player.

Think about it: Uncertainty quantification is our world. Who in our society better understands the concept of uncertainty? Who in the highest levels of our society and the governance of our society understands the impact and the role of uncertainty? Our job is to find vehicles to communicate what we know about problems that are important to the world. Communication is incredibly important. Learn when you are being effective, and recognize when you are not. Know when to make another attempt at explaining your position, and know when to stop. Know when to get help from people in other areas who are effective at communication.

One of the difficulties with improving communication skills is that when we teach communication, we frequently teach it as a single, isolated course: "Go take a technical writing class so you can learn how to do a presentation or learn how to put your materials together." We need to focus on communication, not as something that we do every Tuesday, but as something we do every day in all aspects of life. Communication needs to be integrated into everything you do, in all your studies and in every mode of your life.

CAC: ***Do you have any other advice for students in statistics that might help shape their studies and early careers?***

SKM: I guess my advice is to look around at the breadth of our field, to meet lots of people, and to engage and start building a network of colleagues. Get to know the people in college with you. These are your future colleagues. Introduce yourself to established statisticians and take advantage of meetings—the Joint Statistical Meetings and other conferences—to network with your peers across the profession. Whether they are in industry, academia, or government does not matter. Build a network of people who know your strengths with whom you can discuss technical matters and turn to for career advice. Another thing that is important is to maintain a curiosity in science, engineering, and everything around you.

People frequently ask me what I look for when I hire somebody. There are three key things. THE FIRST thing I look for is strong training in some area of statistics. If the person has that and is smart, then they are going to be able to learn other aspects of our field. THE SECOND thing I look for is whether they have great computational skills. In statistics today, every aspect of methodological development and applications reaches way beyond the preprogrammed packages that you learn in school. You will need to have some flexibility in terms of being able to develop and implement new methodology, so computational skills are critical. THE THIRD thing I look for is somebody who has an incredible curiosity for science. If they have that, they are going to continue to ask questions, engage, and integrate our field into science, engineering, and society.

So, I encourage all students of statistics to focus on both breadth and depth in their training. Be a "T-person," with good depth in at least one statistical area and breadth across many others. Continue to work on your communication skills and never lose your love of science. That is a combination that will lead to success. ◕

A Permutation Test of the *Challenger* O-Ring Data

The disastrous loss of the *Challenger* space shuttle in 1986 gave rise to what has become one of the classical datasets of statistics. *Challenger* was launched at the unusually low temperature of 29°F, in spite of evidence—judged to be inconclusive beforehand—that the failure of O-rings is associated with low temperatures. Afterward, leakage of fuel around the O-rings was implicated in the explosion of *Challenger*.

There were 24 previous flights. Of those, four were launched when the temperature was below 65°F and 20 were launched at higher temperatures. The number of O-ring incidents on each of the previous flights is given in Table 1. Is there an association between O-ring incidents and temperature?

TABLE 1. Temperature and O-Ring Incidents

Temperature at Launch	Number of O-Ring Incidents
Below 65°F	1, 1, 1, 3
Above 65°F	0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 2

PHOTO courtesy of the NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

VIEW OF THE O-RING. This is a close-up photograph of the O-ring in the top of the aft segment of the right solid rocket booster (SRB) flown on Space Shuttle Mission 51-L. The photograph was released following a hearing on the space shuttle accident.

This is a standard two-sample situation. We want to test the null hypothesis that temperature makes no difference against the alternative that lower temperatures tend to be associated with more O-ring failures. We usually would use the two-sample t-test, but it is not appropriate here because the few observations available do not suggest *normal* populations. Moreover, nonparametric tests—such as the Wilcoxon rank sum test—are not applicable because they assume *continuous* data with few ties, while the O-ring data consist of integers with lots of ties.

Let's take a permutation approach. We can use a *permutation test* to find a p-value by computing the proportion of possible groupings of the observed data that produce a test statistic as extreme—or more extreme—than what we have observed. We proceed by considering all possible arrangements of the 24 observed numbers with four in one group and 20 in the other. The total number of these arrangements is "24 choose 4," or 10,626. For the test statistic, we will use the total number of O-ring failures and compute this test statistic for each arrangement. The most extreme values of the test statistic would occur if the lower temperature group had outcomes 1, 1, 2, 3 (sum of seven) or 0, 1, 2, 3 (sum of six), in addition to the observed 1, 1, 1, 3 (sum of six).

By a combinatorial argument, we can show there are 10 ways to get the first arrangement: Multiply the 10 ways to select two 1s out of five 1s, multiply the one way to select one 2, and multiply by the one way to select one 3. Similarly, there are 85 ways to get 0, 1, 2, 3 and 10 ways to get 1, 1, 1, 3.

This gives a total probability of (10 + 85 + 10) / 10,626 = 0.0099. So the p-value of the permutation test of our hypothesis is .0099, or about 1%. This is fairly strong evidence that low temperature is associated with an increased chance of O-ring failures.

While it is not difficult to find the p-value of the permutation distribution of our test statistic in this simple situation, it can be very difficult to do so for larger datasets. Thus, it is common practice to approximate the permutation distribution by simulation.

Applied to our data, the simulation procedure would be to make random permutations of the 24 observations, take the first four observations in each permutation to be the low-temperature group, and compute the test statistic for the result. Repeating this procedure many times, we take the observed test statistic value to be the simulated p-value. With 100,000 iterations, our result is p-value ≈ 0.0096—very close to the exact value obtained above. ◕

Permutation Distribution Compared with t(22)

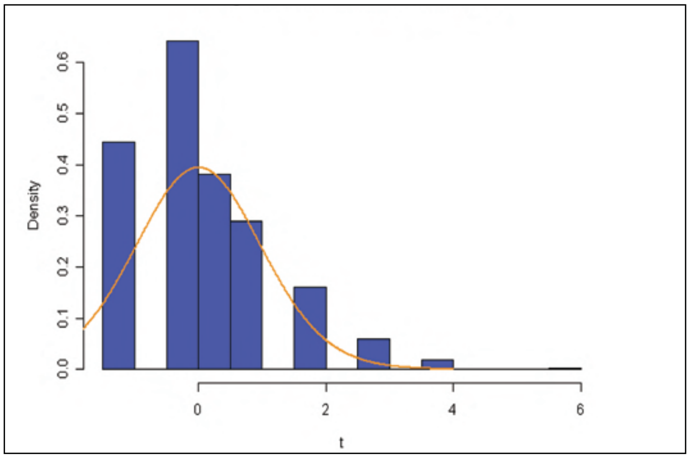


FIGURE 1. Shown above is a histogram of the simulated permutation distribution. The superimposed curve is the density function of the t distribution with 22 degrees of freedom shown for comparison.

Lesson Learned

We can use a permutation test to estimate the p-value with combinatorics or simulation, even when the conditions necessary for other tests are not met.

TRY THIS

STAT•DOKU

The 9 x 9 grid used in Sudoku puzzles is a special type of Latin square, which is an n x n matrix filled by n symbols so each symbol appears only once in each row and column. It was named by Leonard Euler, the great eighteenth-century mathematician. He used Latin characters to fill his squares, hence the name "Latin square."

In Sudoku, the symbols are the numerals 1 through 9, but a Sudoku grid has the added constraint of the nine nonoverlapping 3 x 3 subgrids also containing 1 through 9 without repeating any of the numerals.

Now let's look at STAT•DOKU. The same basic rules apply, but the nine symbols are the letters S-T-A-T-I-S-T-I-C. Some of the letters are used more than once, but every row, column, and subgrid must contain those nine letters in the exact frequency as in the word "statistic."

Here is a STAT•DOKU puzzle. See if you can solve it.

S	T				S	T		C
		C	I			A	T	T
T		I		A	T	I		
	A	S		I	T		I	T
		S		S		I		
I	I		T	T		S	S	
	C	T	S		I			I
A	I	I				T		
S		T	A				C	I

By the way, the solution is shown on Page 25.

After you have solved the puzzle, think about the following questions:

Should the puzzle be easier, or harder, to solve when the symbols are letters, rather than numbers?

What is the effect on the difficulty of a puzzle when the letters are used more than once?

Do you use the same strategies to solve a STAT•DOKU puzzle as a Sudoku puzzle?

How is the statistical concept of "degrees of freedom" related to solving this kind of puzzle?

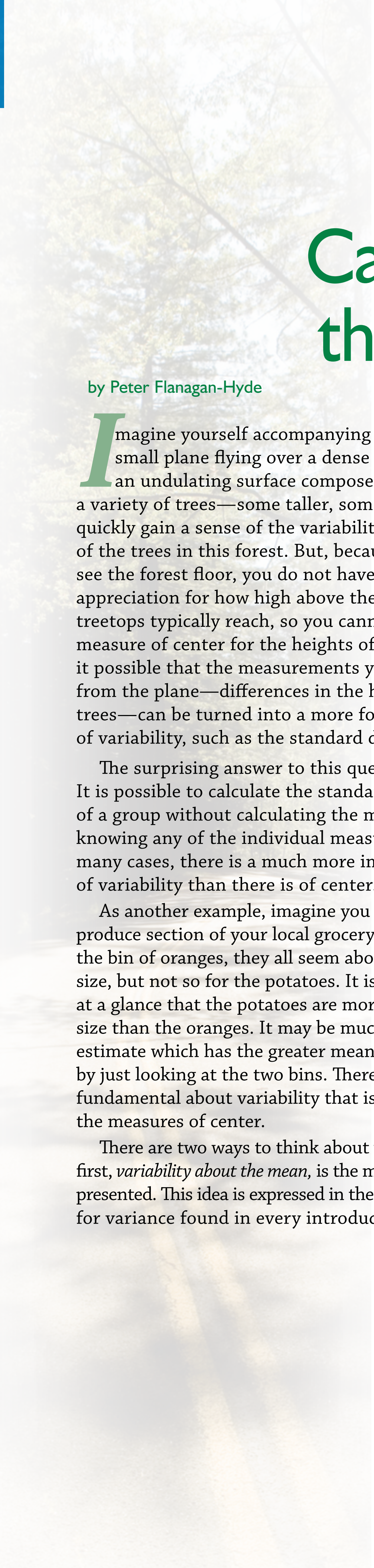
Send your answers to *STATS* Editor Paul J. Fields at *pjfields@byu.edu*; if it's correct, you will be entered into a drawing for an ASA T-shirt.

