



# Beware the LURKING Variable

Understanding Confounding from Lurking Variables Using Graphs

**Random Numbers** from  
Nonrandom Arithmetic

**Observational Studies:**

The Neglected Stepchild in the Family of  
Data Gathering

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

# Call for Papers



## **Statistical Computing and Statistical Graphics Sections American Statistical Association Student Paper Competition 2007**

The Statistical Computing and Statistical Graphics Sections of the ASA are cosponsoring a student paper competition on the topics of statistical computing and statistical graphics. Students are encouraged to submit a paper in one of these areas, which might be original methodological research, a novel computing or graphical application in statistics, or any other suitable contribution (for example, a software-related project). The selected winners will present their papers in a topic-contributed session at the 2007 Joint Statistical Meetings. The Sections will pay registration fees for the winners and as a substantial allowance for transportation to the meetings and lodging (which in most cases covers these expenses completely).

Anyone who is a student (graduate or undergraduate) on or after September 1, 2006, is eligible to participate. An entry must include an abstract, a six page manuscript (including figures, tables, and references), a blinded version of the manuscript (with no authors and no references that easily lead to identifying the authors), a CV, and a letter from a faculty member familiar with the student's work. The applicant must be the first author of the paper. The faculty letter must include a verification of the applicant's student status and, in the case of joint authorship, should indicate what fraction of the contribution is attributable to the applicant. We prefer that electronic submissions of papers be in Postscript or PDF. All materials must be in English.

All application materials **MUST BE RECEIVED** by 5:00 p.m. EST, Monday, December 18, 2006, at the address below. They will be reviewed by the Student Paper Competition Award committee of the Statistical Computing and Graphics Sections. The selection criteria used by the committee will include innovation and significance of the contribution. Award announcements will be made in late January 2007.

Additional important information about the competition can be accessed at [www.statcomputing.org](http://www.statcomputing.org). A current pointer to the web site is available from [www.amstat.org](http://www.amstat.org). Inquiries and application materials should be emailed or mailed to:

### **Student Paper Competition**

c/o J. R. Lockwood

The RAND Corporation

4570 Fifth Avenue, Suite 600

Pittsburgh, PA 15213

[lockwood@rand.org](mailto:lockwood@rand.org)

# Call for Papers



#### Editor

Paul J. Fields  
pjfields@stat.byu.edu

Department of Statistics  
Brigham Young University  
Provo, UT 84602

#### Editorial Board

Peter Flanagan-Hyde  
peterfh@mac.com

Mathematics Department  
Phoenix Country Day School  
Paradise Valley, AZ 85253

Schuyler W. Huck  
shuck@utk.edu

Department of Educational  
Psychology and Counseling  
University of Tennessee  
Knoxville, TN 37996

Jackie Miller  
jbm@stat.ohio-state.edu

Department of Statistics  
The Ohio State University  
Columbus, OH 43210

Chris Olsen  
colsen@cr.k12.ia.us

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

Bruce Trumbo  
bruce.trumbo@csueastbay.edu

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

#### Production

Megan Murphy  
megan@amstat.org

American Statistical Association  
732 North Washington Street  
Alexandria, VA 22314-1943

Valerie Snider  
val@amstat.org

American Statistical Association  
732 North Washington Street  
Alexandria, VA 22314-1943

*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; Web site [www.amstat.org](http://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: \$15.00 a year for ASA members; \$20.00 a year for nonmembers; \$25.00 Library subscription.

Ideas for feature articles and material for departments should be sent to the editor 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 directly to the ASA office.

Copyright (c) 2006 American Statistical Association

## Features

### 3 **Glide Testing: a Paired Samples Experiment**

*Catherine Elizabeth Cavagnaro*

### 14 **Understanding Confounding from Lurking Variables Using Graphs**

*Milo Schield*

## Departments

#### 2 **Editor's Column**

#### 6 **Statistical Snapshot**

What Does a t-Test Test?

#### 8 **Ask *STATS***

Why Is  $n = 30$  'Magic'?

*Jackie Miller*

#### 12 **Statistical Snapshot**

Should Outliers Be Deleted?

#### 19 **AP Statistics**

Observational Studies:  
the Neglected Stepchild in  
the Family of Data Gathering  
*Peter Flanagan-Hyde*

#### 22 ***STATS* Puzzler**

Infinite Wisdom  
*Schuyler W. Huck*

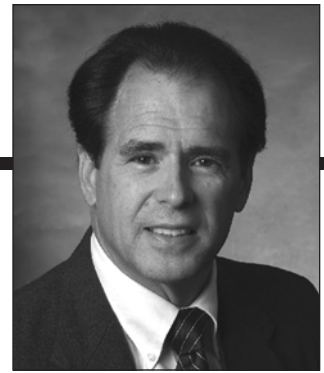
#### 23 **R U Simulating?**

'Random' Numbers from  
Nonrandom Arithmetic  
*Bruce Trumbo*

#### 28 **Statistical –sings**

I Salute You  
*Chris Olsen*

# EDITOR'S COLUMN



Paul J. Fields

If you were flying along in a single-engine airplane and then the engine quit, I bet it would get your attention. You would probably wonder what you could do to increase your chances of safely getting back on the ground. Catherine Cavagnaro is an experienced pilot, a flight instructor, and a statistics professor. In our lead article, she shows how a matched-pairs t-test could help you know what to do. She explains how she designed and conducted an experiment to learn about the glide characteristics of her airplane. Then, using the matched-pairs t-test, she found a result that might surprise you. It surprised me! She also asks us to test a claim by the aircraft manufacturer against the data from her experiment. See what you conclude when you do the test.

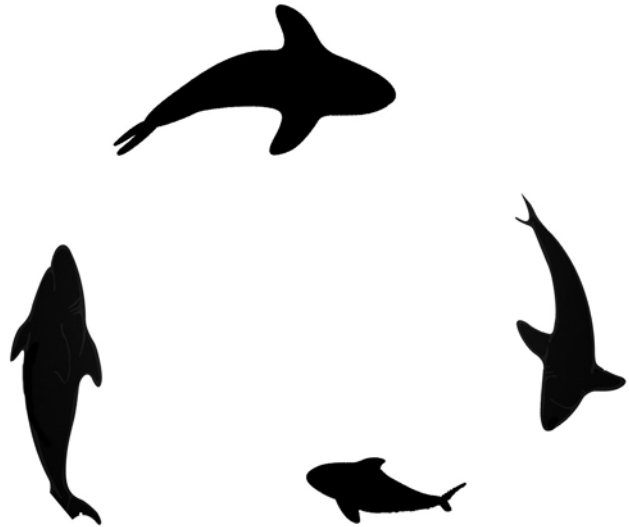
Speaking of t-tests, check out the first “Statistical Snapshot” in this issue. It asks: What does a t-test test?

When you studied t-tests, you probably heard about that *magic* number,  $n = 30$ . Have you ever wondered what is so magic about  $n = 30$ , anyway? Well, *STATS* asked David Moore to tell us.

He answered, and then graciously provided some questions (and answers) of his own. I am sure you will find his list of frequently asked questions a ready reference worth coming back to many times as you study and use statistics.

One of the questions in Moore’s FAQs is about outliers. There is a second “Statistical Snapshot” in this issue, and it looks at what we should do with outliers. Should we delete them?

Our feature article is by Milo Schield. He explores with us the mysteries of confounding and the possible paradoxical consequences due to lurking variables. He presents an illuminating graphical approach to seeing what is going on. Not only is his technique eye-opening, it’s also easy to use. Schield also gives us a problem to solve using his technique, so try it and then compare your answer to his.



In “AP Statistics,” Peter Flanagan-Hyde discusses various types of observational studies and their importance in research. The role of observational studies can be misunderstood, and he helps us see how it all fits together.

And since Cavagnaro started us off thinking about what is wise to do, consider “*STATS* Puzzler’s” Infinite Wisdom. See if you can solve this puzzle—I bet the solution will surprise you.

Have you ever wondered where random numbers originate? Out of a hat, maybe? Seriously, how can we get random numbers from nonrandom numbers? In Bruce Trumbo’s “R U Simulating?” column, we’re shown how. Try your numerical analysis skills on his challenges and I promise you will learn something new.

In “Statistical  $\mu$ -sings,” Chris Olsen offers a salute to teachers and students of statistics. As this new school year gets under way, all of us at the American Statistical Association join him in his salute—keep up the great work!

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

Paul J. Fields

# Glide Testing: a Paired Samples Experiment



Catherine Elizabeth Cavagnaro

On July 23, 1983, Air Canada Flight 143—a brand new Boeing 767—grew eerily quiet as it traveled above the Canadian countryside. In ordering fuel for the flight, the pilot had made a unit conversion error and, consequently, received an insufficient fuel supply. With no airport in the vicinity, the pilot directed his aircraft at a speed of 220 knots toward an abandoned airbase in Gimli, Manitoba, and braced for an emergency landing.

What can a pilot in command of such an unintentional glider do to reach the most forgiving terrain? Upon engine failure, a powered airplane does not just fall from the sky. Rather, the craft becomes a glider, albeit a rather inefficient one. To maximize the horizontal distance traveled, or “glide distance,” a pilot must use the cockpit yoke control to achieve the optimal airspeed. Furthermore, any reduction in the drag force that opposes motion through the air also will increase the glide distance. For example, the landing gear and the flaps should be retracted until needed for landing. But, what about the propeller on a propeller-driven airplane? Should the pilot let it spin freely or stop it from spinning?

## Flying and Gliding

First, let’s take a look at how an airplane flies. See Figure 1. Propeller-driven aircraft use an engine and rotating propeller to generate a thrust force parallel to the flight path that moves the wings through the air. Drag acts opposite to the thrust. As an airplane wing moves through the air, it creates a lift force perpendicular to the path of the airplane. Once the plane is airborne, thrust is not

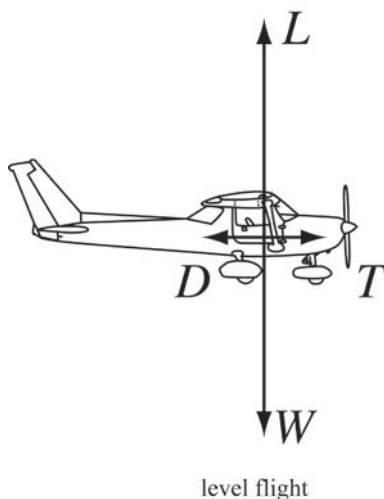


Figure 1. In level flight, the opposing forces of lift ( $L$ ) and weight ( $W$ ) and thrust ( $T$ ) and drag ( $D$ ) cancel each other.

necessary to create the lift, but, in its absence or without an air current rising from below, the plane will descend necessarily as the weight force pulls it toward the Earth.

By Newton’s law of motion for straight, unaccelerated flight including climbs and descents, the lift, weight, thrust, and drag forces sum to zero. In straight and level flight, lift is approximately weight and thrust is approximately drag. With engine failure, as we see in Figure 2, the thrust goes to zero and the weight can be decomposed into the component that opposes drag,  $W \sin \gamma$ , and that which opposes lift,  $W \cos \gamma$ , where  $\gamma$  represents the glide angle or angle between the flight path and horizontal and  $W$  is the weight of the airplane. For small angles  $\gamma$ , we have that  $\sin \gamma$  is approximately  $\gamma$  (denoted  $\sin \gamma \sim \gamma$  and  $\cos \gamma \sim 1$ ). Letting  $L$  denote lift

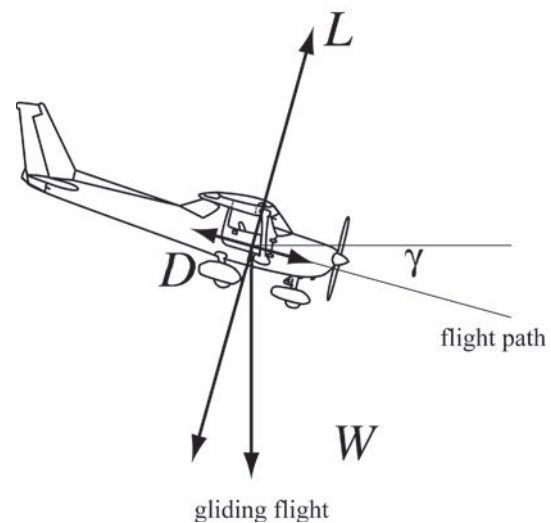


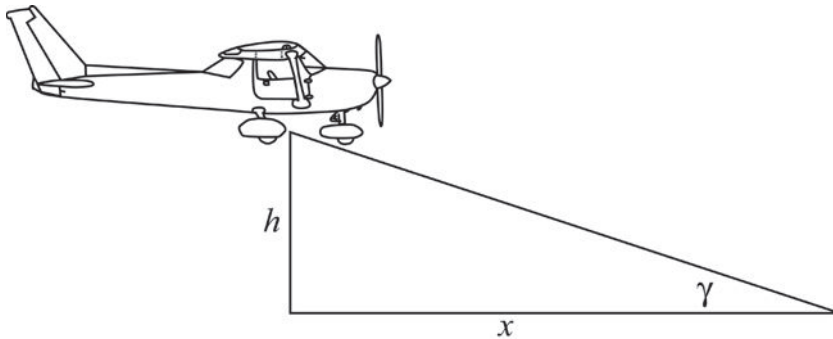
Figure 2. In gliding flight, with zero thrust in gliding flight, the components of weight parallel and perpendicular to the flight path oppose drag and lift, respectively.

and  $D$  denote drag, then  $L \sim D$ ,  $\gamma \sim D/L$ , and any reduction in drag or drag to lift ratio will reduce the glide angle and extend the glide distance.

Suppose an airplane experiences an engine failure at height  $h$  feet above the ground, as in Figure 3. The airplane will glide toward the ground with an airspeed—the rate traveled along the glide path—governed by the yoke control in the cockpit. Pushing the yoke forward moves the nose down and increases the airspeed, and pulling

back affects the opposite. To increase the likelihood of reaching an airport or a suitable alternative place to land, the pilot needs to maximize the glide distance  $x$ . We can see that, again for small angles,  $\tan \gamma \approx h/x$ , so maximizing the glide ratio  $x/h$  involves maximizing the lift-to-drag ratio  $L/D$ . For each aircraft at a specific weight, there is one airspeed,  $v_{bg}$ , or best glide airspeed, that achieves this goal.

Although manufacturers of small, single-engine aircraft are, in fact, required to make this determination, the Code of Federal Regulations does not require that



**Figure 3:** To maximize the glide ratio  $x/h$ , maximize the lift-to-drag ratio  $L/D$ .

$v_{bg}$  be determined for transport category aircraft. Thus, the pilot of the “Gimli Glider,” as it is now called, had to take a guess that the best glide airspeed for a Boeing 767 is 220 knots.

## Stopping the Propeller

Besides holding the yoke to achieve the best glide airspeed, is there anything else a pilot can do to extend the glide? Reducing the drag force can help. Upon engine failure in a propeller-driven aircraft, the propeller will continue to turn, or “windmill,” as air passes over it. Aviation literature has long reported that drag from this turning propeller is responsible for a considerable decrease in performance. In fact, twin-engine aircraft are equipped with a mechanism to stop the propeller of an inoperative engine. Although single-engine aircraft are not designed with such capability, it’s reasonable to expect that drag reduction from a stopped propeller will increase the horizontal distance achieved per unit of altitude loss – the glide ratio. For single-engine airplanes, the glide ratio is approximately 10:1. Any increase in this ratio offers an unfortunate pilot beset with engine failure a greater likelihood of reaching an airport or terrain suitable for an emergency landing. By pulling the yoke control back, the pilot can force the propeller to stop by slowing the airflow over it. Once it has stopped, internal engine friction guarantees that only speeds much higher than  $v_{bg}$  will allow it to turn once more.

