

STATS

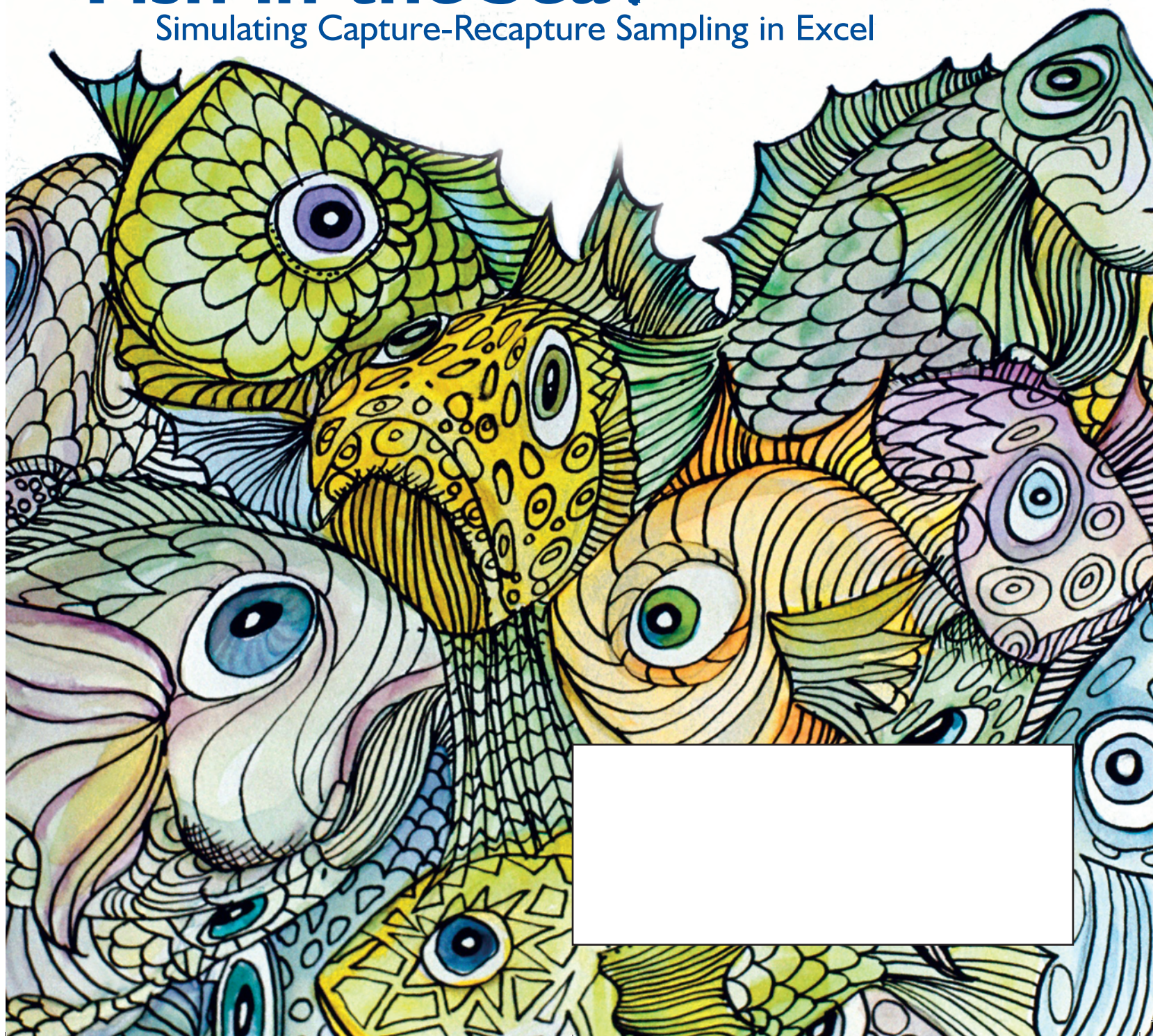
THE MAGAZINE FOR STUDENTS OF STATISTICS :: ISSUE 49

Stephen M. Stigler
remembers Karl Pearson
after 150 years

Lawrence Lesser
shares more fun ways to
learn stats

How Many Fish in the Sea?