Win a T-shirt

Two final questions:

Can you find any place in the puzzle where the letters come in the right order to spell "statistic"?

Is this a unique solution, or are there more?



Can You See the TREES for the FOREST?

by Peter Flanagan-Hyde

Imagine yourself accompanying a biologist in a small plane flying over a dense forest. Below is an undulating surface composed of the tops of a variety of trees—some taller, some shorter. You quickly gain a sense of the variability in the heights of the trees in this forest. But, because you cannot see the forest floor, you do not have an immediate appreciation for how high above the ground the treetops typically reach, so you cannot find a measure of center for the heights of the trees. Is it possible that the measurements you can make from the plane—differences in the heights of the trees—can be turned into a more formal measure of variability, such as the standard deviation?

The surprising answer to this question is "Yes!" It is possible to calculate the standard deviation of a group without calculating the mean, or even knowing any of the individual measures. In fact, in many cases, there is a much more immediate sense of variability than there is of center.

As another example, imagine you are in the produce section of your local grocery. Examining the bin of oranges, they all seem about the same size, but not so for the potatoes. It is easy to see at a glance that the potatoes are more variable in size than the oranges. It may be much harder to estimate which has the greater mean circumference by just looking at the two bins. There is something fundamental about variability that is separate from the measures of center.

There are two ways to think about variability. The first, *variability about the mean,* is the most commonly presented. This idea is expressed in the usual formula for variance found in every introductory statistics textbook: The population variance is the mean squared deviation of each measurement from the mean, or $\sigma^2 = \dfrac{\sum_i (x_i - \mu)^2}{n}$.

The second way of thinking about variability is the *variability among the individuals in the group.* In many situations, this is really more natural. At the produce bin, what is obvious about the potatoes is that there is a big difference in size between some of the potatoes sitting next to each other. We really would not imagine a mean size and then compare the potatoes in the bin to this imagined mean. It is more about looking at differences of the form $x_i - x_j$, where the indices *i* and *j* reference different individuals in the population. Taking a cue from the variance formula, does the mean-squared difference among the individuals provide a useful measure of variability? If so, is it related to the variance as calculated above? Let's find out.

A Weighty Example

Let's suppose we have a class of 10 students. We would like to know the variance of their weights, but people's weights are a sensitive subject, so it would be nice to be able to complete this calculation without actually revealing an individual's weight. For now, though, we need these values, so Table 1 shows their weights in pounds:

TABLE 1. Students' Weights Measured in Pounds

Student	1	2	3	4	5	6	7	8	9	10
Weight	95	124	131	121	104	97	116	151	122	139

PETER FLANAGAN-HYDE has been a math teacher for 27 years, and has taught AP Statistics since its inception in the 1996–1997 school year. 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.

By the usual formulas, you can find that the mean weight is 120 pounds and the variance—the mean squared deviation from the mean—is 289 pounds^2. The population standard deviation is 17 pounds.

For this class of students, what is the mean squared *difference* of their weights, and does this relate to any of the numbers we have calculated above? Below is a table of all the squared differences. It is important to note that the interior of the table (in italics) could be calculated easily without weighing any of the students. Imagine a long board balanced on a small rod at its midpoint. One student stands on the left end of the board and another on the right. Because none of the students in the class has the same weight, the lighter one will be off the ground. Hand that student one-pound weights until he or she balances and the difference in weights is found. Square this number and enter it into Table 2. When we are done, notice we have no measures of individual weights and no estimation at all about the mean weight in the class.

Notice there are 100 values in Table 2, one for each ordered pair of students. It is easy enough to calculate the mean of these 100 numbers; it is 578 pounds^2. It is not too hard to see that this is exactly twice the variance. Is this a strange coincidence of this class, or is this always the case?

A Little Algebra

It will take only a little algebra to show that the mean squared difference is always exactly twice the mean squared deviation from the mean. As we are adding a large number of variable terms, this is a good opportunity to practice working with summation notation, too.

To make the notation simpler, we will express everything we do in terms of three quantities: n, the number of individuals; T, the sum (or total) of all the values; and SS, the sum of the squares of all the values. In simplified summation notation, $T=\sum_i x_i$ and $SS=\sum_i x_i^2$. It is worth noting that the mean $\mu=\frac{T}{n}$ and that the index, i, is arbitrary; it's equally true that $T=\sum_j x_j$ and $SS=\sum_j x_j^2$.

One student stands on the left end of the board, and another on the right. Because none of the students in the class has the same weight, the lighter one will be off the ground.

TABLE 2. Squared Pair-Wise Differences in Students' Weights Measured in Pounds

Student	1	2	3	4	5	6	7	8	9	10
1	*0*	*841*	*1296*	*676*	*81*	*4*	*441*	*3136*	*729*	*1936*
2	*841*	*0*	*49*	*9*	*400*	*729*	*64*	*729*	*4*	*225*
3	*1296*	*49*	*0*	*100*	*729*	*1156*	*225*	*400*	*81*	*64*
4	*676*	*9*	*100*	*0*	*289*	*576*	*25*	*900*	*1*	*324*
5	*81*	*400*	*729*	*289*	*0*	*49*	*144*	*2209*	*324*	*1225*
6	*4*	*729*	*1156*	*576*	*49*	*0*	*361*	*2916*	*625*	*1764*
7	*441*	*64*	*225*	*25*	*144*	*361*	*0*	*1225*	*36*	*529*
8	*3136*	*729*	*400*	*900*	*2209*	*2916*	*1225*	*0*	*841*	*144*
9	*729*	*4*	*81*	*1*	*324*	*625*	*36*	*841*	*0*	*289*
10	*1936*	*225*	*64*	*324*	*1225*	*1764*	*529*	*144*	*289*	*0*

Let's start with the usual formula for the variance $\sigma^2 = \frac{\sum_i (x_i - \mu)^2}{n}$ and see if it can be expressed in terms of these three quantities.

The tricky part is the numerator:

Expand the binomial
$$\sum_i (x_i - \mu)^2 = \sum_i (x_i^2 - 2x_i\mu + \mu^2)$$

Break the sum apart and factor constants from sums
$$\sum_i x_i^2 - 2\mu \sum_i x_i + \mu^2 \sum_i 1$$

Substitute for summations
$$SS - 2\mu T + \mu^2 n$$

Substitute for μ
$$SS - 2\frac{T}{n}T + \frac{T^2}{n^2}n$$

Simplify
$$SS - \frac{T^2}{n}$$

Then, the variance can be written as
$$\sigma^2 = \frac{SS - \frac{T^2}{n}}{n} \text{ or } \sigma^2 = \frac{SS}{n} - \frac{T^2}{n^2}.$$

Now, let's tackle the mean of the squared differences of all pairs of individuals. To begin, we will require a double summation to work through both members of the pairs. (Another way to think about this is that we have to work through both the rows and columns of the table above). We start with MSD representing the mean squared difference:

$$MSD = \frac{\sum_i \sum_j (x_i - x_j)^2}{n^2}$$

To simplify the double summation, we start with the inside and work out. Again, we focus on the numerator:

Expand the binomial
$$\sum_i \sum_j (x_i - x_j)^2 = \sum_i \sum_j \left(x_i^2 - 2x_i x_j + x_j^2\right)$$

Distribute the inside sum
$$\sum_i \left[\sum_j x_i^2 - \sum_j 2x_i x_j + \sum_j x_j^2\right]$$

Factors with the other index are constants
$$\sum_i \left[x_i^2 \sum_j 1 - 2x_i \sum_j x_j + \sum_j x_j^2\right]$$

Substitute T and SS
$$\sum_i \left[x_i^2 n - 2x_i T + SS\right]$$

Factor out constants
$$n\sum_i x_i^2 - 2T\sum_i x_i + SS\sum_i 1$$

Substitute
$$n \bullet SS - 2T \bullet T + SS \bullet n$$

Simplify
$$2n \bullet SS - 2T^2$$

Divide by n^2 to get
$$\frac{2n \bullet SS - 2T^2}{n^2}.$$

So, $MSD = 2\frac{SS}{n} - 2\frac{T^2}{n^2}$ $MSD = 2\left(\frac{SS}{n} - \frac{T^2}{n^2}\right) = 2\sigma^2.$

As promised, the mean squared difference is exactly twice the variance.

Back over the Forest…

As you and the biologist are flying over the forest, you can, in fact, make a measurement on a random sample of trees and, through photographic or radar measurements, find the difference in height between them. Adding up the squares of these differences can enable you to make an estimate of the variance and standard deviations of the heights of the trees in the forest. The derivation above can be adapted for sample variances; it is a little messier, but the principle is the same. As with our weighty example, this can be done even if we cannot see the ground (assuming all the trees are standing on level ground). No individual tree's height is known, and no estimate of the mean tree height is made. So, estimates of the mean and variance can be calculated independently. ◕

Using some of the basic concepts of introductory statistics, statisticians Juana Sanchez and Jean Wang attempt to answer the question,

WHICH CAME FIRST THE CHICKEN OR THE EGG?

Chickens play a huge role in our lives. As well as sometimes acting as pets or ending up sitting on plates, chickens show up in pop culture. They had feature roles in the hit movie "Chicken Run"; they frequently guest star in popular games such as "The Legend of Zelda" and "Final Fantasy"; and they take center stage in the classic dilemma of causality: "Which came first, the chicken or the egg?"

Let's use statistics to solve this dilemma. We will not delve into the philosophical issues. For that, you can watch last year's CBS News Video, "Was the Chicken or Egg 1st?" Rather, let's use time series analysis of historical data from the United States chicken industry to shed some light on the direction of causality.

A time series is a sequence of observations that are ordered in time. Some common examples include daily temperatures, weekly stock prices, and monthly employment figures. In a time series, the value of a variable for today often depends on its value in the past. In effect, the variable has some memory of its past and, perhaps, some memory of the past of other variables. Thus, the nature of time series data is different from the data usually studied in statistics courses—the observations are not independent. But, like many other datasets, time series data can be explained with models. In order for a data series to be correctly modeled using time series analysis, it must be stationary. A stationary time series is one who's mean, variance, and covariances do not change over time.