How much will glide ratio increase with a stopped propeller? Although Cessna Aircraft reputedly witnessed a 20% increase in the two-seat 150 model and the same for the four-seat Cessna 172, we found little information on the tests. Barry Schiff reports an increase in glide ratio in a test he conducted using the four-seat Cessna 182 in his AOPA Pilot article, “Stopping the Propeller: Buying the Most Distance When the Engine Quits.”

To estimate the impact of the drag associated with a windmilling propeller, we planned 25 test flights in a 1979 two-seat Cessna 152, the successor to the 150 model, at Franklin County Airport, Sewanee, Tennessee. Sewanee is located on the edge of the Cumberland Plateau in southeast Tennessee. We selected days with calm winds so disturbances caused when higher winds meet the plateau were minimized. A matched-pairs experimental design held the promise of minimizing the effects of these and other atmospheric phenomena, assuming conditions would not vary too much for consecutive runs.

## Don’t Try This at Home

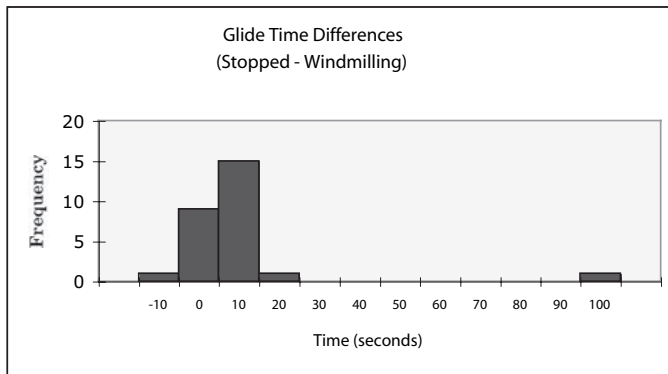
Stopping the propeller of a single-engine airplane is not for the faint of heart. In our tests, with the throttle closed and the plane’s nose held high, the airframe of the Cessna 152 shuddered as its once invisible propeller was coaxed to a halt. The airspeed at which this is possible is much lower than the best glide airspeed, and is close to the airspeed at which a plane does, in fact, fall from the sky. After the propeller came to a halt, the nose was lowered to achieve the desired airspeed. After the test glide was completed, we lowered the nose to achieve an airspeed sufficient to start the propeller turning again. Pilots wishing to experience such a flight condition should attempt the procedure only with an instructor who has such experience in that make and model of aircraft.

## The Test

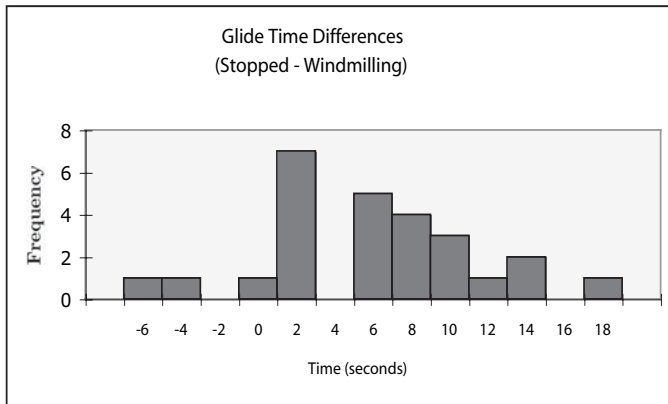
Glide distance in a no-wind situation can be shown to be proportional to the time spent in the descent. Therefore, the test compares the difference in time to descend in the two conditions. We climbed above 8,000 feet mean sea level (MSL), stabilized at the best glide airspeed of 60 knots with the propeller either windmilling or stopped, and recorded the time to descend to 7,200 feet MSL. We then repeated the glide with the propeller in the other condition. The order that the propeller was either windmilling or stopped was randomized. Blinding was not possible in our experiment, as it is impossible to keep that condition from the pilot because a stopped propeller is difficult to ignore and results in an airplane that is uncharacteristically quiet. The times for 27 paired trials are shown in Table 1.

**Table 1.** Glide Times for 27 Trials and Their Paired Differences for all 27 trials

Trial	Windmilling	Stopped	Difference	Trial	Windmilling	Stopped	Difference
1	73.4	82.3	8.9	15	64.2	82.5	18.3
2	68.9	75.8	6.9	16	67.5	81.1	13.7
3	74.1	75.7	1.6	17	71.2	72.3	1.1
4	71.7	71.7	0.0	18	75.6	77.7	2.1
5	74.2	68.8	-5.4	19	73.1	82.6	9.5
6	63.5	74.2	10.8	20	77.4	79.5	2.1
7	64.4	78.0	13.6	21	77.0	82.3	5.3
8	60.9	68.5	7.6	22	77.8	79.5	1.7
9	79.5	90.6	11.1	23	77.0	79.7	2.7
10	74.5	81.9	7.4	24	72.3	73.4	1.1
11	76.5	72.9	-3.6	25	69.2	76.0	6.8
12	70.3	75.7	5.4	26	63.9	74.2	10.3
13	71.3	77.6	6.3	27	70.3	79.0	8.7
14	72.7	174.3	101.6				



**Figure 4.** Histogram of glide time differences for all 27 trials



**Figure 5.** Histogram of glide time differences after removing trial 14

## Data Analysis

Figure 4 shows the frequency of the difference in times for each pair of trials. We can see that trial 14 constitutes a distinct outlier. We knew something was amiss when we stopped the propeller and glided down 800 feet in almost three minutes, more than twice the usual time. Whether the cause was an unusual updraft from wind hitting the plateau below us or a giant hand holding us up, we wish such good luck on any pilot unfortunate enough to experience a genuine engine-out condition.

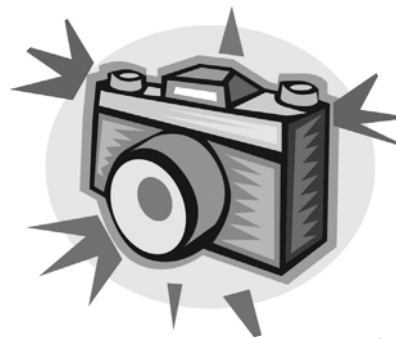
Judging trial 14 to be due to abnormal conditions and, thus, a nonrepresentative event, we can remove it from the data. Then, the data appear to be roughly symmetric with no obvious outliers, so using *t*-procedures is justified (see Figure 5).

Our null hypothesis is that the true mean difference in times to descend 800 feet (stopped minus windmilling) is zero versus the alternative that the difference is positive. Using Microsoft Excel's paired *t*-test procedure, Table 2 shows that a miniscule *p*-value of approximately 0.0000052 allows us to reject the null hypothesis and conclude that the airplane we tested will glide farther with a stopped propeller following engine failure.

**Table 2.** Data Analysis of Means for Paired Samples Using Microsoft Excel

t-Test: Paired Two Sample for Means		
	Stopped	Windmilling
Mean	77.44	71.52
Variance	24.12	25.62
Observations	26	26
Hypothesized Mean Difference	0	
df	25	
t Stat	5.49426	
one-tail probability	0.0000052	
t Critical one-tail	1.70814	

In our test, the mean increase in flight time was 5.9 seconds. A 95% confidence interval for the difference in seconds is [3.7, 8.1]. In this test, we therefore witnessed an increase in the mean glide distance of approximately 8.3%, as glide distance is proportional to time of decent.



## What Does a t-Test Test?

As one step in making a new drug, a pharmaceutical company uses micro-organisms to produce batches of a protein. The target yield of such batches is  $\mu_0 = 100$ . It is a problem if  $\mu$  is either too small or too large. We want to know if the target yield is being achieved.

With observed yields from  $n = 10$  batches, we use the standard t-test of  $H_0: \mu = 100$  against  $H_A: \mu \neq 100$  at the 5% level, rejecting  $H_0$  if the absolute value of

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

exceeds 2.262. Assuming  $\mu_0 = 100$ , for what values of the sample mean  $\bar{X}$  and standard deviation  $S$  is  $H_0$  rejected? You might suppose we reject mainly when a sample happens to have 'too large' a value of  $\bar{X} - \mu_0$ .

Figure 1 plots  $S$  against  $\bar{X}$  for 10,000 simulated samples of size 10 from a normal population with  $\mu = 100$  and  $\sigma = 10$ , emphasizing (by darker dots) the roughly 500 samples for which  $H_0$  is wrongly rejected. Figure 1 shows that values of  $S$  that are too small can play an important role in incorrect rejection. By 'too small,' we are saying the sample standard deviation is less than the population standard deviation, when we are assuming they are equal.

An important fact about normal data is that  $\bar{X}$  and  $S$  are independent. In the same figure, the cloud of all 10,000 simulated points illustrates this principle. There is no clear pattern of association between  $\bar{X}$  and  $S$ . Also, their correlation is 0 within the accuracy of the simulation.

### Try This at Home

Use the data above to see if the difference we witnessed in the Cessna 152 is consistent with the 20% that was reported for other Cessna aircraft. After you have done your analysis, compare your results to our analysis on Page 13.

### Conclusion

So, when the engine quits, stop that prop! We have found highly persuasive evidence that stopping the propeller does improve glide performance and the pilot of a propeller-driven aircraft may want to consider this prospect if altitude and experience permit. Incidentally, our experimental results in a Cessna 152 are consistent with trials conducted in the author's 1973 Piper Cherokee 140.

The captain of Flight 143, an experienced glider and aerobatic instructor, atoned for his mathematical slip by executing a successful emergency landing at Manitoba's air base. On the Gimli Glider, windmilling turbine fans in the jet engines—similar to propellers—created drag that hindered the plane's glide performance. Perhaps a fan-stopping mechanism on jet engines is in order. Fortunately, engine failure incidents, like those resulting from fuel exhaustion, are exceedingly rare in commercial aviation. Still, information on best glide airspeed and procedures that minimize drag would have been useful to the pilot of the Gimli Glider, or any other pilot in such unfortunate circumstances. ■

**Editor's Note:** *The author would like to thank William K. Kershner, who provided the Cessna 152 and the idea for the test. These results appear in his book, The Flight Instructor's Manual.*

### Additional Reading

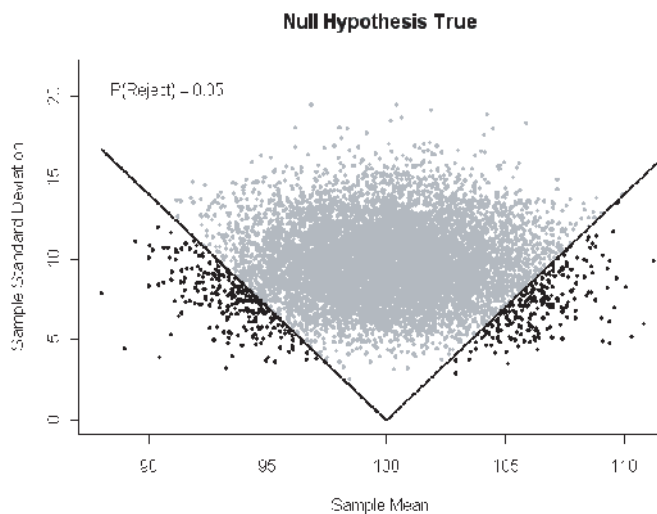
Federal Aviation Administration. (2005). *Code of Federal Regulations, Title 14: Aeronautics and Space*. Available at [http://faa.gov/regulations\\_policies](http://faa.gov/regulations_policies).

Kershner, William K. (2002). *The Flight Instructor's Manual (4th ed.)*. Blackwell Publishing.

Schiff, Barry. (1995). "Stopping the Propeller: Buying the Most Distance When the Engine Quits." *AOPA Pilot*, Aircraft Owners and Pilots Association.

*Catherine Cavagnaro, ccavagna@sewanee.edu, is an associate professor of mathematics at the University of the South in Sewanee, Tennessee, who enjoys teaching elementary statistics. She is also a flight instructor who specializes in aerobatics and spin training at the Franklin County Airport.*





**Figure 1.** Sampling with a sample size of 10 from a normal distribution with  $\mu = 100$  and  $\sigma = 10$  and testing  $H_0: \mu = 100$  against  $H_A: \mu \neq 100$  when the null hypothesis is true, the dark dots show samples that reject  $H_0$  and the pale dots show samples that do not reject  $H_0$ .

In monitoring the pharmaceutical production of protein, it is especially important to know if the actual value of the population mean differs from 100 by more than 10. What are the chances of rejecting  $H_0$  when  $\mu = 110$ ?

To investigate this, we simulate 10,000 samples of size 10 from a normal distribution with  $\mu = 110$  and  $\sigma = 10$ . The results, displayed in Figure 2, show that the power of the t-test in this situation is 80%—when 110 is the true population mean, about 8,000 of the samples rejected  $H_0$ .

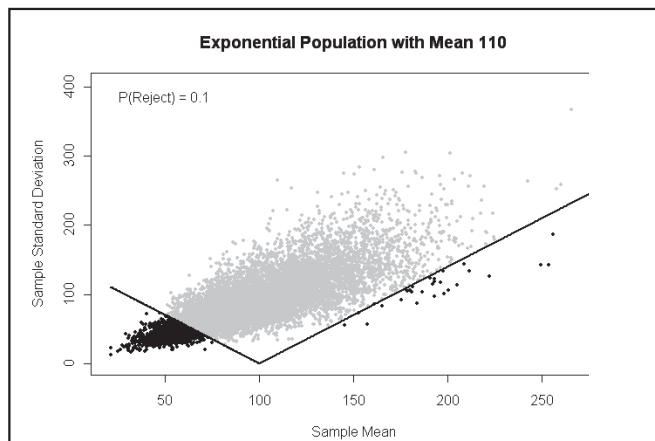


**Figure 2.** Sampling with a sample size of 10 from a normal distribution with  $\mu = 110$  and  $\sigma = 10$  and testing  $H_0: \mu = 100$  against  $H_A: \mu \neq 100$  when the null hypothesis is false, the dark dots show samples that reject  $H_0$  and the pale dots show samples that do not reject  $H_0$ .

Here again, the size of  $S$  is as important as that of  $\bar{X} - \mu_0$ . In Figure 2, some rejecting samples (dark dots) with small values of  $S$  had  $\bar{X}$  around 105, and some not rejecting ones (pale dots) with large sample standard deviations had sample means around 110.

By commonly accepted standards of judging tests, the t-test is the best for normal data when both parameters are unknown. We are just seeing the consequences of testing hypotheses about  $\mu$  when  $\sigma$  is unknown. (If  $\sigma$  was known and the z-test was used to test  $H_0$ , the dividing lines between dots rejecting and not rejecting would be vertical, instead of diagonal, as in our figures.)