Simulating Capture-Recapture Sampling in Excel



Looking for a **JOB?**

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

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

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

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

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

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

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

<http://jobs.amstat.org>

Visit the
ASA JobWeb online
TODAY!



STATS

contents

features

SPECIAL TOPICS



page 14

Remembering
Karl Pearson
After 150 Years 3

Even More FUN
Learning Stats 5

How Many Fish in
the Sea? Simulating
Capture-Recapture
Sampling in Excel 14

COLUMNS

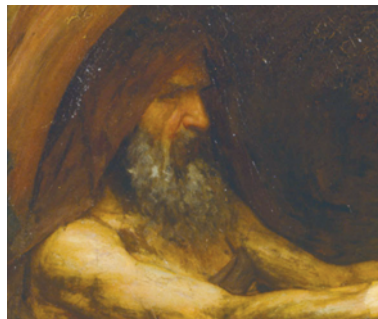
EDITOR'S COLUMN 2

STATS PUZZLER
A Perfect Correlation ...
If Not for a Single Outlier 9

ASK STATS
Busting Statistical Myths 10

STATISTICAL μ -SINGS
Diogenes: The Sampler
from Sinope 24

REFERENCES and
Additional
Reading List 27



page 24



page 20

PUZZLES

EXPLORE MORE
A Hands-On Capture-
Recapture Study Without
Going Outside 20

TRY THIS
STAT•DOKU 26

“

Statistics is the grammar
of science.”

Karl Pearson

guest writers

Remembering Karl Pearson After 150 Years

STEPHEN M. STIGLER, professor of statistics at The University of Chicago, is widely recognized as the world's leading expert on the history of statistics. His books, *The History of Statistics: The Measurement of Uncertainty Before 1900* and *Statistics on the Table*, provide insightful perspective into the fascinating history of statistics. He is the past president of the International Statistical Institute.



Even More Fun Learning Stats



LAWRENCE (LARRY) LESSER is an associate professor at The University of Texas at El Paso and on the Research Advisory Board of the Consortium for the Advancement of Undergraduate Statistics

Education (CAUSE). His web site and contact information are at www.math.utep.edu/Faculty.

How Many Fish in the Sea? Simulating Capture-Recapture Sampling in Excel

BRUNO C. DE SOUSA teaches statistics in the Departamento de Matemática para a Ciência e Tecnologia at the Universidade do Minho, Campus de Azurém, in Guimarães, Portugal. His research interests are in biostatistics, extreme value theory, and statistics education. bruno@mct.uminho.pt



EDITOR'S COLUMN

EDITOR

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



DIRECTOR OF PROGRAMS

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

EDITORIAL BOARD

Peter Flanagan-Hyde
Mathematics Department
Phoenix Country Day School, Paradise Valley, AZ 85253
pflanaga@pcds.org

Schuyler W. Huck
Department of Educational Psychology and Counseling
University of Tennessee
Knoxville, TN 37996
shuck@utk.educolsen@cr.k12.ia.us

Jackie Miller
Department of Statistics
The Ohio State University, Columbus, OH 43210
jbm@stat.ohio-state.edu

Chris Olsen
Department of Mathematics
Thomas Jefferson High School, Cedar Rapids, IA 53403
colsen@cr.k12.ia.us

Bruce Trumbo
Department of Statistics
California State University, East Bay, Hayward, CA 94542
bruce.trumbo@csueastbay.edu

PRODUCTION/DESIGN

Megan Murphy
Communications Manager
American Statistical Association

Val Snider
Publications Coordinator
American Statistical Association

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

STATS: The Magazine for Students of Statistics (ISSN 1053-8607) is published three times a year, in the winter, spring, and fall, by the **American Statistical Association**, 732 North Washington Street, Alexandria, VA 22314-1943 USA; (703) 684-1221; www.amstat.org. *STATS* is published for beginning statisticians, including high school, undergraduate, and graduate students who have a special interest in statistics, and is provided to all student members of the ASA at no additional cost. Subscription rates for others: \$15.00 a year for ASA members; \$20.00 a year for nonmembers; \$25.00 for a Library subscription.