In our study, the question of causality can be rephrased as "Does the chicken depend on the egg, or does the egg depend on the chicken?" We can use a vector autoregressive model of chicken and egg data to answer that question. Although this and other time series data analysis methods are advanced concepts, the basic ideas come from the fundamentals of estimation, hypothesis testing, p-values, and regression analysis.

The United States Chicken Industry

The main products of the poultry industry in the United States are broilers (chicken for meat) and table eggs (eggs for cooking). Being the world's largest producer of poultry meat, 14% of the total United States' annual poultry production is exported. The United States is also the second-largest egg producer in the world.

Although it has become a highly specialized agricultural business nowadays, the commercial poultry industry was made up of millions of small backyard farms before the 1950s, when meat was a byproduct of egg production. Today, poultry products account for about 10% of all farm revenue, and the industry has been transformed almost completely from a fragmented, home-owned industry to a highly organized, vertically integrated industry linking all production decisions from farm to market.

Chicken and Egg Time Series Data

One of the two variables we can study is monthly chickens hatched with the intended purpose of becoming broilers. We will call this variable *hatched*. The other variable we can study is *eggs*, but not table eggs; rather, we will study broiler eggs.

Often the best way to begin a statistical study is to plot the data, and time series analysis is a good example of this. So, let's look at the plots in Figure 1. The data span the years between 1975 through 2002. Because the values of the time series, *hatched* and *eggs*, display an upward trend with inconsistent variability, the time series are not stationary. This is common for real-world time series data. There is also obvious seasonality in the data—a repeating periodic effect at approximately the same time each year.

For many nonstationary time series, the trend can be removed by differencing the data (i.e., subtracting consecutive values of the variable.) Then, the model is built using the changes in the variable from time period to time period, instead of the original values of the variable.

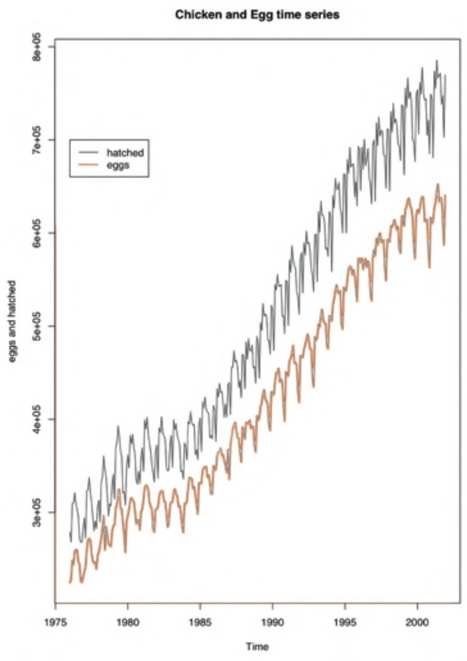


FIGURE 1. Number of broiler chickens hatched (top) in thousands and number of broiler eggs in incubators (bottom), in thousands, on the first day of the month

Differencing

Differencing is an easy and effective method to help stabilize a nonstationarity time series. Simple differences are differences taken one period apart. Seasonal differences are differences taken 12 periods apart. In some time series, we need to do both simple and seasonal differencing.

After simple and seasonal differencing, we can see in Figure 2 that *hatched* and *eggs* fluctuate around a constant mean of zero and the variance looks relatively stable with a few extreme values.

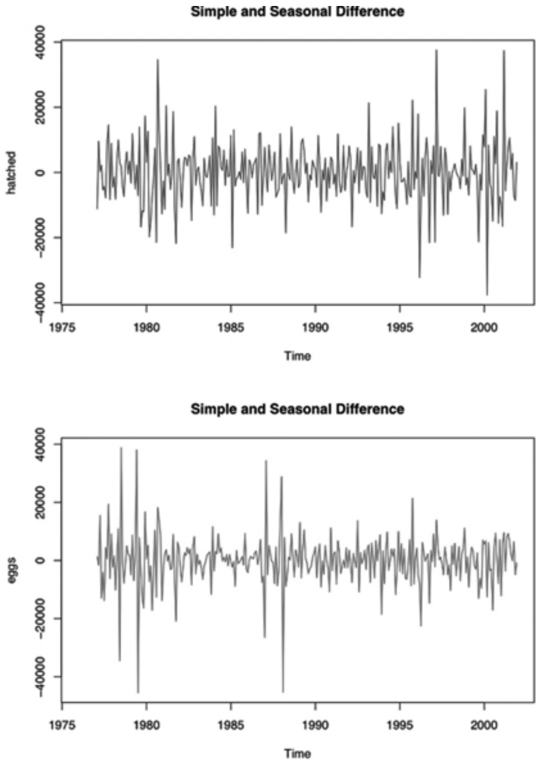


FIGURE 2. Stationary time series after simple and seasonal differencing for *hatched* and *eggs*

But in time series analysis, we do not rely on only our eyes. We also look at two plots that play a prominent role in understanding a time series: the autocorrelation function and the partial autocorrelation function.

Autocorrelation is the association between values of the same variable over time. The autocorrelation coefficient (ρ_k) measures the autocorrelation between two values in a time series k time periods apart. The autocorrelation function (ACF) plots the autocorrelation coefficents values for k from 0 and up.

Partial autocorrelation is the association between times series values separated by k time periods with the effects of the intermediate observations eliminated. A plot of these values is the partial autocorrelation function (PACF).

Autocorrelation Functions

If we want to know whether the "memory" of a time series goes as far back as k months, that is whether the value of *hatched* this month depends on what happened k months ago, then we can test:

$$H_0 : \rho_k = 0$$
$$H_a : \rho_k \neq 0$$

where ρ_k is the autocorrelation between the value of the series at time t and its value k periods before. In the graph in Figure 3, at each lag k (the vertical axis numbers), we test this null hypothesis. The two bands in the graphs represent two standard errors in the sampling distribution of the sample autocorrelation coefficient r_k. If a spike is past two standard errors, this means the p-value for the test at that lag k is smaller than 0.05, and, therefore, we can reject the null hypothesis. If so, the spike is significant, and we say there is memory or correlation between values of the variable at time t and at time t-k.

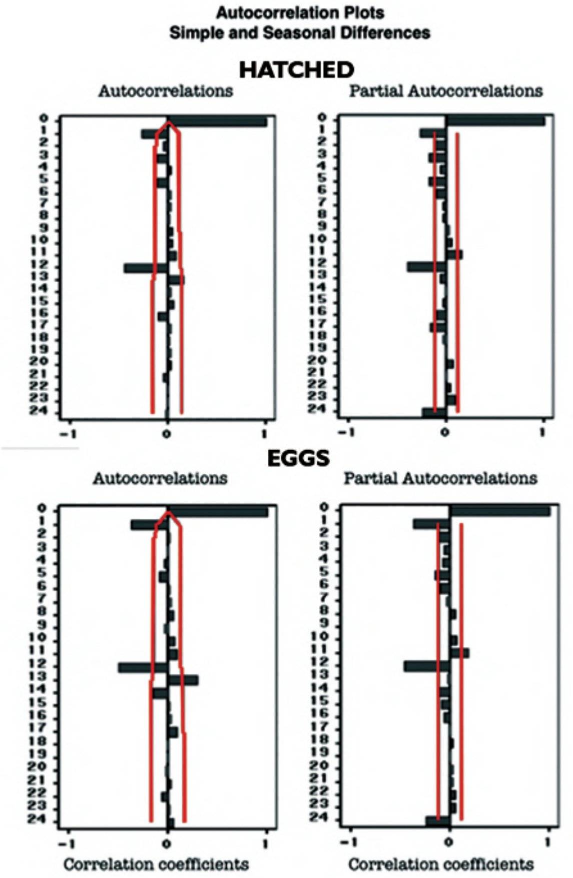


FIGURE 3. Sample autocorrelation and partial autocorrelation functions (sample ACF and sample PACF) for the variables *hatched* and *eggs*

Looking at the sample autocorrelation and partial autocorrelation functions in Figure 3, we can see that the ACF has a significant spike at lag 1 and the PACF exponentially dies down. Therefore, based on these traits, the model for the variable *hatched* is a moving average process of order 1 for nonseasonal lags. However, as there is also a spike in the ACF at 12 months and the PACF shows the seasonality dying down at 24 months, we also have a moving average model at the seasonal lag. Our final model for the variable *hatched* is MA (1, 12), which is shorthand for saying it is a moving average model using lagged variables at one month and 12 months.

Looking at the other time series, *eggs,* the original values display an upward, positive trend, so it also is not stationary (see Figure 1). We can transform *eggs* by taking simple and seasonal differences so the values of the time series fluctuate around a constant mean of zero for stationarity (see Figure 2). The sample autocorrelation and partial autocorrelation functions for the differenced *eggs* data look similar to the ACF and PACF for differenced *hatched* data (see Figure 3). That indicates the model to use for *eggs* is the same as for *hatched:* MA (1, 12).

We are lucky; the change in *eggs* and the change in *hatched* follow the same model. This will make it easier for us to find which comes first, the chicken or the egg. But we are not quite ready for that yet.

Vector Autoregression

The question of which variable leads—precedes in time—in the movement of two stationary time series has been studied often in economics. For example, in the context of predicting stock market prices, one question can be whether the price of a stock that trades in both the United States and, let's say, Germany, is such that the United States price leads the German price or the German price leads the United States price during overlapping trading periods (i.e., during the hours both markets are simultaneously open).

Questions such as this can be answered with a technique used in econometric analysis called vector autoregression (VAR). It is a method that can help us determine the time precedence between variables. If we find that one variable consistently precedes another in time, that would be evidence supporting a possible causal relationship.

In our case, we want to see whether the number of broilers *hatched* causes the number of broiler *eggs* or the number of broiler *eggs* causes the number of broilers *hatched*. For this, we need two regression equations: one to regress *hatched* on *eggs* and the other to regress *eggs* on *hatched*. As this is a time series model, we want to estimate these two regression equations taking into account that there could be causality in either direction. So, we estimate the two equations together with a common variance-covariance matrix for both. This is different from separately estimating them.