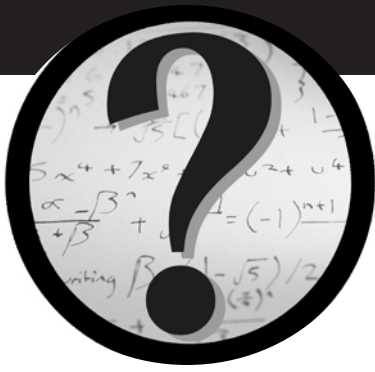
For data from a non-normal population, if we use the t-test, the probability of rejection will not be the  $\alpha$ -value specified in the test. For one reason,  $\bar{X}$  and  $S$  may no longer be independent. As an extreme example, Figure 3 shows what happens if we draw samples of size 10 from an exponential population with  $\mu = 100$  and, thus, necessarily  $\sigma = 100$ . The true significance



**Figure 3.** Sampling with a sample size of 10 from an exponential distribution with  $\mu = 100$  and  $\sigma = 100$  and testing  $H_0: \mu = 100$  against  $H_A: \mu \neq 100$  when the null hypothesis is true but the sampled population is non-normal, the dark dots show samples that reject  $H_0$  and the pale dots show samples that do not reject  $H_0$ .

level of the t-test with critical value 2.262 is only about 10%, instead of 5%. Moreover, the power of this test (not shown), when  $\mu = 110$ , is only about 15%.

So, the t-test is not purely a test of whether  $H_0: \mu = \mu_0$  is true. The results of the t-test can be affected by a sample standard deviation that underestimates the population standard deviation and by departures from normality. ■



# ASK STATS



Jackie Miller

## Why is $n = 30$ 'magic'? ...and Other Frequently Asked Questions



David Moore

---

Jackie Miller ([miller.203@osu.edu](mailto:miller.203@osu.edu)) is a Statistics Education Specialist and auxiliary faculty member in the Department of Statistics at The Ohio State University. She earned both a BA and BS in mathematics and statistics at Miami University, along with an MS in statistics and a PhD in statistics education from The Ohio State University. She is very involved in the statistics education community. When not at school, Miller enjoys a regular life (despite what her students might think), including keeping up with her many dogs!

We asked David Moore, renowned educator and author, a couple of burning questions about statistics. He responded in grand style.

### **1. Why is $n = 30$ the magic number for switching from t to z in inference about a single mean?**

It isn't. There is no magic number. The distinction between z and t procedures for a mean has nothing to do with sample size. Use z if you know the population standard deviation ( $\sigma$ ), and use t if you don't. You almost never know  $\sigma$ , so you should simply always use t.

Of course, the t distributions approach standard normal as the degrees of freedom increase, so inference from z becomes a better approximation to inference from t as the sample size increases. In the dark ages, when we used tables with one line for each sample size, about 30 lines would fit on a page. So, after  $n = 30$ , we were advised to switch from exact t to approximate z. This last made sense when we last used a table to compute square roots.

### **2. Well, then, is $n = 30$ the magic number for safe use of t for nonnormal data?**

No. There is no magic number. Inference based on t, and inference about means in general, is reasonably robust against lack of normality. That's why various t statistics, one-way ANOVA, and the like are useful in

practice. For a given nonnormal population distribution, accuracy of  $t$  inference does improve with  $n$ . But a one-sample  $t$ -test is more robust against nonnormality when the true distribution is approximately symmetric than for skewed distributions. That's why there is no overall magic number. For roughly symmetric distributions, I find the  $t$ -test strikingly robust, so that  $n = 15$  is adequate for practical conclusions.

Three codas to this, minor FAQs in themselves. First, inference about spread for normal populations (such as the  $F$ -test for comparing variances) is notoriously sensitive to nonnormality. I think these procedures should never be used. This is a place to start learning about permutation tests and bootstrap confidence intervals, which aren't based on a specific form for the population distribution. Second, if you have these resampling procedures in your software, you can routinely check the robustness of  $t$  by comparing  $p$ -values or confidence intervals. I do this, and  $t$  is pretty impressive. Third, the variable robustness of  $t$  to varied forms of nonnormality reminds us why a normal quantile plot is much more helpful than any formal test of normality: The plot shows how the data deviate from normality. What you want normality for is as important as how nonnormal the data are. A  $p$ -value for a test of normality doesn't answer the question, "Are these data normal enough for my purposes?"

### 3. You say just always use $t$ for inference about the mean of a roughly normal population. So why does $z$ for means still appear in texts?

A text has to start at the beginning: What is the basic reasoning of inference? What specific procedures operationalize this reasoning? What are the practical

barriers to effective use of these procedures? Starting at the beginning, how shall we introduce beginners to the reasoning of inference? This is a pedagogical issue, not a question of statistics in practice. Some teachers may prefer to start with rank tests (but discrete sampling distributions are awkward and the corresponding confidence intervals are more so). Sometime in the golden future, we will start with resampling methods. I think permutation tests make the reasoning of tests clearer than any traditional approach. For now, the main choices are  $z$  for a mean and  $z$  for a proportion.

I find  $z$  for means quite a bit more accessible to students. Positively, we can say up front that we are going to explore the reasoning of inference in an overly simple setting. Remember, an exactly normal population and a true simple random sample (SRS) are as unrealistic as known  $\sigma$ . All the issues of practice—robustness against lack of normality, application when the data aren't an SRS—are put off until, with the reasoning already in hand, we discuss the practically useful  $t$  procedures. This separation of initial reasoning from messier practice works well.

Negatively, starting with inference for proportions introduces many side issues: (1) No exact normal sampling distribution, but a normal approximation to a discrete distribution; (2) use of  $\hat{p}$ , both in the numerator and denominator of the test statistic, to estimate both the parameter  $p$  and  $\hat{p}$ 's own standard deviation; and (3) loss of the direct link between test and confidence interval.

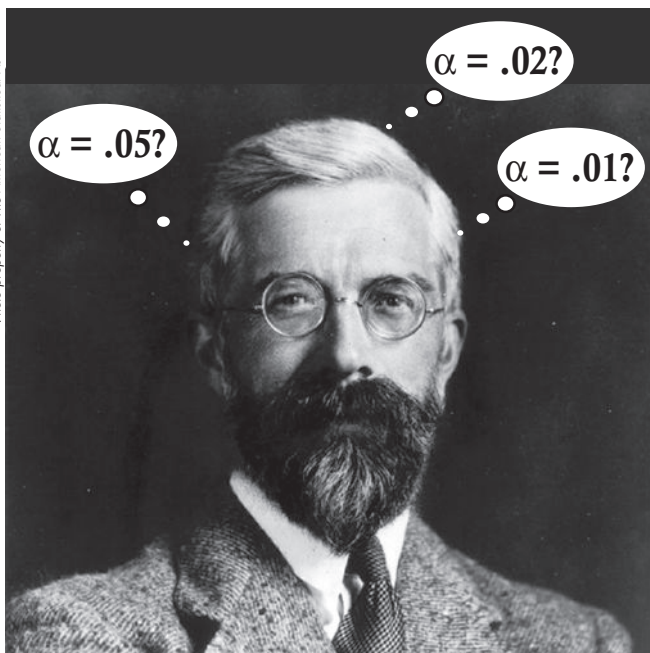
Once upon a time, we had at least the compensation of developing practically useful procedures. Now, the often gross inaccuracy of the traditional  $z$  confidence interval for  $p$  is better understood. It is hard to abandon old friends, but I think a look at the graphs in Section 2 of the paper by Brown, Cai, and DasGupta in the May 2001 issue of *Statistical Science* is both distressing and persuasive. (See also the article "How Much Confidence Should You Have in Binomial Confidence Intervals" by Seuss et al. in the Spring 2006 issue of *STATS*). The standard intervals often have a true confidence level much less than what was requested, and requiring large samples encounters a maze of 'lucky' and 'unlucky' sample sizes until very large samples are reached. There are countermeasures, but this is exactly the kind of practical difficulty we should avoid when beginning to teach the reasoning of inference.

### 4. Where did the popular significance level of 0.05 come from?

From the master, himself, R. A. Fisher:

...it is convenient to draw the line at about the level at which we can say: Either there is something in the treatment, or a coincidence

Photo property of The American Statistical Association



R.A. Fisher

has occurred such as does not occur more than once in twenty trials. ...If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the two percent point) or one in a hundred (the one percent point). Personally, the writer prefers to set a low standard of significance at the five percent point, and ignore entirely all results which fail to reach that level. A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely* fails to give this level of significance. (“The Arrangement of Field Experiments,” *Journal of the Ministry of Agriculture of Great Britain*, (1926) 33: 504.)

Of course, reading his writings more extensively, Fisher was not an advocate of blind use of significance at the 5% level as a yes-or-no criterion. And he was writing before fast and cheap computing made actual p-values immediately available. Never forget that, in practical settings, there is no meaningful difference between  $p = 0.049$  and  $p = 0.051$ . So “ignore entirely” is lousy advice, even if it did come from Fisher.

### **5. I have trouble deciding when an observation is an outlier. Can I just use the $1.5 \times \text{IQR}$ criterion as a rule?**

Outliers are observations that “seem to stand apart” from the overall distribution and, therefore, deserve to be investigated. “Standing apart” is a matter for judgment. Although students would like a recipe for what makes an outlier, there is none. One sometimes sees rules based on standard scores, such as  $|Z| > 2$ . These are flatly wrong. They confuse the most extreme observations in a distribution with outliers, saying for example that the most extreme 5% in a normal distribution are always outliers.

Mainly, the  $1.5 \times \text{IQR}$  criterion is intended for automatic searching and plotting—it identifies only “suspected outliers.” You can calculate that, for normal distributions, the  $1.5 \times \text{IQR}$  criterion flags observations more than about 2.7 standard deviations from the mean. But this misses the point: The standard deviation is not a suitable descriptive measure of spread for skewed distributions. Because the two tails have different spreads, no one number can do a good job. IQR describes the spread of the middle half of the data, so the  $1.5 \times \text{IQR}$  criterion looks at the position of extreme observations relative to the spread of the center part of the distribution. That’s why we use this criterion, rather than one based on the standard deviation.

Nonetheless,  $1.5 \times \text{IQR}$  can point to observations that are not outliers. This happens, for example, in a distribution with a quite compact and symmetric center half but a long tail. It remains a matter of judgment whether an observation is an outlier or just the largest or smallest in a long-tailed distribution. Basically, if it doesn’t jump out at you, it isn’t an outlier. Students need to learn

to see the big effects: major peaks in a distribution, rather than minor ups and downs in a histogram; clear skewness, rather than slight imbalance between the two sides of a distribution; and serious outliers, rather than just the largest and smallest observations.

### **6. Excel’s Data Analysis menu offers both “t-Test: Two-Sample Assuming Equal Variances” and “t-Test: Two-Sample Assuming Unequal Variances.” How do I decide which to use?**

Excel isn’t the only offender here, but the wording “Assuming Unequal Variances” is seriously misleading. There’s an outdated version of the two-sample *t*-test that does assume equal variances. You should never use it, in part because it’s very hard to know if the population variances are equal. Most statistical software packages have an accurate approximation that works very well whether or not the population variances are, in fact, equal. That is, it doesn’t assume anything about variances. Good software should use this version by default, usually with an option to assume that the variances are equal. If your technology offers an equal/unequal choice, just choose unequal.

### **7. I saw an examination item that asked for the primary purpose of blocking. Two of the choices given were “reduce variation” (correct) and “reduce confounding” (not correct). I thought we formed blocks to reduce the effect of lurking variables. So doesn’t this reduce confounding?**

Consider an experiment comparing two treatments for a “Dread Disease.” A completely randomized design assigns subjects at random to two groups. Suppose now that women and men differ systematically in their response to the two treatments. This systematic difference increases the variation of responses in both treatment groups, making it harder to assess overall differences in the mean responses to the treatments. There is, however, no confounding of gender with treatment because randomization will (on average, given enough subjects) balance the groups in gender.

A randomized block design separately randomizes women and men, allowing direct comparison of the two treatments for each gender separately. The variation in responses in each block is less than before. Thus, blocking reduces variation by dealing with a specific cause of variation systematically, rather than leaving it to randomization.

That’s an adequate explanation at the level of introductory statistics, where we treat confounding conceptually, rather than attempting a formal definition. At a more advanced level, we would try to distinguish confounding from interaction (there is an interaction between gender and treatment because the difference in the mean response to the two treatments changes with gender). This can get subtle, and, by some standards, we might say that both confounding and interaction are

present in this example when we do not block. This does not change the fact that the main purpose of blocking is generally to reduce variation by removing the systematic effect of the blocking variable.

**8. I read that “regression to the mean” describes why students who do well on a midterm exam tend to do less well on the final. I looked at the algebra and found this is only true if the slope of the regression line is less than 1.**

That’s almost right. The slope of the least-squares line is  $b = r s_y / s_x$ . So if the two exams have the same scale and the same spread of scores, the slope is  $r$  and is less than 1. But suppose, for example, that the midterm has a 0 to 50 scale and the final a 0 to 100 scale. We don’t expect “do less well on the final” to hold in points. Instead, it holds in the standard scale: The final score will (on the average) be fewer standard deviations above the mean than the midterm score. This also adjusts for exams on the same scale, say 0 to 100, that differ greatly in difficulty.

**9. I see examples of significance tests apparently applied to entire populations. Doesn’t inference draw conclusions about a population on the basis of sample data?**

This is a bit subtle. Confidence intervals do not make sense if we have data on an entire population. If we have the salaries of all full professors, for example, we know the population mean and have no need to estimate it. But tests do make sense. We can ask whether the difference between the mean salaries of female and male full professors is statistically significant. That is, is it so large that it would rarely occur just by chance? Here’s one way to do it. Start with the given set of salaries. If there’s no gender effect, any salary is equally likely to belong to a man or a woman. So look at all possible random assignments of the salaries to the professors and see how often the mean male-female difference is as large as that actually observed. This gives the  $p$ -value for a permutation test of the null hypothesis of no gender effect. The  $t$ -test, by the way, is an approximation to this result, so, in practice, we often just use the  $t$ -test on the population of all salaries.

**10. What’s the most exciting discovery in statistics that you have experienced in your career?**

That’s easy: resampling procedures—bootstrap confidence intervals and permutation tests—made practical by very cheap and very fast computing. These already have appeared in several of my responses above. As understanding and use spreads, they will replace much of the inference now taught in introductory statistics courses.

By the way, “most important discovery” is not the same as “most important trend.” The most important

trend is the movement of statistics somewhat away from mathematics (though you can still never know too much mathematics) back toward its roots in data analysis and scientific inference. Fast and cheap computing is again a driving force. We see much more clearly now that formal probability-based inference, though important, has a more restricted domain of use than data analysis based on data graphics and effective summaries. For very large datasets, this also means “based on good algorithms.” The computer science aspect of statistics will become more important.