Ideas for feature articles and materials for departments should be sent to Editor Paul J. Fields at the address listed above. Material must be sent as a Microsoft Word document. Accompanying artwork will be accepted in four graphics formats only: EPS, TIFF, PDF, or JPG (minimum 300 dpi). No articles in WordPerfect will be accepted.

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

Copyright (c) 2008 American Statistical Association.

In 2007, we celebrated the 150th anniversary of the birth of Karl Pearson. To start off this issue, statistics historian Stephen Stigler remembers Karl Pearson and some of the foundational contributions he made to statistical science as we know it today. Stigler provides fascinating historical insight into one of Pearson's discoveries that will probably surprise you.

In the last issue of *STATS*, Larry Lesser showed us that "Learning Stats is FUN ... with the Right Mode." He returns now to share even more fun ways to learn statistics. Be sure to check out the References and Additional Reading List at the end of the issue to find out about web sites and other fun-filled places to learn statistics. For example, at www.causeweb.org/resources, click on "Fun" and then "Song" to find statistics songs by Lesser (a.k.a. the Mathemusician) and other creative composers, along with a seemingly endless supply of other resources to help you learn statistics.

The *STATS* Puzzler always has something fun for us. Remembering that it was wKarl Pearson who gave us the Pearson product-moment correlation coefficient, Schuyler Huck shows that relying on our intuition alone can sometimes lead us in the wrong direction. It is better to use our knowledge of statistics. Try to solve his statistics puzzle, but think it through carefully.

To help clear up some of the misconceptions people have about statistics, we asked Jessica Utts to be a statistics myth buster. She identifies the eight most common misconceptions and walks us through each.

Our cover feature article comes from Bruno de Sousa, who teaches statistics in Guimarães, Portugal. He asks a fascinating question: "How many fish are in the sea?" This type of problem is called "abundance estimation," and de Sousa explains the basic statistical ideas we need to answer such questions using the sampling technique "capture-recapture." He shows how this technique works using a simple simulation model in Excel. We will be looking more deeply into capture-recapture in the next issue of *STATS*, so stay tuned.

Something new for this issue is a feature we call "Explore More." Jonathan Chipman, a statistics student, was 'captivated' by de Sousa's article and decided to explore capture-recapture sampling through a hands-on experiment, which we've included here. Try your hand at both de Sousa's mathematical simulation approach and Jonathan's physical simulation.

Then, Chris Olsen takes us all the way back to ancient Greece. In his Statistical μ -sings, he found a guy in Athens who also used sampling in an interesting way to try to estimate the size of a population of a rare species.