ECONOMETRICS

Econometrics literally means "economic measurement." It is the branch of economics that uses statistical methods to study empirically relationships in economic data. Regression analysis and time series analysis provide the foundation for econometric investigations.

VECTOR AUTOREGRESSION

Vector autoregression is a technique in the econometrician's tool kit for analyzing the dynamic properties of an economic system. One application is to estimate the direction of a possible causal relationship between two variables, X and Y. Least squares regression is used to examine the autocorrelation ("self-correlation") due to the time-dependence of each of the variables. First, the analysis is performed with X as the dependent variable and its historical values and the Y values as the independent variables. Then, the process is repeated with Y as the dependent variable with its historical values and the X values as independent variables. Comparing the results of the two regressions can show the direction of possible causality.

A simple vector autoregression for our problem would be a model such as this:

$$X_{1t} = \phi_{11} X_{1,\,t-1} + \phi_{12} X_{2,\,t-1} + \varepsilon_{1t}$$

$$X_{2t} = \phi_{21} X_{1,\,t-1} + \phi_{22} X_{2,\,t-1} + \varepsilon_{2t}$$

where X_{1t} is *hatched* and X_{2t} is *eggs*, both variables are stationary with a mean of zero, and ϕ_{ij} are constants we estimate by regression.

Looking at the above model, we notice that if ϕ_{12} is zero, but ϕ_{21} is not zero, there is no feedback from X_2 to X_1. Thus, X_{1t} (*hatched*) does not depend on the lagged value of *eggs*, but X_{2t} (*eggs*) does depend on the lagged value of *hatched*. This would indicate any causality goes in only one direction.

Vector Autoregressive Model for *Hatched and Eggs*

Based on the structure we found in the sample ACF and PACF, the bivariate VAR model is:

$$hatched_t = -0.2391\ hatched_{t-1} - 0.0043\ eggs_{t-1} - 0.3768\ hatched_{t-12} - 0.0956\ eggs_{t-12}$$

p-value = 0.000 p-value = 0.9464 p-value = 0.0001 p-value = 0.1331

$$eggs_t = 0.1237\ hatched_{t-1} - 0.3614\ eggs_{t-1} + 0.0543\ hatched_{t-12} - 0.5001\ eggs_{t-12}$$

p-value = 0.0169 p-value= 0.0001 p-value = 0.2904 p-value = 0.0001

where $hatched_t$ is the value at time t of the seasonal difference of the first difference for the variable *hatched* and $eggs_t$ is the value at time t of the seasonal difference of the first difference for the variable *eggs*. The p-values of the coefficients correspond to the test of the null hypothesis that a coefficient is equal to zero. A p-value ≥ 0.05 means the coefficient is not significantly different from zero.

Comparing the p-values highlighted in orange, we can see that *hatched* last month (one-month lag) affects *eggs* this month, but no lag of *eggs* affects *hatched* in the present month. This means that while the number of broilers previously *hatched* affects the number of broiler *eggs* in incubators now, the number of broiler *eggs* incubating previously does not affect the number of broilers *hatched* now.

In "economic-speak," that means the number of chickens *hatched* is a leading indicator for the number of *eggs*, but the number of *eggs* is not a leading indicator for the number of chickens *hatched*. So, in our dilemma of causality, the chicken comes first!

Our conclusion makes sense in the economic context. A downturn in broilers is probably an indication of a sluggish chicken meat market, perhaps due to factors such as a recession in the economy, maybe some pandemic of avian flu, or some other economic factor. If this is the case, it does not make economic sense to keep the number of eggs in incubators at the previous level. Why incubate eggs that will give chickens that will not be sold? Consequently, we would expect the number of eggs in incubators to go down. So, as the demand for chicken meat goes up or down, the number of eggs in incubators should follow.

Chicken or the Egg?

As we have seen, time series analysis is fun and makes use of the basic concepts we learn studying introductory statistics: estimation, test of hypotheses, p-values, and regression. We just adapt the basic principles to the circumstances present in a time series modeling problem.

For chicken farmers, it is useful to know that the number of chickens hatched is a leading indicator for the number of eggs in incubators. For all of us concerned with the dilemma of causality, it is nice to see how quantitative methods can help us answer a classic dilemma: Which comes first, the chicken or the egg? Conditional on the vector autoregressive model that we used, in the economic decision chain of the United States poultry industry, the data indicate that chickens come first. ◕

How High Can *r* Go?

by Schuyler W. Huck

Imagine you are sitting at a large table. There are three objects on the table in front of you. On your left, there are 10 index cards, each with an *X* written on its visible side. On your right, there are 10 more index cards, each labeled with a *Y*.

Directly in front of you, between the two sets of index cards, is a sheet of paper with the following instructions:

Each of the 20 cards in front of you has a whole number written on it. Your task is to turn the cards over, look at the numbers, and then create 10 pairs of cards, with each pair made of one *X* card and one *Y* card, and with no card in more than one pairing.

The numbers on the *X* cards are whole numbers. The lowest of these numbers is zero, and the highest is 10. These *X* numbers have a mean (μ_x) of five and a standard deviation (σ_x) of three.

The range of the *Y* numbers is the same as that of the *X* numbers: zero to 10. Moreover, the *Y* numbers have the same mean ($\mu_y = 5$) and standard deviation ($\sigma_y = 3$) as the *X* numbers.

Within each set, the numbers can be repeated and, thus, can appear on more than one card.

Your goal in forming the 10 pairs of numbers is to create a bivariate set of data such that the Pearson product-moment correlation, *r*, between the *X* and *Y* numbers is positive and as high as you can make it.

OK. Turn over the 20 cards and create your 10 pairs of numbers.

Imagine now that you have performed the task described on the sheet of instructions and your 10 pairs have been analyzed to determine the correlation between the *X* and *Y* numbers. How high can *r* go?

See Page 26 for the solution to this puzzle. ◕

SCHUYLER W. HUCK teaches applied statistics at the University of Tennessee. He is the author of *Reading Statistics and Research*, a book that explains how to read, understand, and critically evaluate statistical information. His books and articles focus on statistical education, particularly the use of puzzles for increasing interest in and knowledge of statistical principles.

Story First, Analysis Second

Suppose someone gives us the data in Table 1, telling us the numbers represent 10 bivariate observations *x* and *y*, where *y* is the dependent variable.

TABLE 1. Ten Bivariate Observations

Obs#	1	2	3	4	5	6	7	8	9	10
x	0.000	0.043	0.086	0.129	0.172	0.215	0.258	0.301	0.344	0.387
y	0.000	0.135	0.252	0.352	0.434	0.496	0.538	0.568	0.580	0.566

A Hasty Start

We are eager to get started on our statistical analysis. First, we find the correlation between the two variables; it is $r = 0.94$, which seems to indicate a high degree of linear association. Based on this information, we decide to find a regression equation to predict values of *y* from corresponding values of *x*. The regression equation is

$\hat{y} = 0.108 + 1.47x$

The p-value for the regression model is smaller than 0.0005, indicating the *x*-values are useful in explaining the *y*-values. Finally, based on this equation, we venture to predict the value of *y* corresponding to $x = 0.5$, which is

$\hat{y} = 0.108 + 1.47(0.5) = 0.842$.

The corresponding prediction interval is (0.620, 1.063).

All the computations we have done are correct, but considered from the point of view of appropriate statistical analysis, everything we have done is incorrect and misleading. Why is this not a useful analysis?

The Story behind the Data

These observations are from a standard physics experiment in which a ball is thrown into the air and photographed with a strobe light. From multiple images of the ball in the photograph, we can find the height of the ball in meters (*y*-values) at time increments spaced 0.43 seconds apart (*x*-values).

Figure 1 shows the 10 points corresponding to the data in Table 1 as solid dots—along with our useless regression line from above. The open circles show what happened after each of six additional intervals of 0.43 seconds—data not recorded in the table or used here.

As soon as we know the story, we know the ball must eventually start to come down, and, thus, that it is not appropriate to use a linear model for height as a function of time. A linear function would correspond to an interesting alternate reality with no gravity: It would be a lot easier to launch rockets into space, but we would all have to live in caves—presumably wearing helmets—to keep from launching ourselves into space.

Lesson Learned

Make sure you understand the story behind the data, then do the analysis—story first, analysis second.

If we know a little physics—or, for that matter, just take a moment to look at a plot of the data—we know the path of the ball is a curve, so a better model would be

$y= \text{ß}_0+\text{ß}_1x + \text{ß}_2x^2.$

A Useful Analysis

By doing a regression of y on two variables, *x* and x^2, we can find the best-fitting parabola. When we do this, we find that the estimate of ß_0 is 0.000, so there is no constant term. The best-fitting parabola is

$\hat{y} = 3.357x - 4.884x^2$.

If we plot this parabola (not shown) through the points in the figure, the fit appears to be almost exact. Also, this equation corresponds well with the standard formula from physics for the height of a ball thrown upward, $h = v_0t + (½)\, gt^2$, where *h* is the height of the ball after *t* seconds, v_0 is the initial velocity of the ball in meters per second (m/s), and $g = -9.8$ m/s^2 is the known value of the acceleration due to gravity. Our equation gives $g = 2(-4.884) = -9.77$, which is consistent with the known value. Also, we see that the ball in our photograph must have been thrown into the air with an initial velocity of about 3.4 m/s.