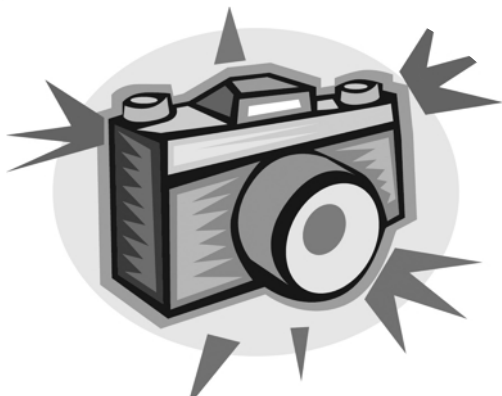
**11. What do you find to be exciting new areas of statistics that will be “hot topics” in the near future?**

I’ve mentioned the interplay between statistics and computer science. My perception of “most important trend” implies that the most exciting new areas will be just that: effective application of statistical ideas and tools to new scientific problems. Think of the explosion of statistical applications in molecular biology, genomics, and proteomics. New applications will, as in this example, drive development of new techniques. It’s a great time to be a statistician. ■

---

**About David Moore:** Moore is Shanti S. Gupta Distinguished Professor of Statistics, emeritus, at Purdue University. He received his AB from Princeton (1962) and PhD from Cornell (1967), both in mathematics. He has written many research papers in statistical theory and served on the editorial boards of the Journal of the American Statistical Association, Technometrics, the International Statistical Review, and the Journal of Statistics Education. Moore served as program director for statistics and probability at the National Science Foundation, as a member of the National Research Council’s Committee on Applied and Theoretical Statistics, and as a member of the oversight committee for NRC’s Mathematical Sciences in the Year 2000 project. He was president of the American Statistical Association in 1998.

In recent years, Moore has devoted his attention to the teaching of statistics. He was the content developer for the Annenberg/Corporation for Public Broadcasting college-level telecourse, “Against All Odds: Inside Statistics,” and other video series. He also is the author of influential articles on statistics education and of several leading texts, including the immensely popular *The Basic Practice of Statistics*, now in its fourth edition. Moore has served as the first president of the International Association for Statistical Education and as a member of the National Research Council’s Mathematical Sciences Education Board. He also has received the Mathematical Association of America’s national award for distinguished college or university teaching of mathematics.



## Should Outliers Be Deleted?

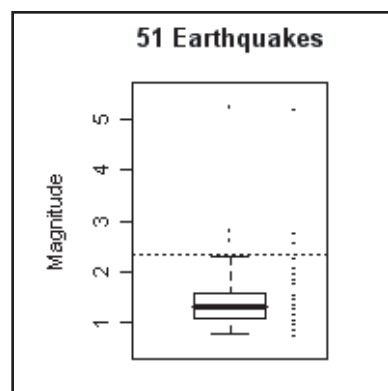
An “outlier” is a value that appears to lie an unusual distance away from the rest of the observations. To put this definition into practice, we need a way to say whether a particular observation is unusual. One way to look for outliers is with boxplots.

Boxplots use the five-number summary of a sample: minimum (Q0), lower quartile (Q1), median (Q2), upper quartile (Q3), and maximum (Q4). The box of the boxplot extends from Q1 to Q3, so its length is the interquartile range  $IQR = Q3 - Q1$  of the data. Many statistical computer packages use the “ $IQR \times 1.5$ ” rule to identify outliers. Boundaries (sometimes called lower and upper fences) are established at the points  $L = Q1 - 1.5 \times IQR$  and  $U = Q3 + 1.5 \times IQR$ . Any point outside the interval from L to U is designated as a possible outlier.

Figure 1 on the following page shows the boxplots of four datasets, the first three of which show at least one outlier. The stripchart at the right of each boxplot shows the actual data values. The upper fence, U, is shown explicitly by a dotted line in each of the three boxplots that show the possible outliers in the upper tail.

Analysts often wonder what to do with outliers. Should they be deleted? Should they be ignored? Do they represent an error that needs to be corrected? Or, are they just extreme, but correct, values? Reasonable answers to these questions always depend on the specific situation. Each of the boxplots in Figure 1 has its own story and its own answers to these questions.

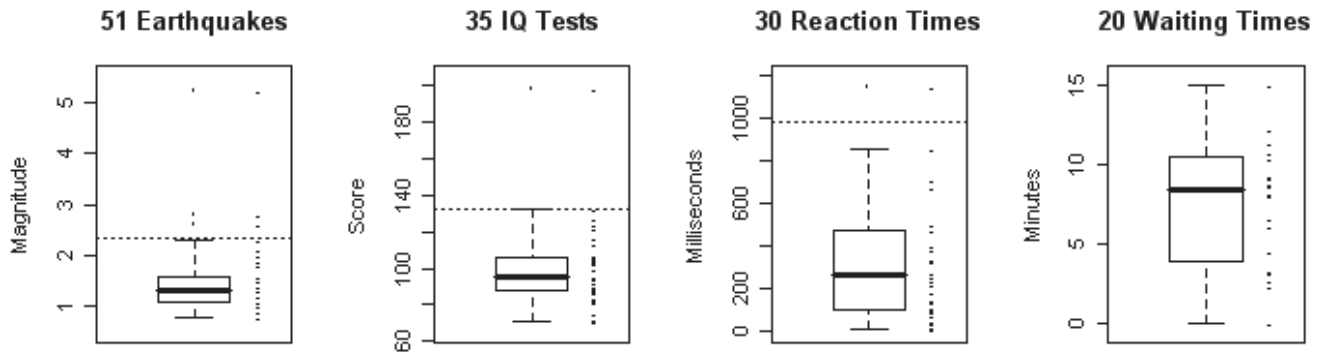
The first boxplot shows magnitudes for 51 earthquakes that occurred in California on September 3, 2000. Most of these earthquakes were so small they could be detected only by delicate equipment, but the three outliers were strong enough to be noticed by people nearby. The largest one, of magnitude 5.2, was a major quake.



The first boxplot

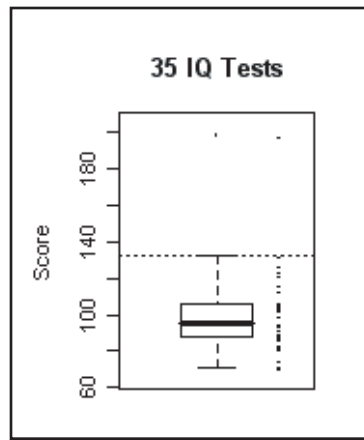
Occurring in the Napa Valley, it injured about 25 people and caused about \$50 million in damage, much of that to wineries and storehouses of fine wines. With earthquakes, only the outliers are of any interest to the general public, and only the most extreme outliers cause any damage. Certainly, a statistician dealing with earthquake data would be foolish to delete these outliers.

Another situation where outliers are not to be deleted or ignored occurs in the biotech industry where there can be a lot of background noise in the measurements. Here, the only measurements that give useful results are the outliers.

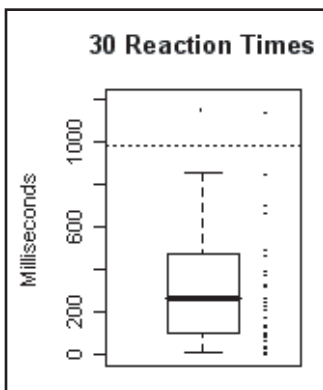


**Figure 1:** Four boxplots are shown of real data from earthquakes, student IQ tests, reaction times in a psychological experiment, and commuter waiting times. Possible outliers are indicated above the dotted line drawn  $1.5 \times$  IQR from the box.

The second boxplot shows IQ scores recorded for a sample of 31 high school students. In this case, the maximum value 198 is an outlier. The researchers recognized instantly that this is an absurd IQ value. Fortunately, the original test sheets were still available, so they were able to find that one score of 98 had been typed incorrectly as 198 and ‘clean’ the data. This is an example of the identification of an outlier giving researchers warning to correct an error before doing further data analysis.



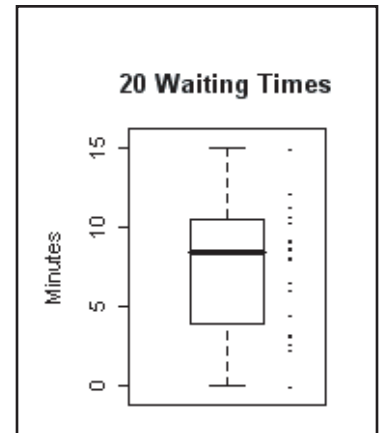
The second boxplot



The third boxplot

were no longer available, the researchers would not have known what to do with the outlier—delete it, ignore it, or correct it. The third boxplot represents reaction times for a student participating in a psychology experiment. Reaction times distributions often are strongly skewed toward higher values. In this case, the subject was supposed to push a button as quickly as possible after seeing a visual cue on a monitor. A blink of an eye, or momentary inattention, can cause an unusually long reaction time. Such data almost always show outliers. Typically in this situation, researcher should take repeated measurements and use the median to summarize the results. In this way, the data will be less sensitive to the occurrence of outliers.

The fourth boxplot illustrates the number of minutes a student waited for a commuter train on 20 trips in a particular month. Although the campus shuttle bus runs irregularly, the commuter trains run on a more controlled schedule. Essentially, possible waiting times for the commuter train are uniformly distributed in the interval from zero to 15 minutes. It would be rare to see any outliers in data from a uniform distribution, and we see none here.



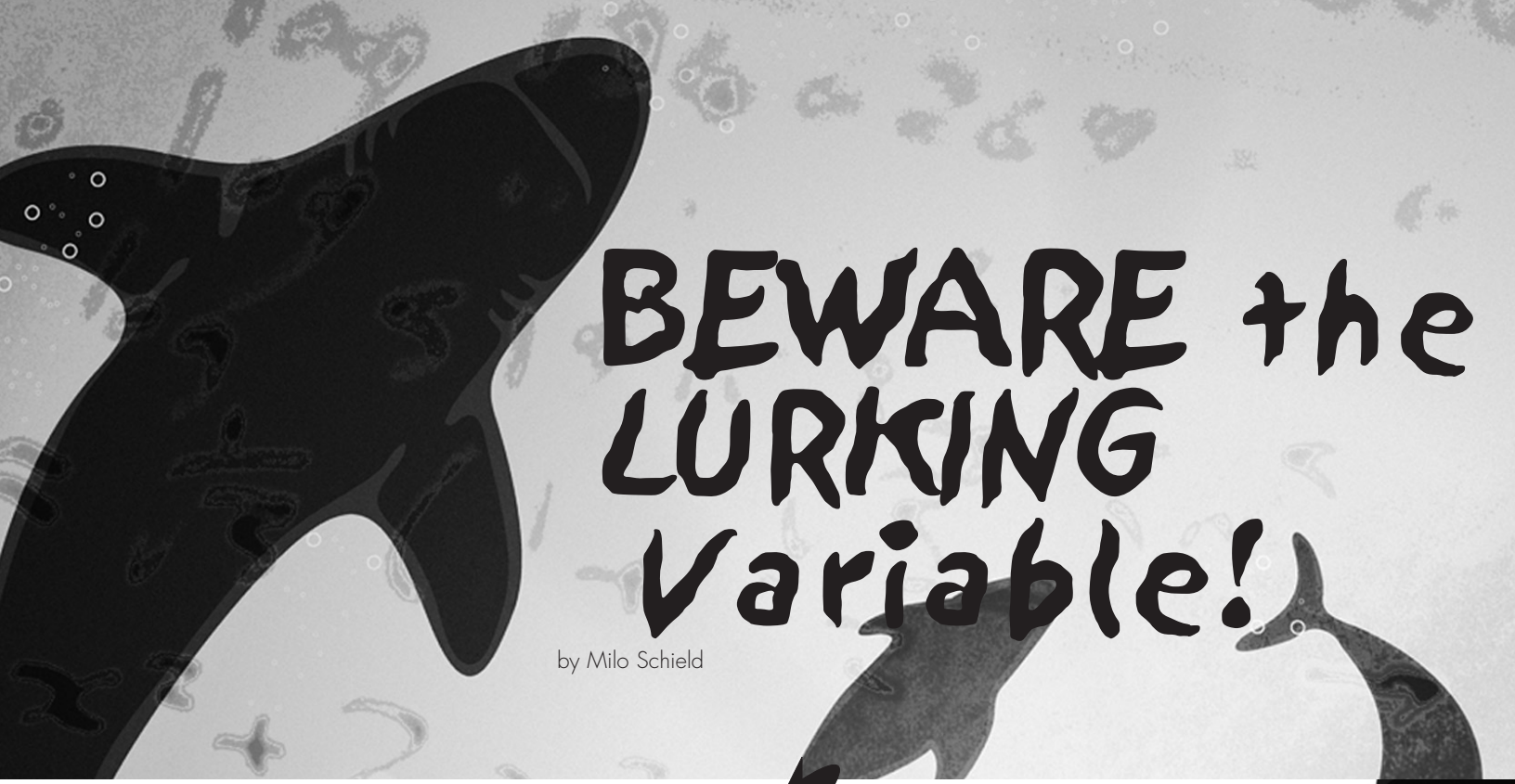
The fourth boxplot

So, we see that outliers may be caused by errors or may be an expected property of the kind of data at hand. Sometimes, it is appropriate to delete an outlier before data analysis, but often it is not. In any case, if we delete an outlier, we need to have a good reason for doing so and we need to document it in the report that goes with the data analysis. ■

## Answer from Page 6 Try This at Home

Remember that the mean increase in mean flight time was 5.9 seconds, and the mean flight time with windmilling was 71.5 seconds. Remember also that glide distance is proportional to time of decent, so the mean percentage increase in glide distance is approximately  $5.9 \text{ seconds} / 71.5 \text{ seconds} = 8.3\%$ .

Now, since the 95% confidence interval is from 3.7 to 8.1 seconds, the 95% confidence interval for glide distance can be approximated as  $3.7 / 71.5$  to  $8.1 / 71.5$ , which gives 5.2% to 11.4%. Therefore, we can state with 95% confidence that the claim of an increase of 20% is inconsistent with the observed data, as 20% does not lie within the 95% confidence interval. We conclude that the increase is not as great as 20%.



# BEWARE the LURKING Variable!

by Milo Schield

## Understanding Confounding from Lurking Variables Using Graphs

Did you know the United States has a higher death rate than Mexico? It's a fact. In 2003, the death rate was 80% higher in the United States than in Mexico (8.4 per 100,000 compared to 4.7 per 100,000).

What does this statistic mean? Does Mexico have better health care than the United States? That seems unlikely. Yet it is difficult to claim that this unexpected result is due to chance, error, or bias. The populations being studied are large, and death is definite, therefore usually counted accurately. You may be perplexed further when you learn that death rates are even lower in Ecuador and Saudi Arabia (4.3 per 100,000 and 2.7 per 100,000).

A possible explanation is **confounding**: "a situation in which the effects of two processes are not separated," according to John M. Last's *A Dictionary of Epidemiology*. Confounding can be due to a **lurking variable**. Often referred to as a confounder, Last says a lurking variable "can cause or prevent the outcome of interest ... and [is] associated with the factor under investigation."

Lurking variables are called "lurking" because they are not recognized by the researcher as playing a role in the study. Although they can influence the outcome of the process being studied, their effect is mixed in with the effects from other variables.

In comparing the death rates in the United States and Mexico, a lurking variable may be the difference in the age distributions within each population. Mexico has a much younger population than the United States. In 2003, there were 59% more people under 15 in Mexico than in the United States (32% of the Mexican population, compared to 21% of the United States population). In addition, there were more than twice as many people 65 or older in the United States as in Mexico (12% compared to 5%).

It's a fact that older people are much more likely to die than younger people. Unless we take age into account, a comparison of the crude (not accounting for age) death rates may be misleading. Mexico's comparatively low death rate is more likely due to its youthful population, rather than to its health care system.

So how can we untangle this confusion? How can we take into account the influence of a lurking variable that confounds an association?