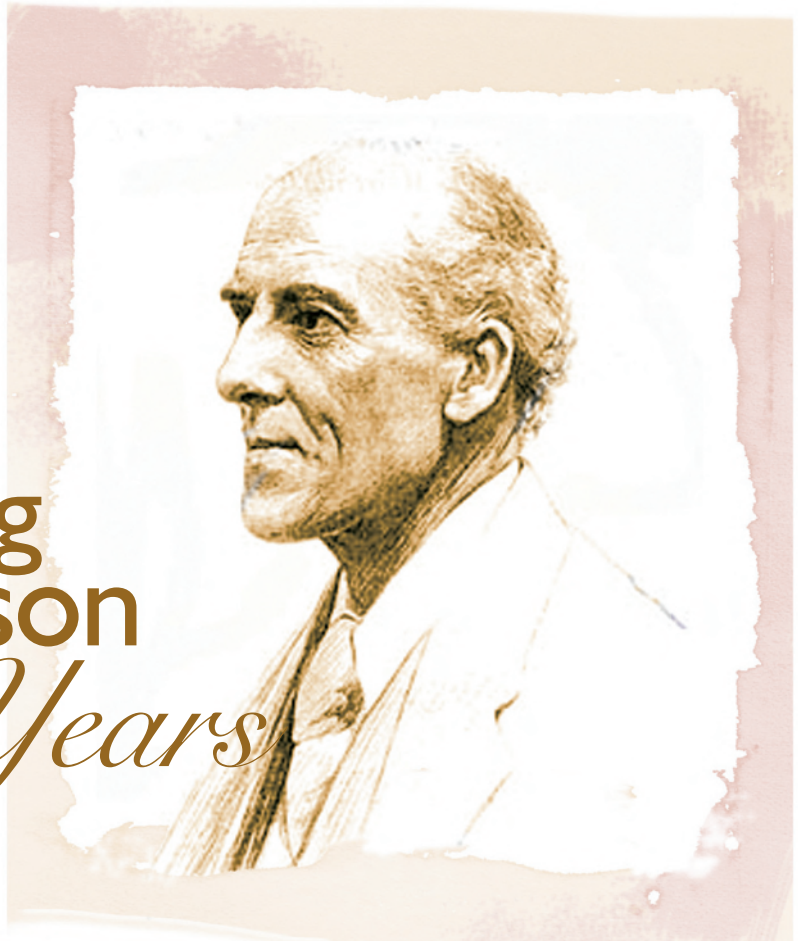
We end this issue with two new STAT·DOKU puzzles contributed to the Try This section by statistics student Amy White. The solutions are not provided, but if you are the first to send in your solutions with the correct answers to the statistics questions White asks, we will send you an ASA T-shirt (of course it will be a "Student's t " shirt).

And remember, we are interested in publishing articles that illustrate the many uses of statistics to enhance our understanding of the world around us. We are looking for engaging topics that will inform, enlighten, and motivate you. So, send us a detailed description of your concept for a feature article and you, too, could become a published author.

Paul J. Fields

Remembering Karl Pearson *After 150 Years*

by Stephen M. Stigler



KARL PEARSON, pencil drawing by F.A. de Bieden Footner, 1924

*K*arl Pearson was born 150 years ago on March 27, 1857; he died April 27, 1936. He made important contributions to statistics, applied mathematics, and genetics (when it was known as eugenics). He is usually remembered in statistics for two path-breaking achievements: his “product-moment” estimate of the correlation coefficient (dating from 1896) and his chi-square test (introduced in 1900). But on the 150th anniversary of his birth, let us recall a small but striking discovery he made in 1895 that is virtually unknown today, yet well worth knowing.

Everybody knows, as the standard lead-in goes, that the binomial distribution is “like” a normal distribution if the number of independent trials (n) is large and the probability of success (p) on a single trial is not too near 0 or 1. Everybody knows this because Abraham De Moivre told us so, and even proved it to be true in 1733. Some form of that statement has by now crept into every statistics text, elementary or advanced. Every student knows (or should know) that for most practical purposes, if you want to calculate the probability that a binomial count X will fall between two limits, $P[a \leq X \leq b]$, you would be foolish to do other than use a normal approximation, such as

$$P[a \leq X \leq b] \approx \Phi\{(b - \mu_n)/\sigma_n\} - \Phi\{(a - \mu_n)/\sigma_n\},$$

or even, to correct for the discreteness of the binomial,

$$P[a \leq X \leq b] \approx \Phi\{(b + .5 - \mu_n)/\sigma_n\} - \Phi\{(a - .5 - \mu_n)/\sigma_n\},$$

where $\mu_n = np$ and $\sigma_n^2 = np(1-p)$ are the mean and variance of X and Φ is the standard normal cumulative distribution function.