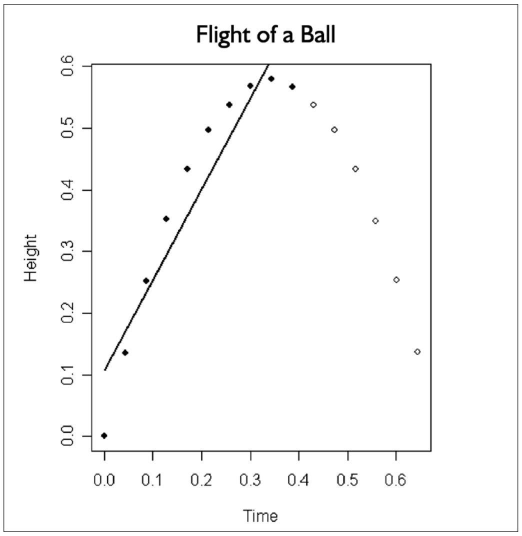
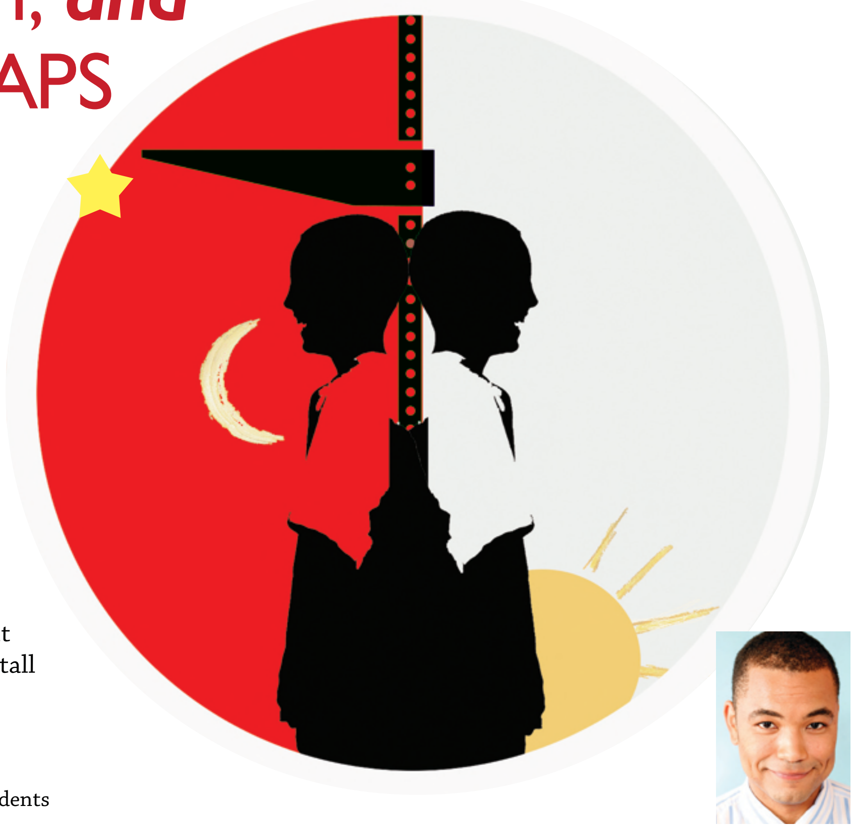


FIGURE 1. Trajectory of the flight of a ball

Using the correct regression on both *x* and x^2, we can estimate the height of the ball at 0.5 seconds to be about 0.46 meters, with a 95% prediction interval of (0.45, 0.47). This agrees well with the points in the vicinity of $x = 0.5$ seconds, not used in this regression. ◕

Shrinking Students, Poisoned Children, *and* BOOTSTRAPS

by Antonio Curtis and Bruce E. Trumbo



One of the classical datasets in the field of statistics resulted from very careful measurements of the heights of 41 students at a boarding school in India in the early 1940s. Four measurements of each student's height were taken in the morning and another four were taken in the evening. Table 1 shows the average morning (AM) and evening (PM) heights for each student. Heights were measured in millimeters (mm). There are about 25.4 mm to the inch, so someone who is 5' 8" tall would also be about 1,728 mm tall.

TABLE 1. Average Morning and Evening Heights of 41 Students

Student	AM	PM	Student	AM	PM
1	1728.75	1720.25	21	1688.75	1677.00
2	1538.25	1528.50	22	1688.75	1681.00
3	1462.25	1452.50	23	1620.75	1613.50
4	1782.50	1776.50	24	1679.00	1668.25
5	1671.00	1667.00	25	1557.25	1550.25
6	1581.75	1571.00	26	1704.50	1696.50
7	1673.75	1664.50	27	1632.75	1619.00
8	1721.75	1708.50	28	1587.00	1581.50
9	1646.50	1636.00	29	1598.75	1590.50
10	1793.75	1781.75	30	1592.25	1583.50
11	1825.25	1814.00	31	1719.50	1709.25
12	1801.50	1787.00	32	1807.50	1795.00
13	1742.50	1729.75	33	1624.00	1619.50
14	1720.75	1711.50	34	1705.25	1694.50
15	1728.25	1717.25	35	1692.75	1686.00
16	1753.75	1742.75	36	1795.25	1782.00
17	1725.50	1716.75	37	1643.50	1628.75
18	1598.00	1592.25	38	1677.25	1668.25
19	1756.25	1747.00	39	1647.75	1641.50
20	1674.00	1662.50	40	1620.00	1608.25
			41	1727.50	1721.25

Students' Heights Decrease During the Day

A quick inspection of these measurements shows that every one of the 41 students measured taller in the morning than in the evening. So, if we are willing to consider these students as randomly chosen from some population, there is overwhelming evidence that people in that population tend to be taller in the morning. If morning and evening heights do not differ in the population, there is only one chance in 2^{40} (more than a million million) that every difference would have the same sign.

It turns out that similar decreases in height from morning to evening have been seen in many other groups of people. A likely explanation is that the shrinkage occurs mainly along the spine as the cartilage between vertebrae becomes compressed during the day.

ANTONIO CURTIS
When he submitted this article, Antonio Curtis was a master's student in statistics at California State University, East Bay. He is now a mathematics instructor at Riverside Community College. His statistical interests center on social and governmental statistics.

BRUCE TRUMBO is a professor of statistics and mathematics at California State University, East Bay (formerly CSU Hayward). He is a Fellow of the American Statistical Association and a holder of the ASA Founder's Award.

Estimating the Amount of Shrinkage

If we compute the 41 differences between morning and evening measurements in Table 1, we find that these differences average about 9.60 mm, with a standard deviation of about 2.74, and thus a standard error of $2.74/\sqrt{41} = 0.43$. Assuming the data to be normally distributed, we find the 95% confidence interval to be (8.73 mm, 10.47 mm). This is based on 9.60 ± 2.02(0.43), or 9.60 ± 0.87, where 2.02 is the 97.5 percentile of student's *t* distribution with 40 degrees of freedom. The normality assumption is discussed below.

In practical terms, this confidence interval—centered at about 10 mm—indicates it ordinarily would not make sense to try to measure someone's true height with more precision than about the nearest 10 mm, or 1 cm (between ⅜" and ½"). Height varies by about that much during a day.

Indeed, just at any one given time, it seems measurements cannot be reproducibly made within less than a few millimeters. For example, even though the measurements in Table 1 were done very carefully, the four measurements that averaged to 1728.75 for Student 1 in the morning were 1727, 1728, 1730, and 1730. The margin of error for estimating this student's true morning height from these four measurements is about 2.4 mm. This lack of precision is reflected in the margin of error of our confidence interval for the population mean shrinkage in height. But a much larger component of that margin of error arises from differences in the amount of shrinkage from one student to another.

Empirical Cumulative Distribution Functions

One can view the density function of the distribution of a random variable as the smoothed histogram of a large sample from a population with that distribution. So for a large sample, we can use a histogram to judge whether the data fit a particular density function. But this does not work so well for samples of a small or moderate size. For example, Figure 1 shows a histogram of the 41 height differences with the density curve for NORM(9.60, 2.74), the best-fitting normal distribution. This best normal fit is not terrible, but density curves of some non-normal shapes would fit at least as well.

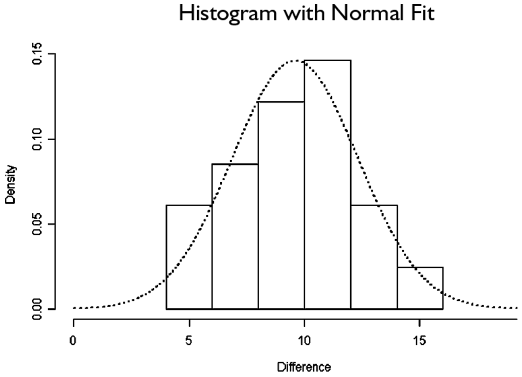


FIGURE 1. From their histogram, it is not clear how well the 41 height measurements fit a normal density function (dotted curve).

By contrast, Figure 2 shows the empirical cumulative distribution function (ECDF) of these 41 differences, along with the theoretical cumulative distribution function (CDF) of NORM(9.60, 2.74). Beneath the ECDF is a strip chart showing the values of the 41 individual differences. Notice that the ECDF starts at 0 on the left and increases to 1 toward the right. The ECDF has a jump of i/n at an observation value, where i is the number of tied observations at that value. It is relatively easy to see that there is a pretty good fit of the ECDF to the theoretical CDF. Of course, because of randomness, one cannot expect all the black dots of the ECDF to lie exactly on the smooth dotted CDF curve, but none of them lies far away.

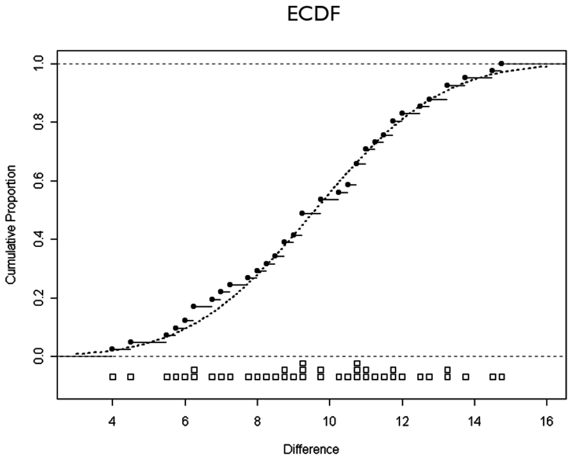


FIGURE 2. It is easy to see that the ECDF of the 41 differences provides a reasonably good fit to the best-fitting normal CDF.

The ECDF contains exact information about the sample of 41 differences. But the histogram was made by sorting these 41 values into six fairly large "bins," and so the histogram is based on information that is only approximate. For relatively small sample sizes, ECDFs work better than histograms in judging the goodness of fit to a particular distribution because ECDFs do not waste any information.