### Standardizing

Standardization is used in demography to 'take into account' the distribution of ages within a population. It can take into account the influence of a related factor





when comparing ratios for two groups so we are not comparing “apples to oranges.” When the death rates of Mexico and the United States are standardized for age, the death rate in Mexico is higher than that in the United States.

Standardization also can take into account the influence of a related factor when comparing ratios over time for the same group. For example, according to the *2001 United States Statistical Abstract*, the crude death rate due to pneumonia was 7.4% higher in 1996 than in 1990 (33.4 per 100,000 compared to 31.1 per 100,000). But when standardized on the 1940 United States population distribution, the age-adjusted death rate due to pneumonia was 5.1% lower in 1996 than in 1990 (13.0 per 100,000 compared to 13.7 per 100,000). In this case, standardizing actually reversed the direction of the association.

### Standardizing Ratios Graphically

To ‘see’ standardization, it would be nice to have a simple technique—ideally graphical—that will take into account or ‘adjust for’ the influence of a lurking variable.

In an article that appeared in *The Roles of Representation in School Mathematics*, Lawrence Lesser featured a graphical technique for showing how an association can be influenced when the lurking variable has just two values. The graph shows how a weighted average can be obtained easily without algebra. Howard Wainer did the same in a 2002 *CHANCE* article, “The BK-Plot: Making Simpson’s Paradox Clear to the Masses.” Milo Schield used this technique to illustrate standardization in “Three Graphs To Promote Statistical Literacy,” presented at the 2004 International Congress on Mathematical Education. To see how it works, let’s consider some examples.

### Patient Death Rates by Hospital

Table 1 and Table 2 present the underlying data (hypothetical) for two hospitals: Rural Hospital and City Hospital. Patients in good condition can walk in; patients in poor condition are carried in.

**Table 1.** Death Rates of Patients by Hospital and by Condition

Death Rate	Patient Condition		
	Good	Poor	All
Rural	2.0%	7.0%	3.5%
City	1.0%	6.0%	5.5%
All	1.5%	6.5%	

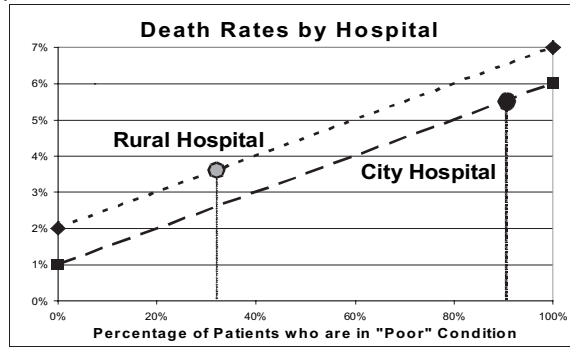
We want to analyze the association between hospital (predictor) and death rate (outcome). First, we plot the data from Table 1 in Figure 1. City Hospital has a death rate of 6% for patients in poor condition and 1% for patients in good condition. Connecting these data values gives the heavy dashed line. Rural Hospital has a death rate of 7% for patients in poor condition and 2% for patients in good condition. Connecting these data points gives the light dashed line.

**Table 2.** Number of Patients by Hospital and by Condition

Number of Patients	Patient Condition		
	Good	Poor	All
Rural	700	300	1,000
City	100	900	1,000
All	800	1,200	2,000

From Table 2, we can see that 90% of the patients in City Hospital are in poor condition, while only 30% of those at Rural Hospital are in poor condition. Plotting

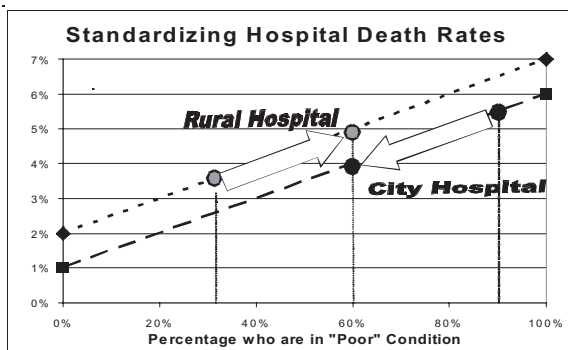
these percentages in Figure 1 gives the death rates at City Hospital and Rural Hospital.



**Figure 1.** Hospital death rates by percentage of patients in poor condition

The death rate is much higher at City Hospital (5.5%) than at Rural Hospital (3.5%). Based on this, Rural Hospital would seem like a better hospital than City Hospital. But notice that City Hospital has a lower death rate than Rural Hospital for patients in good condition and those in poor condition. This is an example of Simpson’s Paradox. Simpson’s Paradox occurs when an association has one direction at the group level, but the opposite direction in each of the sub-groups.

Before we shut down City Hospital as “the hospital of death,” we need to consider whether City’s higher death rate could be confounded by patient condition. Note that patient condition is associated with the outcome of interest (death) and with the predictor (hospital). Being in poor condition is positively linked with dying. Dying is more likely for patients in poor condition (6.5%) than for those in good condition (1.5%). See Table 1. Being in poor condition is positively linked with City Hospital. The percentage of patients who are in poor condition is



**Figure 2.** Hospital death rates standardized based on patient condition

greater at City (90%) than at Rural (30%). See Table 2.

To make a fairer comparison of these hospitals, we need to standardize their mix of patients. Let’s standardize both hospitals on their combined mix (60%). Using the group average as the standard emulates the desired outcome in a randomized experiment where the

goal is for each group (exposure and control) to have the same percentage of confounder as found in the overall population.

Standardizing the mix in both groups at 60% increases the expected death rate at Rural Hospital and decreases it at City Hospital, as shown in Figure 2. The standardized death rate is lower for City Hospital than for Rural Hospital (4% compared to 5%). In this case, the direction of the association between the standardized rates is the reverse of that between the crude rates—and we have a fair comparison of the two hospitals; we are comparing “apples and apples.”

## Family Incomes by Race

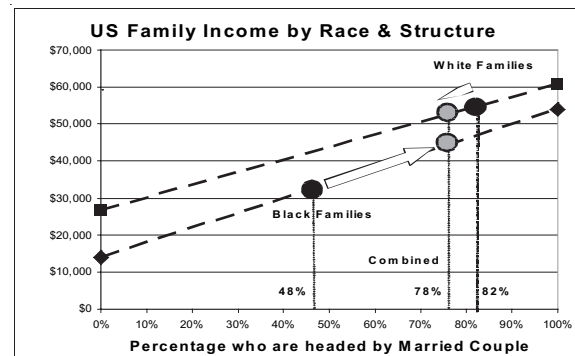
Here is another case. Suppose that in the United States in 1994, mean family income was 66% more for whites than for blacks (\$54,500 compared to \$32,900, as estimated based on the *United States Statistical Abstract*). (See Table 3) Is the black-white income gap fully explained by only race? The \$21,600 white-black income gap could be confounded by a related factor: family structure.

**Table 3.** Estimated Family Incomes by Race and Family Structure

Family Income	Head of Family		
	Unmarried	Married	All
White	\$26,700	\$60,600	\$54,500
Black	\$14,000	\$53,900	\$32,900
All	\$23,000	\$60,100	\$51,900

Family income is higher for married-couple families (\$60,100) than for single-parent families (\$23,000). In order to standardize data, we need the distribution of families by family structure within each race, as shown in Table 4.

Based on Table 4, families headed by a married couple is more likely among whites than among blacks (82% compared to 47.5%). Figure 3 summarizes this data so it can be standardized graphically.



**Figure 3.** Hospital death rates standardized based on patient condition

To take into account the influence of family structure, let’s standardize the mix of family types to a standard mix: the overall percentage of families who

**Table 4.** Number of Families by Race and Family Structure

Families, 1994	Head of Family		All
	Unmarried	Married	
White	10,539	47,905	58,444
Black	4,251	3,842	8,093
All	14,790	51,747	66,537

are married (78%). Standardized family income is 18% more for whites (\$53,000) than for blacks (\$45,000). Standardizing on family structure decreases the black-white income gap by 65%, from \$21,600 to \$8,000. Thus, 65% of the black-white family income gap is explained by family structure.

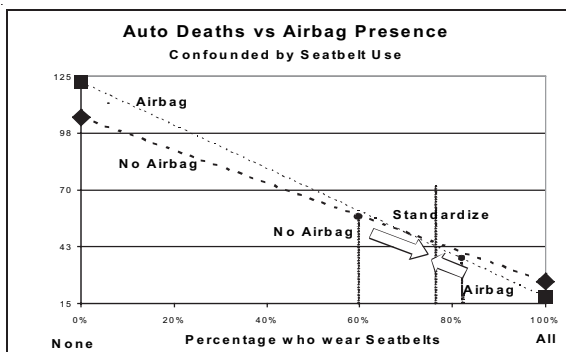
### Auto Death Rates by Airbag Presence

We generally think airbags are good. That conclusion is supported by the data in Table 5, which appeared in Mary C. Meyer and Tremika Finney’s *CHANCE* article, “Who Wants Airbags?”. For the occupants of automobiles in accidents, the death rate is lower for those with an airbag than for those without (37 per 10,000 compared to 60 per 10,000).

**Table 5.** Death Rate per 10,000 Automobile Accident Occupants

Death Rate	Seatbelt Used		All
	No	Yes	
Airbag	No	Yes	All
No	105	26	60
Yes	122	18	37
All	111	21	

But wait! For those not using a seatbelt (left column), the death rate was higher for those with an airbag than for those without (122 per 10,000 compared to 105 per 10,000). The association between airbags and death rate may be confounded by seatbelt usage. Consider the distribution of automobile accident occupants as shown in Table 6.



**Figure 4.** Automobile accident death rates with and without an airbag, standardized based on seatbelt usage

Using a seatbelt is positively associated with a lower death rate, as the death rate was much higher for those who didn’t use a seatbelt at all than for those who used one (111 per 10,000 compared to 21 per 10,000 in Table 5). And using a seatbelt is positively associated with having an airbag, as the percentage using a seatbelt is greater among people in cars with airbags than in cars without (85% compared to 60% in Table 6).

Let’s standardize on the overall percentage of people in accidents who were wearing a seatbelt (73%), as shown in Figure 4.

**Table 6.** Automobile Accident Occupants Using Seatbelts and/or Having Airbags

Number (1,000)	Seatbelt Used		All
	No	Yes	
Airbag	No	Yes	All
No	1,952	2,903	4,855
Yes	871	4,872	5,743
All	2,823	7,775	10,598

The standardized death rate of occupants in auto accidents is slightly higher for those with air-bags than for those without (47 per 10,000 compared to 46 per 10,000). So, do airbags save lives? Not on average for this mix of occupants. This situation is complex because there is an interaction between having airbags and using seatbelts. We can see this because the lines cross. The main point is that seatbelts make a bigger difference in saving lives! Without taking into account the effect of seatbelts, the effect of airbags is almost masked due to the confounding and interaction.

### Analysis of Confounding

Now that we have seen how a lurking factor can confound our understanding of a statistical association, it is good to reflect on what causes these situations and what we can do to avoid them.

Notice what is common to the three examples we have examined. In each case, the researcher was an observer. The researchers did not (and could not) assign patients to a particular hospital, determine which families were headed by a married couple, or determine which car owners bought cars with an airbag. Studies in which the researcher is passive in assigning subjects to exposure and control groups are called observational studies. While the influence of chance decreases as sample size increases, the influence of a confounder remains unchanged in observational studies. The influence of confounding can be a major problem—if not the main problem—in social sciences research, according to Stanley Lieberon in *Making It Count*.

Confounding also can arise in any study—observational or experimental—where the response

due to a factor is observed or measured at a single level and the choice of level influences the association. Because most studies in the news are observational, understanding confounding is absolutely necessary to being statistically literate.

## A Problem from Baseball

To test your understanding of this graphical technique, try working out this problem from baseball.

Ted and Sam are on the same baseball team. Both players have been to bat 100 times. Sam had 26 hits and Ted had 34 hits. So, Sam's batting average is .260 (26%) and Ted's is .340 (34%). But the coach thinks Sam is the better hitter. Could this be due to Simpson's Paradox? Could the strength of the pitcher be a factor? Could the percentage of times each player faced a strong pitcher be a lurking variable? If the pitcher was weak, Sam hit 50% of his times at bat while Ted hit 40% of the time. When facing strong pitchers, Sam hit 20% of the time while Ted hit only 10% of the time. Who is the better hitter? To answer this question, standardize their averages as if each faced strong pitchers 50% of the time. After you have worked this problem, check your answer with the answer on Page 21.

## Conclusion

The influence of context on comparisons of ratios can be profound. Context is an essential difference between statistics and mathematics. To understand the influence of context on a statistic or a statistical association, it helps to understand the confounding effect of lurking variables.

Confounding from lurking variables is the reason that "association is not necessarily causation." With this understanding, we have a stronger reason to be careful in using statistical association as evidence for casual connections. A statistical association is only the first step in establishing causation.

Confounding and standardization are two of the most important ideas in statistics. Once we recognize that standardizing (taking into account confounding) can change the size of a comparison—and may even reverse the direction (Simpson's Paradox)—we have taken a big step toward being statistically literate.

Viewing confounding as the influence of context increases our statistical literacy and provides a link between statistics and other areas of study, including the social sciences and the humanities. For more about this, see Schield's "Statistical Literacy and Liberal Education at Augsburg College," available at [www.StatLit.org/pdf/2004SchieldAACU.pdf](http://www.StatLit.org/pdf/2004SchieldAACU.pdf). ■

**Editor's Note:** *The author would like to thank the W. M. Keck Foundation for their grant "to support the development of statistical literacy as an interdisciplinary curriculum in the liberal arts" and Tom Burnham, Cynthia Schield, and Marc Isaacson for editorial assistance.*

## Additional Reading

Lesser, L. (2001). "Representations of Reversal: Exploring Simpson's Paradox." in Albert A. Cuoco and Frances R. Curcio (Eds.) *The Roles of Representation in School Mathematics (2001 Yearbook)*. National Council of Teachers of Mathematics, 129-145.

Last, J. (1995). *A Dictionary of Epidemiology*. Third Edition, Oxford University Press.

Lieberson, S. (1985). *Making It Count*. University of California Press.

Meyer, M. and Finney, T. (2005). "Who Wants Airbags?" *CHANCE*, 18(1):3-16.

Schild, M. (1999). "Simpson's Paradox and Cornfield's Conditions." *Proceedings of the 1999 Joint Statistical Meetings*, Section on Statistical Education, 106-111.

Schild, M. (2004a). "Three Graphs To Promote Statistical Literacy." Presented at the 2004 International Congress on Mathematical Education, Copenhagen. Available at [www.StatLit.org/pdf/2004SchieldICME.pdf](http://www.StatLit.org/pdf/2004SchieldICME.pdf).

Schild, M. (2004b). "Statistical Literacy and Liberal Education at Augsburg College," Peer Review, 6(3). Available at [www.StatLit.org/pdf/2004SchieldAACU.pdf](http://www.StatLit.org/pdf/2004SchieldAACU.pdf).

Wainer, H. (2002). "The BK-Plot: Making Simpson's Paradox Clear to the Masses." *CHANCE*, 15(3):60-62.