Because we are always reminded that this is an approximation (even De Moivre described it in those terms), there is a nagging doubt as to how close the binomial and normal distributions really are. They are, in fact, very different distributions (one is discrete and the other is continuous), and it would not be out of place to worry that the agreement may be at least a bit off. Indeed, some theoretical works emphasize the discrepancy and provide very complicated improvements. But Pearson discovered something quite remarkable: There is a case and a sense where the agreement is much, much closer than anyone would have a right to expect. In fact, if one particular definition of “agreement” is adopted, and if $p = 1/2$, the “agreement” is actually exact for all n (even for $n = 1$), providing one minor fudge factor is allowed.



KARL PEARSON in his office, 1934

Pearson's remarkable discovery involved the fundamental shapes of the two distributions, where by "shape" we mean the way the normal density function $f(x) = \{1/\sqrt{2\pi\sigma^2}\} \exp\{-(x-\mu)^2/2\sigma^2\}$ and the symmetric binomial probability function

$$P[X=k] = p(k) = \binom{n}{k} / 2^n \text{ change as } x \text{ and } k \text{ vary.}$$

To permit comparison, Pearson considered the binomial distribution as described by a polygon: Plot a dot at height $p(k)$ above each integer k , and then connect the dots to get the polygon. Figure 1 gives the polygon for $n = 4$.

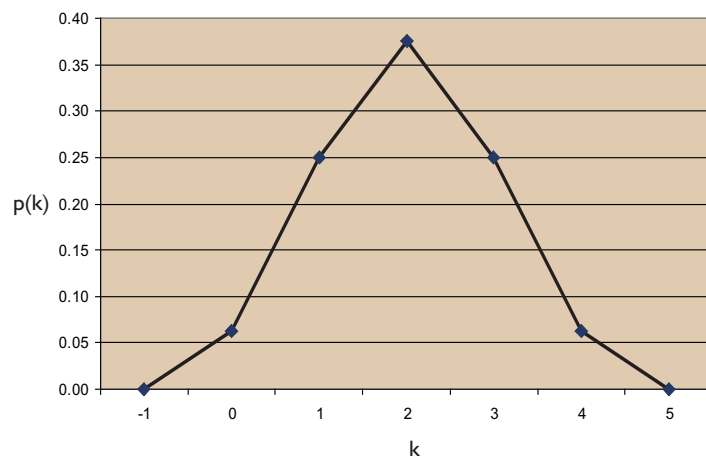


FIGURE 1. Binomial polygon of $p(k)$ when $n = 4$ and $p = .5$

Pearson looked at the relative rate of change—how the slope divided by the height changed as x and k varied. As Pearson noted, for the continuous normal density, this is given by the derivative of the natural logarithm of $f(x)$ and is a simple function of x , μ , and σ , namely

$$\frac{d}{dx} \log(f(x)) = \frac{f'(x)}{f(x)} = -\frac{(x-\mu)}{\sigma^2}.$$

The relative slope changes as a simple linear function of x , with intercept μ (for $x = \mu$, the slopes of both $f(x)$ and $\log(f(x))$ are zero), and at a rate of change that is inversely proportional to the variance, σ^2 . The surprise, as Pearson discovered by considering the binomial polygon, was that the binomial probabilities $p(k)$ satisfy the analogous difference equation *exactly*. As we are dealing with unit intervals $(k, k+1)$, the slope of the polygon in each interval will be the change $p(k+1) - p(k)$, and the height of the polygon at the midpoint will be the average $\{p(k) + p(k+1)\}/2$. Then, at each midpoint between a pair of successive integers, k and $(k+1)$, we have that the ratio of slope to height is, after some algebra,

$$\frac{\text{change } p(k) \text{ to } p(k+1)}{\text{average of } p(k) \text{ and } p(k+1)} = -\frac{\text{midpoint of } (k, k+1) - \mu_n}{\sigma_{n+1}^2},$$

$$\text{or } \frac{p(k+1) - p(k)}{(p(k+1) + p(k))/2} = -\frac{\left(k + \frac{1}{2}\right) - \frac{1}{2}n}{(n+1) \cdot \frac{1}{2} \cdot \frac{1}{2}}$$

for all $n \geq 1$ and all $0 \leq k \leq n$. [Challenge: Can you verify this?]