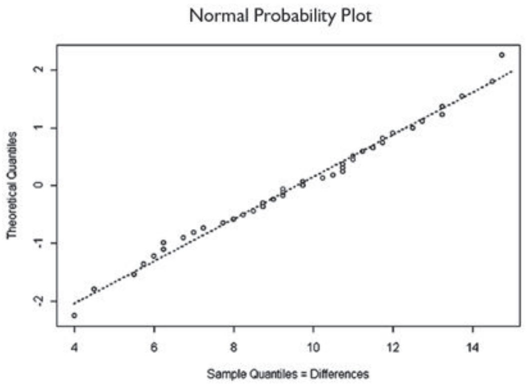


FIGURE 3. Here, the vertical axis of the ECDF plot has been transformed, so the theoretical CDF curve becomes a straight line.

A normal probability plot, based on an ECDF, often is easier to interpret than the ECDF, itself. The idea in making a normal probability plot is to distort the vertical scale of the ECDF so the theoretical normal CDF becomes linear. Specifically, the transformation is the standard normal quantile function, which is the inverse of the standard normal CDF. Then, we can judge relatively easily whether the points of the distorted ECDF lie along a straight line, and it is not necessary to draw the CDF curve for reference. Figure 3 shows a normal probability plot of the 41 height differences. In Figure 3, we show the theoretical line, which is the straightened version of the CDF curve of Figure 2. In this particular case, the equation of the straight line is $y = (d - 9.60)/2.74$.

A Nonparametric Bootstrap Confidence Interval

Above, we used a *t* confidence interval (CI) to estimate the population mean difference in student heights between morning and evening. Provided the population is normally distributed, there is no loss of information if we summarize the data into just two numbers—the sample mean and sample standard deviation—to find the *t* confidence interval. This CI is easy to compute, but if the data are not from a normal distribution, it may not give the best possible result.

The nonparametric bootstrap procedure is a computationally intensive method of estimation. It is based on all the available information, and it does not rely on assuming normality about the data, or that they even adhere to any other particular distributional family. The bootstrap is based on the idea that, because the ECDF contains all the information in the sample, it is the best available imitation of the theoretical CDF. This is true whether or not the CDF is normal. A sample must be of at least moderate size before the ECDF can be relied upon to give a good approximation of the CDF.

Specifically, the bootstrap procedure treats the *n* observed data values as a substitute population. We know for sure that the actual population contains these values, and we do not know for sure whether it can produce any other particular values. We take many bootstrap samples of size *n* from the substitute population. Sampling is done with replacement, which means data values can occur more than once in the bootstrap sample. From each bootstrap sample, we find the mean. This gives us a simulated sample distribution of bootstrap means. Cutting off the top and bottom 2.5% from this bootstrap distribution gives a 95% *nonparametric bootstrap confidence interval.*

Based on a bootstrap with 10,000 bootstrapped samples from among the 41 height differences, we obtain a 95% confidence interval (8.77, 10.43), which is just a little shorter than the *t* CI (8.73, 10.47) obtained above. Figure 4 shows the

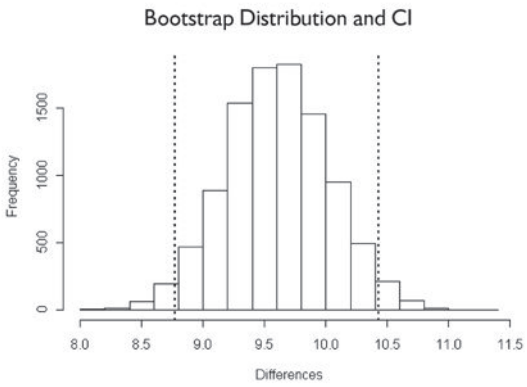
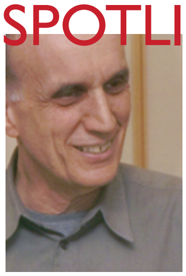


FIGURE 4. Bootstrap distribution of differences in heights. The vertical dotted lines show the 95% bootstrap confidence interval.

SPOTLIGHT



Bradley Efron, a past president of the American Statistical Association, has made important theoretical and applied contributions to statistical science during his long career as a professor at Stanford University. He is perhaps best known for developing the bootstrap resampling technique. Bootstrapping is a method for estimating the sampling distribution of a test statistic by sampling repeatedly—"resampling"—with replacement from the originally drawn sample. For more information, visit *www-stat.stanford.edu/~brad.*

Bootstrap confidence intervals were introduced by Bradley Efron about 25 years ago as a way to use intensive computation in solving a variety of statistical estimation problems that would be difficult or impossible to solve analytically. During the last 20 years—as computing power has become cheaper and more readily available—the bootstrap has become a standard method of assessing the accuracy of estimates.

> “The impact of the bootstrap has transcended both theory and applications. The bootstrap has shown us how to use the power of the computer and iterated calculations to go where theoretical calculations cannot, which introduces a different way of thinking about all of statistics.”
>
> *- George Casella,*
> *introducing a recent collection of papers*
> *about the history and current importance of the bootstrap*

histogram of the bootstrap distribution from this run. Because this is a simulation process, each run will give a slightly different histogram and CI; endpoints of the CI may differ slightly in the second decimal place. If the bootstrap distribution were not so nearly symmetrical, some adjustment might be necessary in the resulting confidence interval, but we will not discuss such adjustments here.

Because the height differences appear to be nearly normally distributed, the *t* confidence procedure correctly uses nearly all the information in the sample of 41 observations. The bootstrap CI is a simulation procedure. While it requires no assumption of normality and uses all the information in the sample, its endpoints are subject to a small fluctuation from one run to another that decreases as the number of iterations in the simulation increases.

The term “bootstrap” comes from the expression about pulling oneself up by one’s own bootstraps in the face of adversity. Here, the adversity would be ignorance of the distribution of the population, which implies ignorance of the distribution of the sample mean.

Lead Poisoning in Children

Now, we look at another study involving differences. But here, it does not seem reasonable to assume the differences are normally distributed.

At a factory in Oklahoma, batteries were manufactured using lead. Lead is a serious neurotoxin that is especially dangerous to young children. Although workers at the factory were told to take showers and change clothes and shoes before going home, the concern remained that lead dust might be transported from the factory to the home, where it could contaminate the children of the workers.

Thirty-three children whose fathers work at the factory are the principal subjects of a study. Blood samples are taken from them, and the amount of lead in their blood is determined. But if lead is found in their blood, this does not necessarily mean it came from the battery factory or the bodies or clothing of their fathers. Many other sources of lead contamination exist—lead in water pipes and paint (applied before its use in paint was banned), for example.

As a control group, a “matching” child is found for each of the principal subjects. Matching is based on neighborhood (similar possibilities for environmental lead poisoning) and age (lead poisoning is cumulative over time). Children in the control group also are tested for blood levels of lead. For each of the 33 pairs of children, Table 2 shows the blood level of lead (in µg/dl) for the potentially “exposed” child and the “control” child, along with the difference, exposed minus control.

TABLE 2. Lead Levels for Children Whose Fathers Work in an Industry Where Lead Is Used (Exps), for Children Selected as Matched Controls (Cont) and Differences (Diff)

Pair	Exps	Cont	Diff	Pair	Exps	Cont	Diff
1	38	16	22	17	15	24	-9
2	23	18	5	18	10	13	-3
3	**41**	18	23	19	**45**	9	36
4	18	24	-6	20	39	14	25
5	37	19	18	21	22	21	1
6	36	11	25	22	35	19	16
7	23	10	13	23	**49**	7	42
8	**62**	15	47	24	**48**	18	30
9	31	16	15	25	**44**	19	25
10	34	18	16	26	35	12	23
11	24	18	6	27	**43**	11	32
12	14	13	1	28	39	22	17
13	21	19	2	29	34	25	9
14	17	10	7	30	13	16	-3
15	16	16	0	31	**73**	13	60
16	20	16	4	32	25	11	14
				33	27	13	14

The first thing to notice here is that, in the exposed group, there are some really serious cases of lead poisoning. Although no amount of lead is desirable, according to standard guidelines, the children above 40 µg/dl (printed in bold in Table 2) need medical treatment and those above 60 µg/dl should be hospitalized immediately.

Figure 5 shows a normal probability plot of the differences in blood lead levels. The points

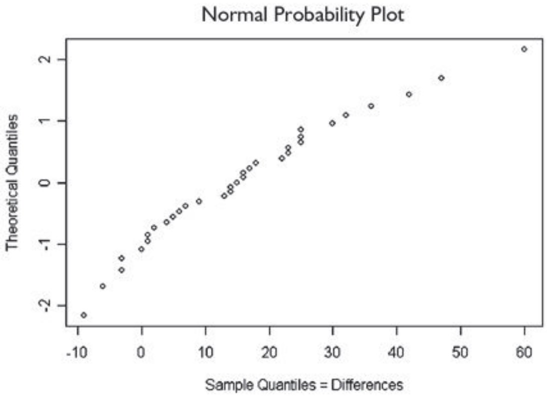


FIGURE 5. Normal probability plot of the differences in Table 2. The pattern of points is noticeably nonlinear, indicating possible non-normality.

seem to fit a curve, rather than a straight line. In particular, the distribution seems to be skewed somewhat to the right. The *t* confidence interval procedure is fairly tolerant of moderate departures from normality, so the CI (10.34, 21.59) it produces is probably not grossly misleading. But, for these data, a nonparametric bootstrap CI seems a better choice.

The bootstrap CI from one run of 10,000 iterations is (10.9, 21.4), which is shorter than the *t* interval. The study that yielded the data in Table 2 also investigated the amount of lead exposure of workers and their adherence to the hygiene rules intended to prevent spread of lead dust from the workplace to home. The children with the highest levels of lead tended to have fathers with both high exposure in the workplace and a lax attitude about the rules on showering and changing clothes.