Milo Schield

*Milo Schield (schield@augsb.org) has taught statistics and statistical literacy for more than 20 years at Augsburg College in Minneapolis, MN. With a BS from Iowa State, an MS from the University of Illinois, and a PhD from Rice University, he has pursued a variety of professional interests, including operations research at a large property-casualty insurance company. Currently, he is the director of the*

*W. M. Keck Statistical Literacy Project at Augsburg College. See Schield (2004b) and [www.statlit.org](http://www.statlit.org) for details.*



Peter Flanagan-Hyde

## Observational Studies: the Neglected Stepchild in the Family of Data Gathering

Gathering data is one of the first topics discussed in an introductory statistics course. Along with the effective display and summary of sets of data, the issues surrounding the source of data are an important part of our understanding. In most presentations, there are two valid sources for data, either random sampling or a randomized experiment. Other sources of data, including observational studies, often are used only as examples of a flawed source of data. Due to the possible presence of confounding factors in observational studies, we often are told that the results cannot be counted on, especially in establishing causal relationships.

The characterization of an observational study as a poor stepchild to random samples and experiments in producing data, however, doesn't do justice to the important place it can serve in gathering useful data. To be sure, we should learn the potential pitfalls in using observational studies, but this should fall short of the complete rejection that is most typical. Let's explore why observational studies are a critically important source of data in a number of settings and how the design of the study can maximize useful information while minimizing the chance of confounding spoiling any conclusions.

### Why an Observational Study?

First, let's understand why anyone would choose a potentially problematic method of data collection (an observational study) instead of one that features random selection or random assignment. Most of us have some exposure to the notion that there are times when a

randomized trial is either impossible to do or unethical to do. A randomized trial might be impossible if, for example, researchers are studying whether genetic factors are related to a response. The most obvious example of these genetic factors is gender—you can't randomly assign male or female to the participants in your study—and other genetic factors that share this trait. In addition, it might be unethical to do a randomized trial if the study is about the effects of known risk factors that can't be imposed on the subjects without potentially harming them.

In addition to these reasons, in an article in the *New England Journal of Medicine*, Kjell Benson and Arthur Hartz note that "observational studies have several advantages over randomized, controlled trials, including lower cost, greater timeliness, and a broader range of patients." The issue of cost is always a factor in the real world, and generating new data can be quite costly. Timeliness is a potential problem for randomized trials if the response takes a long time to develop, such as cancers or other medical issues. The scope of inference of a study may be limited to the population from which the participants are selected, and often is a result of relying on volunteers.

### Different Types of Observational Studies

Thus we see that using observational studies to gather data offers advantages as well as risks, and there is ongoing research about how to do observational studies so as to minimize the risks involved while exploiting the advantages. Different approaches to observational studies balance these competing forces in different ways, so let's look at the major types of observational studies, when they are most helpful, and what cautions should be employed with their use.

Because much of the current research about observational studies takes place in the fields of medical practice, the example we'll use to illustrate the different approaches is a medical one that is common but not

---

*Peter Flanagan-Hyde (peterfb@mac.com) has been a math teacher for 27 years, the most recent 15 in Phoenix, Arizona. With a BA from Williams College and an MA from Teachers College, Columbia University, he has pursued a variety of professional interests, including geometry, calculus, physics, and the use of technology in education. Flanagan-Hyde has taught AP Statistics since its inception in the 1996–1997 school year.*

typically life-threatening: appendectomies. Let's start with the weakest types of observational studies and work up to the putative gold standard—the randomized, controlled trial.

### DESCRIPTIVE STUDIES

In a descriptive study, the only goal is to either estimate the incidence of a condition or determine if there is any evidence of a difference in treatment outcomes. In our example, you might search through hospital records in a given city to make a list of patients who had laparoscopic surgery for abdominal pain and those who had the more traditional open surgery. If the outcomes in one group are more favorable, it may be that the treatment is generally better. However, in a descriptive study, you can never make a valid causal association, as it may well be that patients who had more severe symptoms typically were given one of the treatments. Descriptive studies do have a role, however, in formulating hypotheses that can be checked by more rigorous designs.

### CROSS-SECTIONAL STUDIES

A cross-sectional study is a snapshot of a population at one moment in time. Various conditions are measured on the population or a sample from the population, and associations between potential risk factors and the incidence of illnesses are examined. Because a cross-sectional study measures all of the variables (risk factors and illness) at the same moment, it is impossible to determine the order of any of the conditions, thereby rendering causal relationships impossible to assess. In our example, another drawback of a cross-sectional study is evident: at any given moment, the number of patients with appendicitis is likely to be small, and this issue is of even greater concern for diseases that are rare in a given population.

### CASE-CONTROL STUDIES

A case-control study is a retrospective study in that it looks back in time through existing records. Unlike a descriptive study, however, there is a deliberate matching of subjects who have a given disease with other members of the population who don't. Each of the disease-free members in the study are selected to be as similar as possible to one of the members with the disease in terms of demographic and other variables that might affect the condition. An examination of the records is conducted to determine any systematic differences in the two groups. This has the potential to quickly identify causal factors, and it was through case-control studies that several notorious causal connections were made, including the association with the drug diethylstilbestrol (DES), taken in pregnancy, and cancers that developed in the offspring, and the initial association between smoking and lung cancer.

In our example, a case-control study might look at a number of individuals who have an adverse outcome

following an appendectomy, such as infection. Matching these individuals with others who did not have the adverse outcome and examining the type of surgery they had allows us to estimate the relative risk of infection for the two surgical procedures.

### COHORT STUDIES

A cohort study is the most highly respected observational study. A large, disease-free population (the cohort) is selected and then consistently monitored for a long period of time, with many characteristics of each member of the population recorded. If there is a particular risk factor of concern, a second matched cohort can be selected that does not have the risk factor but is similar in other respects. Over time, some of the members will develop a variety of diseases, and the characteristics that distinguish these members from those who remain disease-free can be determined.

Cohort studies are prospective, and because there is a temporal development of the data, it can be seen clearly whether a given factor precedes the development of the disease, unlike a cross-sectional study. Large cohort studies were instrumental in strengthening the link between smoking and lung cancer. In our example of appendectomies, it is possible that there may be longer-term differences in the patients who have the two treatments. A cohort study would be the only way to discover this.

Cohort studies, though, must include many subjects to be effective, and are, therefore, expensive to conduct. They also are not very effective in studying rare diseases, unless extremely large, as the cohort may or may not have any members who develop the disease. On the other hand, they can be effective at studying groups whose risk factors are rare. For example, are ultramarathon runners (those who run races of 30 miles or more) are more prone to degenerative joint diseases as they age? Selecting a cohort of ultramarathoners (a rare breed in the general population) and a cohort of other avid exercisers could assess this.

In addition to the large sizes and associated costs, the long time frame of cohort studies means there may be difficulties in following the members into the future. If members drop out or are otherwise lost, there may be a nonresponse bias introduced into the study.

In most cases, the design chosen for an observational study is determined by the characteristics of both the population and the issue being investigated, as well as available financing. Descriptive and case-control studies, since they are retrospective, often can be conducted for little money and produce nearly immediate results. Large cohort studies, on the other hand, are expensive and take place over long time frames. In what is perhaps the most famous large cohort study, the Framingham Heart Study, more than 10,000 participants have been monitored since 1948. The study is now working with a third-generation cohort—grandchildren of the original participants. This

study has been instrumental in developing the current understanding of heart disease and has led to more than 1,200 publications about risk factors for coronary disease. The cost of this study over the years, however, is measured in the tens of millions of dollars.

### QUALITY OF INFORMATION FROM OBSERVATIONAL STUDIES

The money spent on observational studies can be well worth it if the studies produce sound results. Otherwise, if confounding is an overwhelming problem, the studies might be leading us in the wrong direction. Several recent efforts have been made to evaluate the quality of results from observational studies, and, in general, the results are promising. It seems that with well-designed observational studies, the risks of confounding can be limited. Concerns have been expressed that observational studies tend to exaggerate treatment effects. A *New England Journal of Medicine* article by J. Concato, N. Shah, and R. Horowitz, "Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs," indicates that "well-designed observational studies (with a cohort or case-control design) did not systematically overestimate the magnitude of the associations between exposure and outcome as compared with the results of randomized, controlled trials." In fact, it seems there can be more variability in the outcomes of the randomized, controlled trials than in the observational studies.

For our example, appendectomies, there have been a number of studies that have compared the results of laparoscopy and open surgery. The consensus wisdom is that laparoscopic surgery, the less invasive alternative, is the preferred method for straightforward cases. This consensus has been established by a combination of both observational studies and randomized, controlled trials. A comparison by K. Benson and A. Hartz of eight observational studies and 16 randomized, controlled trials revealed that seven of the eight observational studies found an advantage for laparoscopy (and the eighth no difference), while, among the 16 randomized trials, eight favored laparoscopy, three favored open surgery, and the remaining five had very close results.

### Conclusion

To use a randomized, controlled trial, the research question must be relatively mature—with some confidence that a treatment is meaningful—before the cost of the trial can be justified. An observational study can set the stage. In many situations, an observational study is a first step toward understanding by roughly identifying associations that can be more closely examined with either a more rigorous observational study or a randomized, controlled trial. The issue of confounding with observational studies is real, but we can benefit from an observational study by having a more developed view that places the value of a randomized, controlled trial in a more meaningful context. ■

### Additional Reading

Benson, K. and Hartz, A. (2000). "A Comparison of Observational Studies and Randomized, Controlled Trials." *New England Journal of Medicine*, (342): 1878-1886.

Concato, J. Shah, N. and Horowitz, R. (2000). "Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs." *New England Journal of Medicine*, (342), 1887-1892.

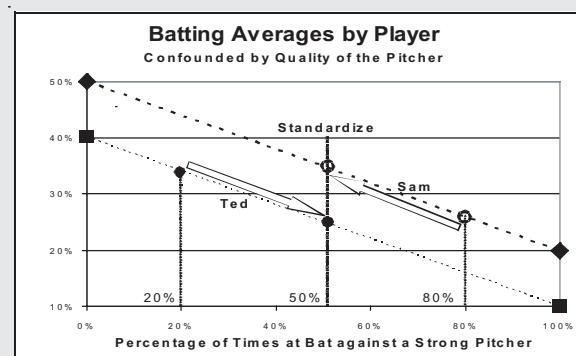
National Heart, Lung, and Blood Institute. (2002). Framingham Heart Study. Available at [www.nhlbi.nih.gov/about/framingham/index.html](http://www.nhlbi.nih.gov/about/framingham/index.html).

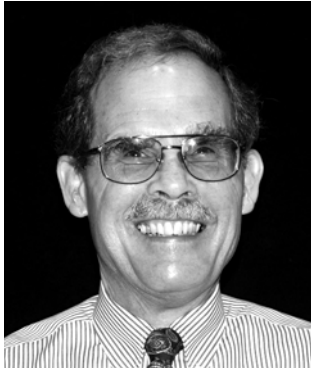
Pai, M. (Accessed 2006). "Observational Study Designs," available at <http://sunmed.org/Obser.html>.

## Answer to Baseball Problem

from Page 18

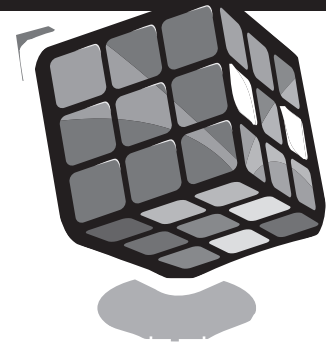
The players' standardized batting averages are .350 (35%) for Sam and .250 (25%) for Ted. After taking into account (controlling for or conditioning on) the strength the pitcher, Sam's batting average is higher than Ted's. So, the Coach is right – Sam is the better hitter.



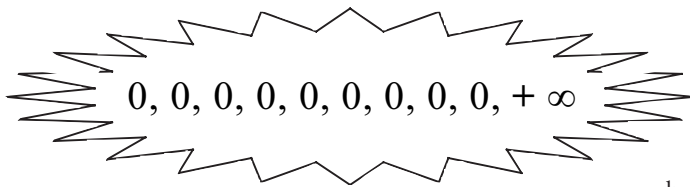


Schuyler W. Huck

# STATS PUZZLER



## Infinite Wisdom



Suppose 10 people are asked to rate themselves in terms of their accumulated wisdom. The first nine people, being quite humble, give themselves a zero. The tenth person is anything but humble and boldly asserts, "I have infinite wisdom!"

Suppose further that the 10 people who supply these wisdom scores are considered to be our population of interest, not a sample. Finally, suppose you're asked to compute the z-score for the tenth person's wisdom score.

A z-score, of course, is computed via the formula,

$$z = \frac{X - \mu}{\sigma}$$

with the result indicating how many standard deviations ( ) a particular score (X) lies above or below the population mean ( ). For example, if a student earns a score of 80 on a test given to a group of examinees in a population with a mean and standard deviation equal to 60 and 10, respectively, the z-score for the student who earns 80 is equal to +2.00.

Okay, it's time for you to put your brain to work on the 10 wisdom scores shown above. As our group of 10 is the entire population of interest, what do you get if you convert the tenth score of infinity into a z-score? Or, stated differently, how many standard deviations is our not-so-humble tenth person above the mean of our little group? After you have worked this through, see Page 29 for the puzzle's solution.

---

*Schuyler W. Huck (shuck@utk.edu) 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.*



# Random Numbers from Nonrandom Arithmetic



Bruce Trumbo

Very simple simulations can be done by mechanical methods such as tossing coins or rolling dice, for example. With real dedication and a lot of spare time, we could roll a pair of fair dice 600 times and notice that we would get a total of 7 ‘about’ 100 times. Suppose we let the random variable  $T$  be the total on any one roll. Then, this experiment could provide a crude demonstration that  $P\{T = 7\} = 1/6 = 0.167$ . This exact result is easy to find: When two dice are rolled, there are 36 possible outcomes, and six of them result in the event  $\{T = 7\}$ .

We say a simulation with 600 tosses is crude because, according to the binomial distribution, we would have only about three chances in four of seeing the event  $\{T = 7\}$  between 90 and 110 times, so there is about one chance in four our estimate of  $P\{T = 7\}$  falls outside the interval  $[90/600, 110/600]$ , or  $[0.150, 0.183]$ . In practice, it usually is not feasible to do useful simulations with such mechanical methods as rolling dice.

Credit for being the first to try computer simulation often is accorded to Stanislaw Ulam and John von Neumann, famous mathematicians and physicists working together on the Manhattan Project in World War II. In 1946, Ulam was pondering the chances of winning a game of solitaire, which begins when the 52 well-shuffled cards of a standard deck are laid out in a particular array. The game is sufficiently complicated that a solution using permutations and combinations seems daunting. So, Ulam wondered about approximating the probability of a win as his fraction of wins in 100 games. The difficulties are that it would be a lot of work to do 100 games and, even then, 100 games would not be enough to get a very good estimate.

At that time, computers were just becoming fast enough that it was reasonable to imagine somehow using a computer to do simulation. Later that year, Ulam and von Neumann began to plan computer simulations on probabilities of neutron diffusion as a substitute for solving very difficult differential equations. (See “Interactive Learning with a Digital Library in Computer Science, *The History of Computing*, John von Neumann” available at <http://ei.cs.vt.edu/~history/VonNeumann.html>.)