The appearance of $n+1$ instead of n in the denominator might be considered a minor fudge, but the equation still demonstrates a fundamental agreement. Upon discovering this remarkable identity, Pearson, himself, wrote,

Hence: this binomial polygon and the normal curve of frequency have a very close relation to each other, of a geometrical nature, which is independent of the magnitude of n . In short, their slopes are given by an identical relation. ... It is this geometrical property which is largely the justification for the manner in which statisticians apply, and apply with success, the normal curve to cases in which n is undoubtedly small.

Pearson discovered in 1895 that the normal and symmetric binomial distributions are more similar than De Moivre realized. I suspect most modern statisticians are equally unaware of this surprising identity, but now you know. ●

Even ^{More} FUN Learning STATS

by Lawrence M. Lesser

In the previous issue of *STATS*, we looked at how learning stats is fun ... with the right mode, but we ran out of room for everything, so we will continue in this issue, starting with a couple of statistics riddles.

- 1 What are a statistician's favorite two breakfast cereals? How do we know this?
- 2 What do statisticians eat for breakfast other than cereal?

The answers are on Page 19. Take a peek and see how you did.

Movies

Next time you take a break from your statistics homework to rent a movie, why not see if there is one with statistical content? The 2006 drama "Statistics," directed and written by Frank Robak, is the story of six everyday individuals who each "become statistics" by the end of the day. Despite the life-changing events, the movie ultimately aims to offer an uplifting message to cherish life today, because our day of death is uncertain.

Some films with a statistics theme are short, such as Tim Robertson's six-minute 2001 film called "The Statistician" that stars Jim Carruthers, Jill Duff, and Sandra Duff. There also is a 7.5-minute,



humorous 2004 film called “Statistically Speaking” that was named the co-winner of best film at the Lingos Film Festival in New York City and features playful uses of statistical-type words. (See the reference section for the web site where you can view the actual film.) This “Statistically Speaking” film is not to be confused with the full-length 1995 comedy of the same title described as “a middle-aged woman braves one outrageous date after another.”

the universe is deterministic or probabilistic, and whether WWII rocket hits in London follow a spatial distribution that is Poisson. What does that mean? Well, let’s imagine a map with a large square area of southern London divided into a 24x24 grid of 576 smaller square areas of 0.25 km² each. Let’s make this even more concrete by looking at some data from Page 481 of R. D. Clarke’s 1946 article “An Application of the Poisson Distribution” in *Journal of the Institute of Actuaries*:

Dear Editor, After reading Larry Lesser’s article, “Learning Stats is FUN ... with the Right Mode” in *STATS*, Issue 48, 2007, I wanted to check out some of his web sites and references. Several of them were quite funny. However, the best one was the statistics rap video, “statz rappers,” found at <http://video.google.com/videoplay?docid=489221653835413043>. My girlfriend and I laughed for a long time about that one, and we even told some of our friends how funny it was. Professor Lesser is right—Learning Stats is FUN ... especially when you put statistical concepts to rap.

John

Rocket Hits (<i>k</i>)	0	1	2	3	4	5	6	7
Observed number of 0.25 km ² areas with exactly <i>k</i> hits	229	211	93	35	7	0	0	1

From these data, let’s now estimate a value for the mean, λ , using the formula for a weighted average (check the answer on Page 19). To do a chi-square goodness-of-fit test, we will need to collapse the rightmost four cells into one, so the table becomes:

Rocket Hits (<i>k</i>)	0	1	2	3	4+
Observed number of areas with exactly <i>k</i> hits	229	211	93	35	8

And, of course, there are biographical-type movies about people who have played major roles in statistics, such as the 1995 film “The Passionate Statistician: Florence Nightingale,” a 25-minute historical re-enactment in which she uses applied statistics to disprove the medical assumptions of her day. (See the reference section for the web site for this statistical classic.) It is a definite “two thumbs up.”