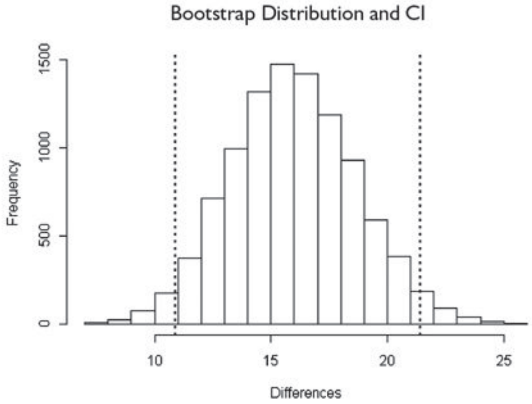


FIGURE 6. Bootstrap distribution of differences in lead levels and their confidence interval

The R code used to make Figure 6 is shown in Figure 7. Although great amounts of computation are required, the structure of the program is straightforward. ◕

```
 dfr  = c(22,  5, 23, -6, 18, 25, 13, 47, 15,
 16,   6,  1,  2,  7,  0,  4, -9, -3, 36, 25,
  1,  16, 42, 30, 25, 23, 32, 17,  9, -3, 60,
 14,  14)

m = 10000; n = length(dfr); d = numeric(m)
for (i in 1:m) {d[i]= mean(sample(dfr, n, repl=T)) }

hist(d, xlab=“Differences”)
bci = quantile(d, c(.025, .975)); bci
abline(v=bci, lty=“dotted”, lwd=2)
```

FIGURE 7. R code for Figure 6. The code for Figure 4 is similar.

CHALLENGES

1. Use the code in Figure 7 to do your own bootstrap confidence interval on the height differences. Is your result similar?

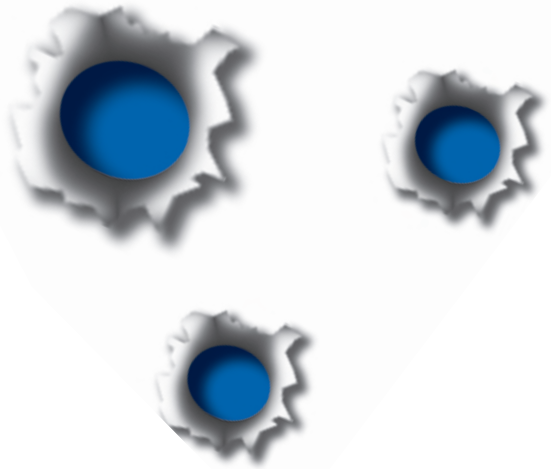
2. Generate mixed normal data with $n = 40$ using

```
set.seed(1)
x = c(rnorm(20, 100, 10), rnorm(20, 130, 15))
```

The population is not normal. Why not? Make a histogram with `hist(x)` and a normal probability plot with `qqnorm(x)`. Can you detect non-normality from the data?

3. Show that the population mean in Challenge 2 is $\mu = 120$. Make a 95% *t* CI and a 95% nonparametric bootstrap CI for μ. Is the true value $\mu = 120$ included in both intervals?

BONNIE & CLYDE meet Bayesian Statistics



by Chris Olsen

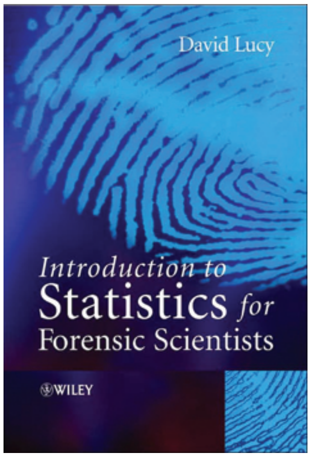
CHRIS OLSEN teaches mathematics and statistics at George Washington High School in Cedar Rapids, Iowa. He has been teaching statistics in high school for 25 years and has taught AP statistics since its inception.

As it is possible that the U.S. attorney general or his designates might be listening to my phone conversations, I suppose I should say at the outset that I was really only joking when I said I was considering turning to a life of crime. Actually, I had been watching a lot of cable television when I noticed there were various crimes of a larcenous nature going unsolved—many perhaps committed by noncollege graduates. I also figured that, should it come down to it, having to live on a teacher's salary would bolster my case for an insanity plea, especially if I got a jury of peers.

My first firm hint that a life of crime might not be the way to go came the other day when I stayed awake long enough to watch the end of the movie "Bonnie and Clyde." Wow, was that a lesson learned! Apparently, after every crime spree, one is forced to buy a new car or at least foot the bill for some significant repairs. This is a serious problem, as I really like my hybrid and I am told I will not be able to get the next one with a standard transmission. Even if they still have standard transmissions, I could not afford one on a teacher's salary.

But what really is going to keep me on the straight and narrow, crime-wise, is this book I just stumbled across by David Lucy, *Introduction to Statistics for Forensic Scientists* (*ISFS*). Of course, teachers of statistics already have a full-blown appreciation for the power of statistics. And I, like all statistics teachers, have used the analogy of jury decisions to Type I and Type II errors when testing hypotheses. But, in all my years of watching "Perry Mason" and "Law and Order," I never actually saw any statistical arguments presented. It seemed nothing vaguely scientific went beyond the practice of the classical Greeks (i.e., consulting oracles). (Well, they do not

“My eyes were really opened by *Introduction to Statistics for Forensic Scientists* when I read the historical, philosophical, and logical discussion of the nature of forensic evidence—and the place of statistics within the whole scheme.”



call them oracles anymore; the current politically correct term is “expert witnesses.”) I remembered Perry Mason got all his clients off, and my chances seemed about even on “Law and Order.” Think again! My eyes were really opened by *ISFS* when I read the historical, philosophical, and logical discussion of the nature of forensic evidence—and the place of statistics within the whole scheme. In my mind’s eye, I imagined lots of budding forensic types reading *ISFS*. It now seems to me a life of crime might be a seriously risky venture.

This book was written for future forensic scientists, and I do have to admit there were parts that caused my eyes to glaze over. For example, where the forensic types might be pardoned for skipping over the details of a derivation of Bayes’ Theorem, I am hoping my understanding of the plot of *ISFS* was not interrupted during my deer-in-the-headlight stages when I would read passages such as “The rhomboid fossa is a groove which sometimes occurs on one end of the clavicle as a result of the attachment of the rhomboid ligament…” (Apparently, the point is that this rhomboid thing helps tell whether skeletal remains are male or female. What I learned, however, is that, in the unlikely event I stumble on a skeleton, the secret of its gender will be safe from my prying.) I was actually able

to follow the discussion of DNA and its role in identifying individuals from evidence left at crime scenes and in establishing paternity. I was really in my comfort zone when the discussion got around to the evidential value of trouser fibers, shoe types, and firearms. So the statistics part was not all I understood while reading this.

What really scared me out of an anticipated life of crime, though, is the decidedly Bayesian character of the analyses presented in *ISFS*. As a confirmed frequentist and merely budding criminal, I believed that until I committed a number of crimes all looking alike—the number 30 sticks in my mind—the power of statistics could not be brought to bear against me. However, I now understand that strong evidential arguments can be marshaled against me, even if I commit fewer than two crimes. Those Bayesians really know how to hurt a guy who is just trying to make a criminal living.

So, in closing, I would like to thank Lucy for writing this tome. His scribbles have prevented me from embarking on what, with my luck, would have been a short life of crime and subsequent long stretch behind bars. With a heavy heart and a light bank account, but with possession of the moral high ground, I will stick to my life of teaching mathematics and statistics. I am sure my hybrid will love me for it. ◕

SPOTLIGHT

THOMAS BAYES (1702–1761) was a British mathematician and Presbyterian minister. He is credited with setting forth Bayes’ Theorem, which can be used to update a prior probability assessment to a posterior probability based on newly observed evidence.

Solution to STAT•DOKU puzzle on Page 8

STAT•DOKU

S	T	A	T	I	S	T	I	C
I	S	C	I	S	T	A	T	T
T	T	I	C	A	T	I	S	S
T	A	S	S	I	T	C	I	T
C	T	S	I	S	A	I	T	T
I	I	T	T	T	C	S	S	A
T	C	T	S	T	I	S	A	I
A	I	I	T	C	S	T	T	S
S	S	T	A	T	I	T	C	I

How High Can r Go?

When confronted with this puzzle's question, many people respond by saying they could create 10 pairs of data such that the correlation coefficient, r, will turn out equal to +1.00. That may or may not be true. It depends on the distributional shape of the two sets of numbers.

In order for you to be able to produce an r equal to +1.00, the distributional shape of the X numbers must be identical to the distributional shape of the Y numbers. The two sets of numbers do not need to be normally distributed; however, for r to "max out" at +1.00, both sets of numbers must have the same distributional shape. The distributions can be positively skewed, negatively skewed, rectangular, bimodal, or anything else, as long as they are the same.

If the X numbers have a distributional shape that is different from the distributional shape of the Y numbers, then r has a maximum possible value that is smaller than +1.00. Consider, for example, the following data where the X and Y numbers have the same range, mean, and standard deviation:

X: 0, 1, 2, 5, 5, 6, 6, 7, 8, 10
Y: 0, 2, 3, 4, 4, 5, 5, 8, 9, 10

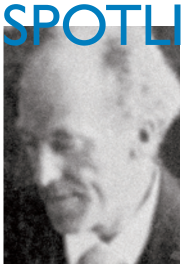
Regardless of how you might create 10 pairs of data using these numbers, you will not be able to get r to equal +1.00. That is because the X numbers are positively skewed and the Y numbers are negatively skewed. Because the skewness in each dataset is minor, the maximum value of r is quite high: ±0.96.

If the two sets of numbers being correlated have distributional shapes that are radically different, the maximum positive value of r may be far away from the widely presumed limit of +1.00. To illustrate, consider these data:

X: 1, 1, 1, 1, 5, 5, 9, 9, 9, 9
Y: 1, 5, 5, 5, 5, 5, 5, 5, 5, 9

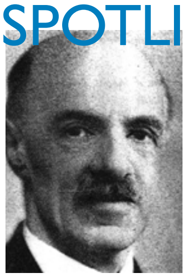
Here, the maximum positive value of r is +0.50. So, we can see that the maximum correlation between two sets of numbers can be much less than +1 (or much greater than –1), if the distributions of the variables differ greatly in shape. In such situations, a researcher should use Spearman's rank correlation (r_S), instead of Pearson's correlation (r).