---

*Bruce Trumbo (bruce.trumbo@csueastbay.edu) is a professor of statistics and mathematics at California State University, East Bay (formerly CSU Hayward). He is a Fellow of the ASA and a holder of the ASA Founder’s Award.*

## Computerized ‘Sin’

Doing modern simulation depends on having available a supply of computer-generated “random numbers.” You may wonder how it is possible for a computer, programmed to follow fixed arithmetical rules, to produce random numbers. You would be right to wonder because, strictly speaking, it cannot be done. In fact, von Neumann became so discouraged with early attempts, he said at a conference on simulation in 1951, “Anyone who considers arithmetical methods of producing random numbers is, of course, in a state of sin.”

Some early simulations were done by capturing numerical results of natural phenomena thought to be inherently random, such as noise in electronic circuits, radioactive decay, and so on. As early as 1955, statisticians began to use published tables and decks of computer cards containing such random numbers to make random assignments of subjects to treatment groups in designed experiments and for small-scale simulations.

However, a lot of ‘sinning’ has gone on since 1951, and it seems von Neumann was too quick to give up on arithmetical methods. While it is not possible to generate truly random numbers with a deterministic formula on a computer, statisticians, mathematicians, and computer scientists have learned that—by being clever and careful—it is possible to generate “pseudorandom” numbers. These are sequences of numbers that (we hope) cannot be meaningfully distinguished from random ones. Our purpose here is to explore briefly how we can be clever enough and careful enough to generate the pseudorandom numbers that make modern simulation possible.

## Congruential Generators

Consider the problem of shuffling the 52 cards in a deck. For our purposes, it is convenient to number them 1, 2, ..., 52. One way to shuffle these 52 cards is repeated use of the equation

$$r_{i+1} = (27r_i) \text{ mod } 53.$$

Start with any one of these numbers for  $r_1$ , say  $r_1 = 9$ . This arbitrary starting number is called the seed.

Here is how the equation works:  $27r_1 = 27(9) = 243$ . But 243 does not correspond to any of the cards, so divide 243 by 53 and you get 4 with a remainder of 31. Using the remainder as the next number, we set  $r_2 = 31$ . That's what "mod 53" means: Use the remainder upon division by 53 when the number of interest exceeds the modulus 53. In the language of the mathematical field called "number theory," we say that 243 is congruent to 31 modulo 53.

Now repeat the process to get  $r_3$ . The remainder is 42 when  $27r_2 = 27(31) = 837$  is divided by 53, so  $r_3 = 42$ . Continuing in the same way, you can verify that the next three values in the sequence are  $r_4 = 21$ ,  $r_5 = 37$ , and  $r_6 = 45$ . This process produces all of the 52 numbers before getting back to 9 with  $r_{53} = 9$ . And then the sequence repeats itself. The output in Figure 1 shows the results of 60 iterations.

Of course, if we start with the seed  $r_1 = 21$ , then we get  $r_2 = 37$ ,  $r_3 = 45$ , and so on. So this formula shuffles the cards in a fixed rotation, where the seed  $r_1$  determines the starting card. More generally, a linear congruential generator is based on the equation

$$r_{i+1} = (ar_i + b) \bmod d.$$

Above, our choice of the constants  $a = 27$ ,  $b = 0$ , and  $d = 53$  was dictated by the need to shuffle the numbers 1, 2, ..., 52, corresponding to the 52 cards in a deck.

```
a <- 27; b <- 0; d <- 53
m <- 60; r <- numeric(m); r[1] <- 9
for (i in 1:(m-1))
{
r[i+1] <- (a*r[i] + b) %% d
}
r; length(unique(r));
plot(r[1:59], r[2:60])
> r
[1] 9 31 42 21 37 45 49 51 52 26 13 33 43
[14] 48 24 12 6 3 28 14 7 30 15 34 17 35
[27] 44 22 11 32 16 8 4 2 1 27 40 20 10
[40] 5 29 41 47 50 25 39 46 23 38 19 36 18
[53] 9 31 42 21 37 45 49 51
```

Figure 1. R code for implementing a linear congruential generator

The behavior of a generator depends on intricate rules of number theory (and, unfortunately, on other principles that are not completely understood). For example, if we use  $a = 8$ ,  $b = 0$ , and  $d = 53$ , then we get a fundamentally different order for the 52 cards.

```
[1] 9 19 46 50 29 20 1 8 11 35 15 14 6
[14] 48 13 51 37 31 36 23 25 41 10 27 4 32
[27] 44 34 7 3 24 33 52 45 42 18 38 39 47
[40] 5 40 2 16 22 17 30 28 12 43 26 49 21
```

However, if we used  $a = 7$ ,  $b = 0$ , and  $d = 53$ , then only half of the 52 numbers are generated before the values begin to be repeated.

```
[1] 9 10 17 13 38 1 7 49 25 16 6 42 29
[14] 44 43 36 40 15 52 46 4 28 37 47 11 24
[27] 9 10 17 13 38 1 7 49 25 16 6 42 29
[40] 44 43 36 40 15 52 46 4 28 37 47 11 24
```

If a generator has  $b = 0$ , it is called multiplicative because there is no additive term or increment in the equation. When  $b = 0$ , it is not possible to have  $r_i = 0$ , but if  $b > 0$ , then  $r_i = 0$  is possible.

## Period of a Generator

The number of distinct values a generator produces before it starts to repeat is called its period. If  $a = 27$  or 3, we get the largest possible period 52, but if  $a = 7$ , the period is only 26. Clearly, the period cannot be larger than  $d$  (or if  $b = 0$ , not larger than  $d - 1$ ). Several values of  $a$  give the full period 52 and fundamentally different orders of the cards. But, there clearly are not anywhere near enough values of  $a$  to generate a representative sampling of the  $52 \approx 8 \times 10^{67}$  possible orders of shuffling 52 cards.

We started by showing some simple generators with  $d = 53$  because they show the idea of shuffling cards and involve numbers of manageable size. However, serious simulation—even for card games—requires a generator with a huge period and, hence, a huge value of  $d$ .

## A Bumpy Histogram and Grid Patterns

Before we show generators with larger periods, we will use the simple generators with  $d = 53$  to illustrate a few more important ideas used to check the usefulness of congruential generators. In practice, we would like random numbers to be independent and uniformly distributed throughout the interval (0, 1). So, it is common to use values  $u_i = r_i / d$  because they lie between 0 and 1.

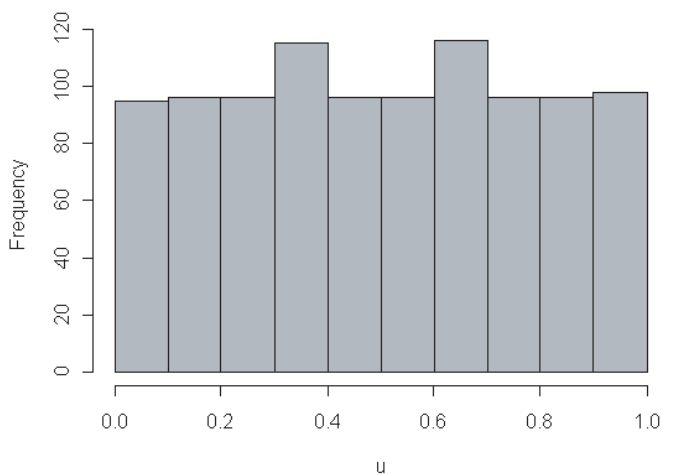


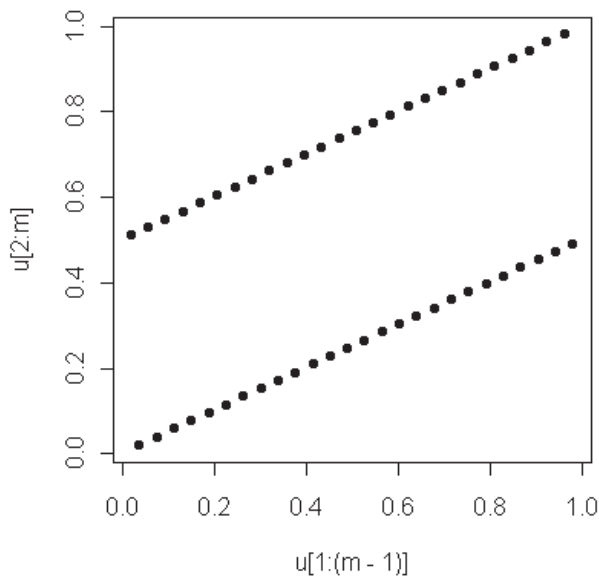
Figure 2. Histogram of 1,000 numbers  $u_i = r_i/d$  from the congruential generator in Figure 1. The histogram is 'bumpy' because there are only 52 values among the 1,000 numbers generated.

As a first check whether the,  $u_i$  are uniformly distributed, we could make a histogram, which should look reasonably flat. Figure 2 shows a histogram of  $m = 1000$  values from the generator of Figure 1. Of course, there are really only 52 different numbers  $u_i = r_i / d$  repeated over and over again in rotation. The histogram is bumpy, rather than flat, mainly because it happens that eight of the bins each include five of the 52 values (accounting for 40 values so far) and the remaining two bins include six numbers each.

Many congruential generators with large periods yield histograms that are as smooth as would be expected from a true random sample from a uniform distribution. But some generators that pass the histogram test behave very badly in other, more subtle ways. For example, the  $u_i$  values may be associated in ways that make them useless for simulation.

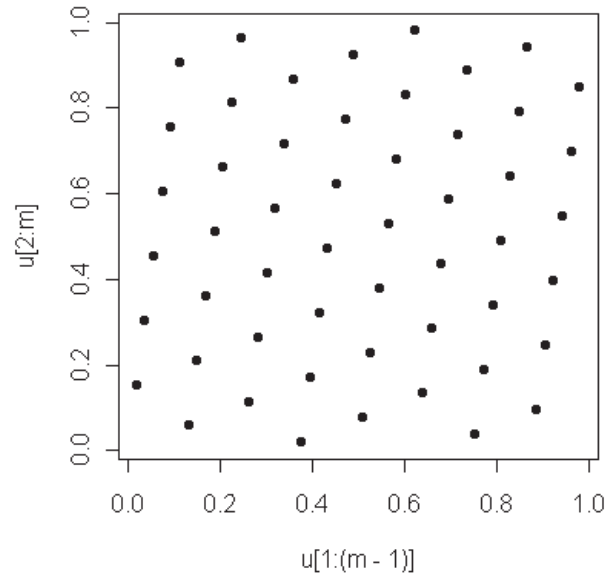
If we plot the pairs  $(u_i, u_{i+1})$  in the unit square, we can check for pairwise independence. Ideally, these points would fall randomly throughout the unit square without showing any apparent pattern. Unfortunately, in such plots, all congruential generators yield points that lie on a grid.

Figures 3 and 4 show the remarkably different results for generators with  $a = 27$  and  $a = 8$ , respectively. Comparing Figures 3 and 4, we clearly see  $a = 8$  makes a more satisfactory grid. In Figure 4, coverage is about as smooth as possible with a period as small as 52. Unfortunately, it is possible to have a generator with a huge period but with a very coarse grid pattern, similar to the one in Figure 3. A useful generator will have a large period and a grid pattern so fine we will not notice it in practice.

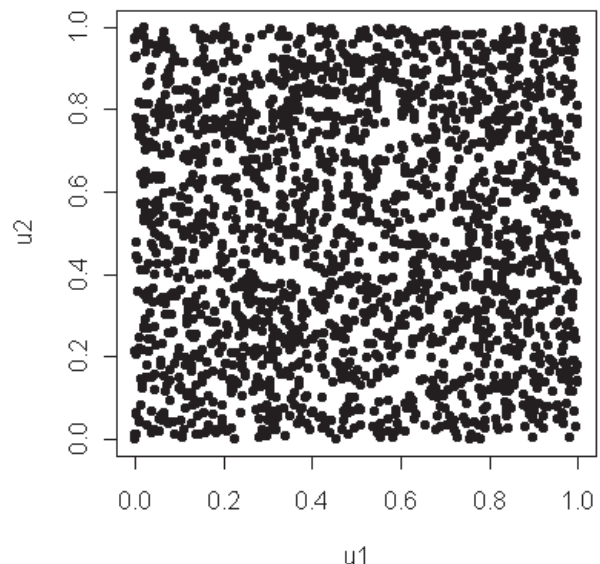


**Figure 3.** Plotted in the unit square, successive points of the linear congruential generator with  $a = 27$ ,  $b = 0$ , and  $d = 53$  lie on a 'coarse'  $2 \times 26$  grid.

A useful generator also will have nonintrusive grid patterns in dimensions higher than two. The generator with  $a = 65539$ ,  $b = 0$ , and  $d = 2^{31}$  shows a pattern that is not distinguishable from random when 2,000 of its points are plotted in a square in Figure 5, using the same method as for Figures 3 and 4. But a disastrously coarse grid appears when it is plotted in 3-D. In the unit cube, all of its very large number of points lie in only a few widely separated parallel planes.



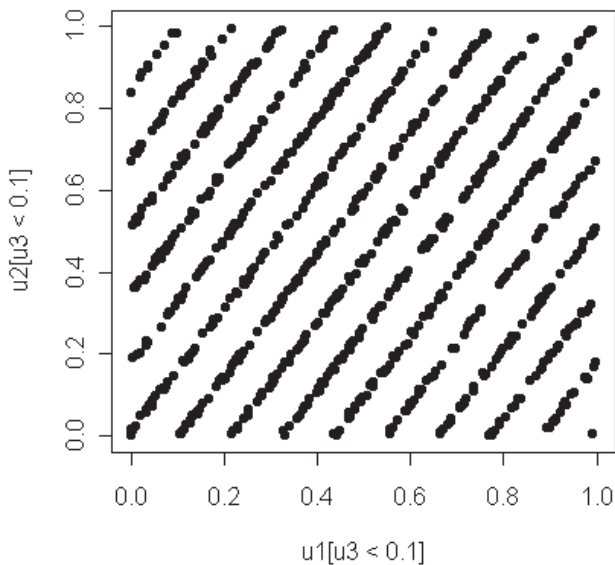
**Figure 4.** Plotted in the unit square, pairs of adjacent values of the generator with  $a = 8$ ,  $b = 0$ , and  $d = 53$  lie on a relatively fine  $8 \times 9$  grid. Compare with the coarse grid of Figure 3.



**Figure 5.** Pairs of successive  $u_i$  values from the RANDU generator plotted in the unit square. In 2-D, the grid is so fine that 2,000 points show no pattern that would contradict randomness.

Figure 6 shows the approximately 1,000 points that lie in the front one-tenth of the cube (out of 10,000 points in the cube altogether.)

This generator, known as RANDU (for the function call to use it on IBM mainframe machines), was perhaps the most popular in the world for many years. But then people started to notice it gave incorrect simulation results for some problems with known answers. Then, its bad 3-D behavior was discovered. Lesson learned! Now it is considered mandatory to validate generators in high dimensions before using them for serious simulation work.



**Figure 6.** When 10,000 triples of adjacent  $u_i$  values from RANDU are plotted in the unit cube, we see they all lie in a few widely separated planes. The figure shows about 1,000 points that lie within 0.1 of the front surface of the cube.

The default pseudorandom number generator in R, accessible with the function `runif`, is more complicated than the congruential generators discussed here. It has a period of  $2^{19,937} - 1 \approx 4.32 \times 10^{6,001}$ . (Within the precision of R, there are about 4.31 billion distinct values.) Its geometry has been ‘twisted’ so that even its very fine grid is nonlinear, and it has passed tests for independence in 623-dimensional space.