Books and Videos

We all know the book is always better than the movie, so let’s talk about statistics books next. Books that use humor to explain big statistics ideas include *Statistics for People Who (Think They) Hate Statistics* and *PDQ Statistics*. Or, if you like the humorous approach in video form, you might enjoy something like the Standard Deviants series on statistics. Their web site is listed in the references, also.

There are statistically rich works of nonfiction, such as *Beat the Dealer*, the 1962 autobiography of statistics professor Edward O. Thorp, who showed that card-counting techniques could turn the casino’s advantage in blackjack into an advantage for the player.

And fiction fans might enjoy Thomas Pynchon’s *Gravity’s Rainbow*, which *Mathematics Magazine* described as “a complex novel set during World War II in which the characters search for meaning in a world dominated by the V-2 rocket.” Pynchon uses probabilistic imagery as characters debate whether

Use the value you calculated for λ to find the expected counts using the Poisson expression $(576)(e^{-\lambda}\lambda^k/k!)$. But wait. How will that formula handle the open-ended class (“4+”) ? No problem. Subtract the other four estimated expected frequencies from the total sample size of 576. What did you get? We can now verify that the Poisson distribution fits the dataset very well by plugging the observed and expected frequencies into the goodness-of-fit test statistic formula: $\chi^2 = \sum (O-E)^2 / E$. Check the answers and *p*-value on Page 19.

Are You Feeling Lucky?

Speaking of probability distributions, why not add spice to your statistics parties with a distribution of probabilistic refreshments? Consider a can of “Lucky Nuts” made by Dave’s Gourmet, Inc., whose label warns that every 10th peanut is (habanero) fiery hot, but the nuts all look alike.

First of all, it is interesting to learn about Scoville heat units (SHU), the original scale of measurement for the spice heat of a chili pepper. A typical habanero SHU rating of 300,000 means its extract has to be diluted 300,000 times before tasters can no longer detect the capsaicin chemical that stimulates our nerve endings. For comparison, a



jalapeño pepper is fewer than 10,000 SHU.

Okay, back to probability. What is the probability of getting at least one hot nut if you grab a handful of 10 nuts? It is not 100%; it is 65%. (Feeling lucky enough to grab that handful yet?) Let's use the binomial distribution and verify this. Recall that the probability of exactly k successes (perhaps we are 'nuts' to think of eating an habanero nut as a success) out of n trials that each have a probability of success p is given by the formula $\binom{n}{k} p^k (1-p)^{n-k}$.

With a handful of 10 nuts, $n = 10$, and since Dave said, "every 10th peanut is hot," $p = .10$. Consequently, $P(\text{Exactly } 0 \text{ habanero nuts in } 10 \text{ trials}) = .35$ and $P(\text{More than } 0 \text{ habanero nuts}) = 1 - .35 = .65$.