SPOTLIGHT

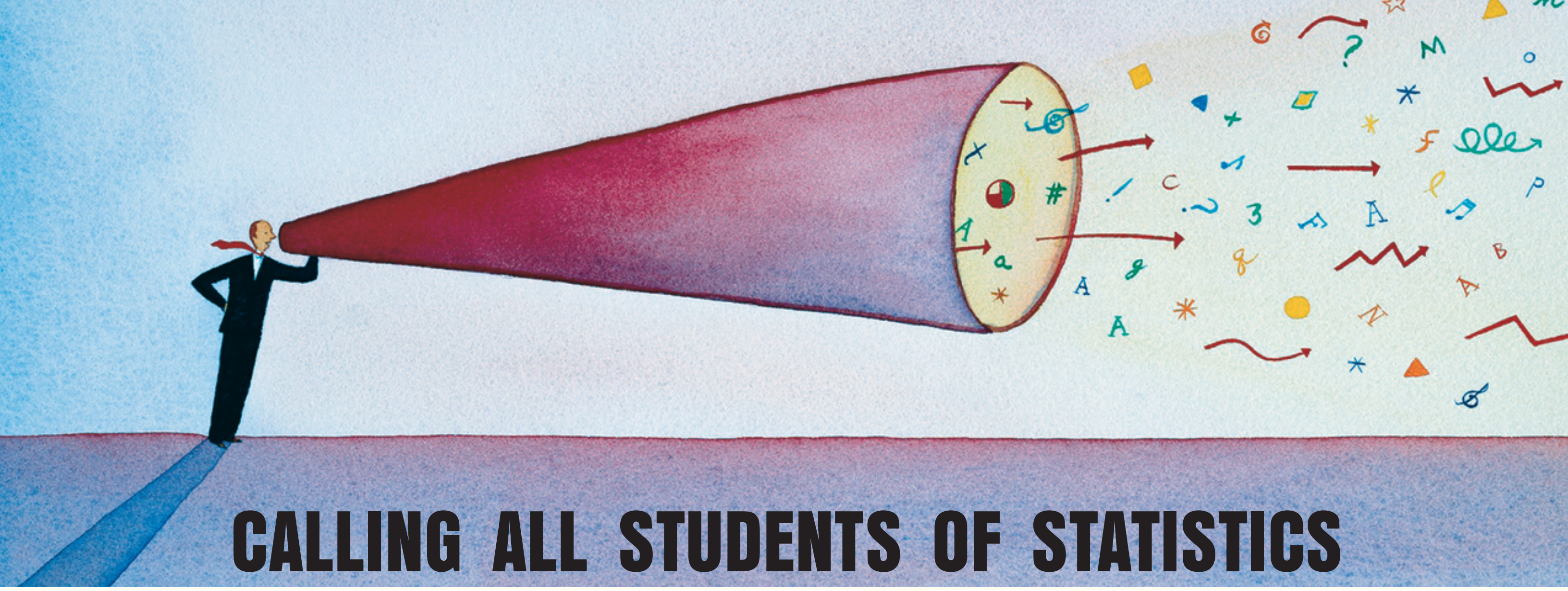


KARL PEARSON (1857–1936) made seminal contributions in statistics and laid much of the foundation for the statistical analysis we use today. He developed Pearson's product-moment correlation coefficient (r), commonly called the "correlation coefficient." It measures the tendency for two variables, X and Y, to increase or decrease together. In other words, it measures their co-relation[ship]. He also developed Pearson's chi-square test, a statistical test of significance for categorical data.

SPOTLIGHT



CHARLES SPEARMAN (1863–1945) was a pioneer in several statistical areas. He is best-known for his development of Spearman's rank correlation coefficient (r_S)—a nonparametric measure of correlation. Also know as "Spearman's rho," it uses the ranks of the data, rather than the data values. It is, therefore, the correlation between the ranks of the Xs and the ranks of the Ys in each pair of numbers. While Pearson's (r) requires data measured on an interval or ratio scale, Spearman's (r_S) can be used for variables measured on an ordinal scale.



CALLING ALL STUDENTS OF STATISTICS

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 datasets, programs, and examples

INTERVIEWS WITH PRACTICING STATISTICIANS working on intriguing and fascinating problems

STATISTICIANS IN HISTORY and the classic problems they studied

HOW TO USE particular probability distributions in statistical analyses

EXAMINING SURPRISING EVENTS and asking, "What are the chances?" and then providing the answers

REVIEWS OF BOOKS about statistics that are not textbooks

STUDENT PROJECTS using statistics to answer interesting research questions in creative ways

Send a description of your concepts for feature articles to Editor Paul J. Fields *pjfields@byu.edu*

References and Additional Reading List

The references for each article in this issue of *STATS* are included in the listing below, along with suggestions for additional reading on related topics. The page numbers for the corresponding articles are shown in blue.

3 An Interview with Sallie Keller-McNulty. To learn more about the American Statistical Association, visit *www.amstat.org.*

6 A Permutation Test of the *Challenger* O-Ring Data. The data on the *Challenger* disaster can be found in Chapters 4, 20, and 22 of *The Statistical Sleuth,* by Fred L. Ramsey and Daniel W. Schafer, Thomson Learning, Inc., 2002.

To learn more about permutation tests, see Chapter 5 of *Introduction to Modern Nonparametric Statistics,* by James J. Higgins, Thomson Learning, Inc., 2004.

8 STAT•DOKU. To learn about Latin square designs, see Chapter 4 of *Statistics for Experimenters: Design, Innovation, and Discovery*, by George E. P. Box, J. Stuart Hunter, and William G. Hunter, John Wiley & Sons, Inc., 2005.

9 Can You See the Trees for the Forest? To review the calculation of variance, see Chapter 2 of *The Basic Practice of Statistics*, by David S. Moore, W. H. Freeman and Company, 2007.

12 Which Came First, the Chicken or the Egg? The CBS News Video, "Was the Chicken or Egg 1st?," was shown on November 9, 2005. It can be found at *www.cbsnews.com.*

For a comprehensive look at the United States poultry industry, read "Briefing Rooms: Poultry and Eggs," Economic Research Service, United States Department of Agriculture, 2005. *www.ers.usda.gov/briefing/poultry.*

An outline of the broiler and eggs production cycles can be found in "U.S. Broiler and Egg Production Cycles," National Agricultural Statistical Service, United States Department of Agriculture, September 2005.

For more on time series models in econometrics, see Chapter 16 of *Undergraduate Econometrics,* by R. Carter Hill, William E. Griffiths and George C. Judge, John Wiley & Sons, Inc., 2001.

For more on vector autoregression, see Chapter 3 of *Applied Time Series Econometrics,* edited by Helmut Luetkepohl, Peter C. Phillips, and Markus Kraetzig, Cambridge University Press, 2004.

The National Chicken Council web site is at *www.nationalchickencouncil.com.*

17 How High Can *r* Go? An extensive investigation of the range of the correlation coefficient is reported in "Evaluating Correlation with Proper Bounds," by Weichung Joseph Shih and Wei-Min Huang, *Biometrics* 48, 1207-1213, December 1992.

A biography on Karl Pearson's life and work is *Karl Pearson: the Scientific Life in a Statistical Age,* by Theodore M. Porter, Princeton University Press, 2004.

For more on Spearman rank correlation, see Chapter 5 of *Introduction to Modern Nonparametric Statistics*, by James J. Higgins, Thomson Learning, Inc., 2004.

18 Story First, Analysis Second. The trajectory data were digitized from a photograph in *Physics for Scientists and Engineers,* by Paul A. Tipler, Worth, 1991.

19 Shrinking Students, Poisoned Children, and Bootstraps. For more about the rationale of the bootstrap and additional kinds of applications, see the review article, "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy," by Bradley Efron and Rob J. Tibshirani, *Statistical Science*, Vol. 1, 54-77, 1986.

For an overview of the bootstrap method, see *An Introduction to the Bootstrap*, *Vol. 57*, Bradley Efron and Rob J. Tibshirani, CRC Press, 1994.

The quote from George Casella is from *Statistical Science*, Vol. 18, 133, May 2003—an issue devoted to the "Silver Anniversary of the Bootstrap."

A comprehensive guide to resampling is presented in *Resampling Methods: a Practical Guide to Data Analysis,* by Phillip I. Good, Springer, 2001.

To learn more about permutation tests, see Chapter 5 of *Introduction to Modern Nonparametric Statistics*, by James J. Higgins, Thomson Learning, Inc., 2004.

The data on heights are from D. N. Majumdar and C. R. Rao's, "Bengal Anthropometric Survey, 1945: a Statistical Study," *Sankhya, the Indian Journal of Statistics*, Vol. 19, 296-298, 1958.

The data on lead exposure are from "Lead Absorption in Children of Employees in a Lead-Related Industry," by David F. Morton et al., *American Journal of Epidemiology,* Vol. 115, 549-555, 1982.

Both the heights and lead exposure datasets are discussed at length in *Learning Statistics with Real Data: Hands on Data Analysis*, by Bruce E. Trumbo, Thomson Learning, Inc., 2002.

The R code for these analyses is available on the *STATS* web site, *www.amstat.org/publications/stats.*

24 Bonnie and Clyde Meet Bayesian Statistics. The book that motivated this article is I*ntroduction to Statistics for Forensic Scientists*, by David Lucy, John Wiley & Sons, Inc., 2005.

The basic concepts and applications of Bayesian statistics are presented in *Introduction to Bayesian Statistics*, by William M. Bolstad, John Wiley & Sons, Inc., 2004.

For more about that 'magic' number, *n*=30, see "Ask *STATS*" in *STATS*, Issue 46, 2006.

Attention Professors of Statistics

Help the American Statistical Association introduce the benefits of membership to your students.

ASA student membership includes:

Subscriptions to *Amstat News* and *STATS: The Magazine for Students of Statistics*

Access to a network of more than 17,000 statisticians

ASA publications at special, drastically reduced student subscription rates

Job opportunities at the annual JSM Career Placement Service and via the ASA JobWeb at *www.amstat.org/jobweb*

The ASA is happy to extend support for local events held to encourage student membership. The ASA will reimburse up to $100 for refreshments when you host a membership social.

For more details about hosting a social, please visit *www.amstat.org/membership/ChapterSchoolSocials.pdf*.

STUDENT MEMBERS join ASA *for* only $10

Direct your students to *www.amstat.org/ membership/index.cfm?fuseaction=student* to find more information.