## Validating a Generator

There are plenty of rules, based on experience, for choices of numbers  $a$ ,  $b$ , and  $d$  that should be avoided, but no rules that ensure success. We have discussed only linear congruential generators because they are easy to explain and have been widely used. There are other kinds of generators that have been successful, and there are useful guidelines for properties to avoid in constructing them. We said earlier that cleverness and care are necessary to do successful simulation based on computer algorithms. The clever part is to understand what must be avoided. The careful part is to test a

candidate generator thoroughly. We have mentioned some of the most basic tests:

- Checking to make sure the period is very large. Some modern simulations use millions of pseudorandom numbers, and we do not want the numbers to recycle during the simulation.
- Histograms, which can be backed by goodness-of-fit tests, to test whether results fit a uniform distribution.
- Multidimensional plots to look for association among the generated values. Human “pattern recognition” is useful here. Anything that looks like a pattern means trouble. There are also methods to test numerically for correlation and nonlinear association.

All of these are important methods of screening out bad generators. (See *Random Number Generation and Monte Carlo Methods* for authoritative information about congruential and other generators.) However, perhaps the most important kind of test for a generator is to use it to solve difficult simulation problems with known answers. Standard batteries of such problems have been developed for testing generators.

## About Seeds

If you start a congruential generator with a particular seed, you will always get the same sequence of pseudorandom numbers. There are two ways to handle seeds in practice.

One method is for an unpredictable seed to be supplied by the program that uses the generator. As a simulation starts, R gets this unpredictable number from detailed information on the computer system’s clock. For a generator with a huge period, it would not be feasible to figure out where in the sequence of pseudorandom numbers the simulation started. So, the effect is as if the simulation is based on truly random numbers.

The other method is for us to supply the seed. In R, we could begin our simulation with the statement `set.seed(1212)` to start with seed 1212. This can be useful for debugging simulation programs. Every time we start our simulation with the same seed, we will get exactly the same answer. We use seeds in the next section.

## A Simple Comparison

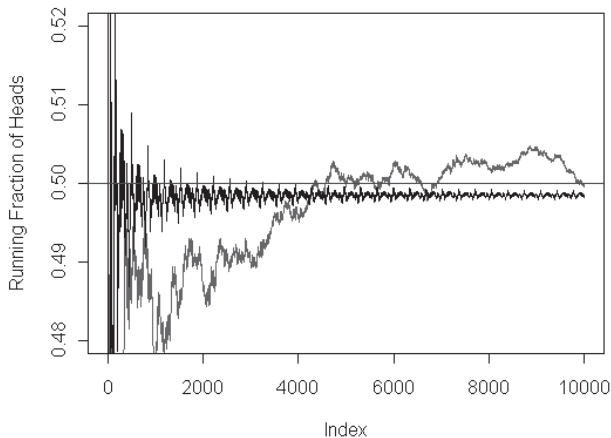
Although the linear congruential generator with  $a = 1,093$ ,  $b = 18,257$ , and  $d = 86,536$  is not good enough for advanced simulations, it is pretty good. It has period  $d$  and has passed some simulation tests. Now suppose someone proposes changing its increment to  $b = 252$ . In the challenges below, we ask you to show various reasons, based on the principles discussed above, for why this is a bad idea. (One reason is that 252 has factors in common with 86,536.) For now, let’s try a simple simulation with the modified generator to see how it performs. Suppose a fair coin is tossed repeatedly and that  $F_n$  is the fraction of heads in the first  $n$  tosses. If we plot  $F_n$  against  $n$ , the resulting “trace” should take a random path that approaches  $1/2$ . (This

is similar to plotting traces of the public opinion polls that we did in the Fall 2005 issue.)

Figure 7 shows traces from two simulations. The ‘saw tooth’ path that converges rapidly to an incorrect value slightly below 1/2 comes from using the bad modified generator with seed 12. The meandering path that approaches 1/2, as we would expect of a random sequence of heads and tails, comes from the R code below. Both simulations are based on 10,000 simulated tosses. It is clear from Figure 7 that the bad generator is not giving values consistent with random numbers. This one picture is enough to discredit the generator.

```
set.seed(1212)
m <- 10000
x <- runif(m); f <- cumsum(x > .5)/(1:m)
plot(f, type="l", ylim=c(.48,.52))
abline(h=.5)
```

In both cases, the simulation assumes the generator produces variables  $X_i$  distributed uniformly on the interval (0, 1), so the events  $\{X_i > .5\}$  can be taken as equivalent to getting a head on the  $i^{\text{th}}$  toss of the coin. The statement  $X_i > .5$  is either TRUE (value 1) or FALSE (value 0). If you use the same seeds and software we did, you should get exactly the same result. (See challenge #3 at the end for hints on plotting the bad trace.)



**Figure 7.** Traces of the proportion of heads after each of 10,000 simulated tosses of a fair coin using a bad generator (regular saw tooth path) that converges too quickly and to the wrong result and a good generator (meandering path) consistent with random coin tosses.

## Nonuniform Random Variables

R uses uniform random variables from its random number generator to obtain random variables with a variety of distributions other than uniform. Here are a few ways this can be done:

- `sum(runif(10) < 1/3)` has a binomial distribution with 10 trials and  $P(\text{Success}) = 1/3$  on each trial. The direct method in R is `rbinom(1, 10, 1/3)`.

- `-log(runif(100))` gives a vector of 100 independent exponential random variables, each with rate 1. This method is exact, but faster methods that do not involve taking logarithms have been proposed. Direct method: `rexp(100, 1)`.
- `sum(runif(12) - 6)` is very nearly a standard normal random variable. Obviously, the result obtained here, based on trusting the Central Limit Theorem for the sum of only 12 observations, cannot fall outside of the interval  $(-6, 6)$ , but an actual standard normal random variable has an extremely small probability of doing so. A better, but more complicated, method of generating standard normal random variables uses a trigonometric transformation. The method shown was used in an earlier era when computer manipulation of trigonometric functions was very slow. Direct (and better) method: `rnorm(1)`.
- `unique(ceiling(52*runif(1000)))[1:5]` gives five distinct randomly chosen numbers between 1 and 52 that can be interpreted as a fairly dealt poker hand. This is a wasteful method and it has a miniscule chance of not working. Direct (and much more efficient and absolutely sure to work) method: `sample(1:52, 5)`.

## Challenges

1. What is the period of the bad generator used in Figure 5? Hint: The period cannot exceed  $d$ . Run a program similar to that of Figure 1 for  $d$  iterations and then find the number of distinct values with `length(unique(r))`. Similarly, show that the period of the pretty good generator we ruined to get the bad generator has period  $d = 86,536$ .
2. Plot 20,000 adjacent pairs of values  $(u_i, u_{i+1})$  in the unit square from each of the generators in challenge #1 and `runif`. For uncluttered views, use the argument `pch="."` in the plot statement. How can you tell `runif` is the best?
3. Make Figure 7 on your own. Use the code shown to make the trace for `runif`. Then, use the bad generator to make a new vector `x` of  $m$  faulty pseudorandom numbers and, from it, a new vector `f`. Then, overlay the bad trace with `lines(f)`. Finally, change the seeds for both generators (maybe use your birth month for both). What changes substantially and what remains essentially the same?
4. (Intermediate) The probability of getting no aces in a fairly dealt poker hand can be obtained by combinatorial methods. Explain why this can be evaluated in R as `choose(48, 5)/choose(52, 5)`. Then, write a program, based on `sample`, to simulate this value. Go through a loop to simulate 50,000 poker hands. Hint: Explain why the number of aces in one poker hand can be simulated as `sum(sample(1:52, 5) < 5)`.
5. (Advanced) Explain the rationale behind the methods suggested in each of the bullets in the last section of this article. ■



## I Salute You

So here's the deal, I collect sayings. Aesop, Confucius, Homer Simpson, all the classic philosophers. I have found on many occasions that a well-placed aphorism can disguise even the most serious dearth of fact and/or logic, and I have no dearth of such dearths. One classic saying is "April showers bring May flowers." I was reminded of this saying when I spent a couple weekends last April—one on the West Coast and one on the East Coast—basically in the rain. Contrary to the view of some of my students, I do know enough to come in out of the rain, and, as it happens, I came in out of the rain to observe some budding May flowers. (Work with me here. I'm attempting to do one of those metaphor things.)

My first coming in out of the rain was in San Francisco in early April. (I guess the old song is correct: It never rains in *Southern* California.) I attended the 2006 annual meeting of the National Council on Measurement in Education. NCME is comprised of those folks who (a) write and analyze national tests such as the ACT, SAT, and AP tests; (b) worry about using tests equitably to make decisions about college admissions, high school diplomas, and No Child Left Behind; and/or (c) study things like posterior predictive model checking for within-item multidimensionality in item response theory. Uh-huh, right.

As I blissfully sat through the three days of the conference, I heard some nice presentations of papers. Almost all of them were based on statistical reasoning and, therefore, pretty interesting. Most of the sessions were presented by folks in the educational measurement community, expert in the ways of convoluted theoretical statistics. Toward the end of the conference, I happened

---

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



Chris Olsen

to hear some presentations by grad students. (Okay, metaphor time. Think May flowers now.) What struck me was the quality of the presentations. True, this was probably not a random sample of students, but there are some statistics professors who should take some bows. All those lectures raining down on their students (April showers? Still with me on this?) were definitely not in vain. One suspects these youngsters learned their PowerPoint from other sources, but what good is PowerPoint without the solid content backed by the explanatory power of statistics? Many of these young men and women were no doubt presenting parts of their dissertations—their advisor's names were listed in the program as co-presenters, although said advisors seemed to not do any co-presenting. One hopes these advisors were present to take well-deserved pride in their student's work—the pride only teachers know.

After taking the red-eye back home and catching up on my day job, I headed for Washington, DC. I had once again missed the cherry blossoms, and—guess what?—it was raining there also. This time, when I came in out of the rain, I heard scholarly papers of a different sort, the general topic being epidemiology. Epidemiology, you may recall, is the study of risks to health at the population level. What factors and behaviors are risky? How are such risks identified, controlled, and perhaps even eliminated? Infectious diseases are the most commonly thought of examples for what epidemiologists deal with, but the topics of epidemiology range from childhood obesity to bicycle helmet use.

Once again, I heard presentations by both experts in the field and by the next generation of potential epidemiologists. And, once again, the younger folks stole the show with their presentations. Here, as in San Francisco, the statistical power of their methodology loomed large in their papers. One young lady invented something called a "bi-orbital rotational swing" as a therapy for treating attention deficit hyperactivity in children. A young man surveyed three high schools to find that vigorous physical activity might offset negative effects of minor mood disorders. And another young lady presented her study of high school students' alcohol drinking behavior and compared the behavior to their parents' perception of what was going on. (Surprise! The parents were clueless.)

These papers, being more applied, did not exhibit the same theoretical statistical depth as the NCME papers. The analyses presented in Washington were more along the lines of one-way and two-way analysis of variance, analysis of covariance, a little multiple regression now and then, and lots of concern about confounding variables. Really, the kinds of stuff one learns in the standard undergraduate statistics program. Oh, I'm sorry. I think I may have forgotten to mention that these young men and women presenting in Washington were all *high school* students.

They were there as regional finalists in a scholarship competition, part of the Young Epidemiology Scholars program, an effort funded by the Robert Wood Johnson Foundation. It was my very great pleasure to have a front-row seat for these presentations; I was one of the judges. In that role, I was able to probe slightly their understanding of the statistical basis for their studies. I am here to tell you they did their statistics teachers proud. The level of their expertise doesn't derive from mindless software twiddling; it is understanding that comes from hard study and reflection with dedicated teachers—in this case, high school math and statistics teachers. These students also are not a random sample of today's high

“And, once again, the younger folks stole the show with their presentations.”

school students, and their teachers may not be able to take credit for their polished PowerPoint presentations either, but, once again, I can report that some serious teacher pride was in order.

So, let's get to the bottom line here. To teachers of statistics everywhere, both high school and college, I offer a well-deserved salute. Though as teachers of statistics, you probably shy away from basking in the limelight with your students, you certainly deserve extensive vicarious credit—your April efforts do bring forth some fine May flowers. And I salute you students of statistics as well. Even though you might sometimes feel you are being 'drenched' in your studies, your hard work will pay off next May and in years to come. ■

## Solution to STATS Puzzler from Page 22

There are two ways to solve this puzzle. The first approach involves discovery and then a logical leap. The second approach involves a little algebra.

To solve this problem via discovery followed by a logical leap, you must first change the tenth wisdom score from infinity to something a bit more manageable. Let's change it to 10. If you now compute the z-score for this tenth score, you get +3.00, because 10 is three standard deviations ( $\sigma = 3$ ) above the group mean ( $\mu = 1$ ). Next, you must change the tenth score to something else, and then compute the new z-score. This time, let's set the tenth score equal to 100. When we now compute the z-score, it again turns out equal to +3.00 because 100 is three standard deviations ( $\sigma = 30$ ) above the mean ( $\mu = 10$ ). Finally, let's change the tenth score one last time, now setting it equal to 1. When we compute the new z-score, it again turns out equal to +3.00 because 1 is three standard deviations ( $\sigma = .3$ ) above the mean ( $\mu = .1$ ).

What this little discovery exercise shows us is that the tenth person's z-score turns out equal to +3.00 no matter how large or small the tenth wisdom score is, so long as it is higher than the other nine scores and they are all identical. If it doesn't matter how far out the "outlier" lies, then it is a legitimate logical leap to conclude that the z-score will be equal to +3.00, even if the tenth wisdom score approaches positive infinity.

The second approach to solving this puzzle involves a small amount of simple algebra. Let's represent the tenth person's wisdom score as  $D$  for "discrepant." Because nine of the 10 wisdom scores are each equal to zero,

$\bar{X} = D / 10$ . Noting that  $s^2$  can be computed as the sum of all raw scores individually squared minus the mean squared,  $s^2$  turns out equal to  $9(D^2) / 100$  and  $s = 3D / 10$ . Inserting the values for  $\bar{X}$  and  $s$  into the formula for  $z$  and substituting  $D$  for  $X$ , we find that the z-score is equal to  $9D / 3D$ , or +3.00. Note that the tenth wisdom score, whatever it might be, cancels out in the last step of our algebraic solution to the puzzle.

It is worth asking what would happen to the last person's z-score if the number of original scores ( $N$ ) had been larger or smaller than 10. So long as all but one of the scores are identical, the z-score for the single outlier will turn out equal to  $\sqrt{N} - 1$ . Thus, if we had started with 26 scores, all identical except one, then the single discrepant score would have had a z-score of +5.00, regardless of whether it appeared—in the original dataset—to be close to or far away from the other 25 scores.

Finally, it should be noted that the  $N - 1$  scores that are identical need not be zero. So long as only one score is different from the rest, with all others being identical, the z-score for the outlier will be equal to  $\sqrt{N} - 1$ , no matter what the common value of the other scores.

Now, put your thinking cap on one more time and try this. What would be the tenth person's z-score if he or she was the humble person and claimed a wisdom score of only zero, while everyone else rated themselves identically with a wisdom score greater than zero? ■

Solution: Using the same algebra and logic, the humble tenth person's z-score would be -3.00.



# Are You a **Student** Majoring in Statistics?

First-time Student  
Members Pay

# \$10

/year

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

## Join NOW!

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

[www.amstat.org/join.html](http://www.amstat.org/join.html).