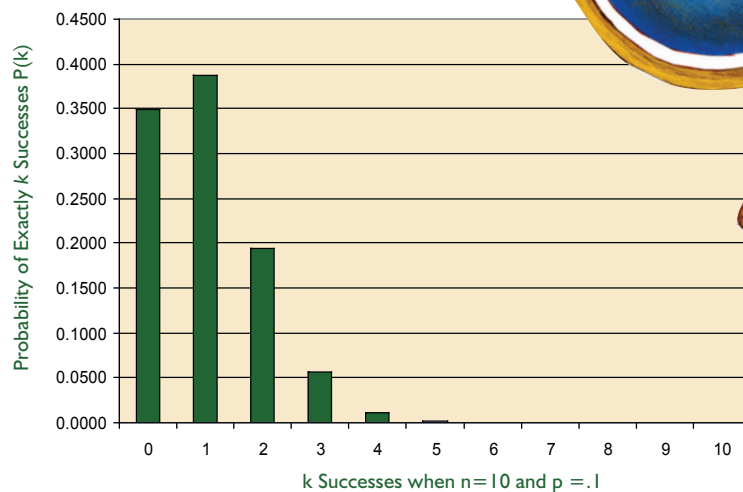


FIGURE 1. Binomial distribution of the probability of k successes in $n=10$ trials when $p=.1$

The binomial distribution ($n = 10, p = .10$) also is interesting because it is clearly right-skewed (see Figure 1). Therefore, we are much more likely to get small numbers of habaneros than large numbers per handful. This might lead us to assume incorrectly the mean must be larger than the median, and yet the mean, median, and mode are all equal. We can readily verify that the mean is 1 by computing np . The simple calculations $P(X = 0) = .9^{10} = .3486$ and $P(X = 1) = (10)(.1)(.9^9) = .3874$ are enough to deduce that the mode and median also are equal to 1. See an article by Paul von Hippel about this misconception of skewness, mean, and median cited in the reference section.

Another question for you and your friends to discuss: Is the can of nuts large enough to satisfy the condition of independent trials necessary to be able to use the binomial distribution?

And speaking of entertainment related to odds, you might find it fun to read through a book such as *The Odds on Virtually Everything*, whose title is self-explanatory.

Places

Another diversion is to get out your atlas and find statistical places to go, such as the following:

Uncertain, Texas

Pearson, Georgia

Independence, Missouri

Normal, Illinois

Speaking of normal, did you know a college founded to train teachers was called a "normal school"? Just be sure to stay on the right side of the median when you are driving. Maybe it would be fun to visit one of these towns and erect a marker explaining the statistical meaning of the town's name. To do that, though, you may need a variance ... a zoning variance.

Statistical Wordplay

Since it is good to know statistics backward and forward, can you make a statistical palindrome—a word or phrase that reads the same in both directions (such as this magazine's very name: *STATS*)? So far, the longest example I have made up is "MO' DNA RANDOM," which was inspired when I read that biostatisticians have found reason to analyze palindrome sequences in virus DNA. (Remember, the DNA alphabet consists of A, G, T, and U: adenine, guanine, thymine, and uracil.) I bet you can make an even better statistical palindrome; let us know.

An easier word game is creating or solving anagrams by rearranging letters of a word or phrase. For example, the word "name" can be rearranged to form the word "mean." You and your friends can try to see who can solve this score of anagrams the fastest (each word or phrase can be rearranged into a statistical word or phrase):

dome	rain cave	its logic
maiden	true oil	in a limbo
said rule	cool terrain	poisons
anger	next premise	persona
level curb	tour skis	dust tents
maples	salsa request	trap oily bib
meet irises	fenced icon	

The answers are on Page 19. Can you make up some fun new ones?

Anagrams are more than just wordplay; they relate to statistics because they are permutations. For example, how many ways are there to arrange the letters in the word “stop”? The four letters can be arranged in $4!$ ways, $4! = 4 \times 3 \times 2 \times 1 = 24$. What is also interesting is that a whopping six of these 24 ways are actually English words. (Can you find them all?) So, for the word “stop,” the probability of a random permutation resulting in a word is $6/24 = 1/4$. That means the odds are 1:3.

Now, what about the number of permutations possible if some letters are repeated? Consider the word “Mississippi.” If the 11 letters were all different, there would be $11! = 39,916,800$ arrangements possible, but swapping the Ps does not create a new arrangement. Neither does rearranging the Ss or Is, so we have to divide $11!$ by $4!$ for the ways to arrange the Is, by another $4!$ for the ways to arrange the Ss, and then by $2!$ for the Ps. So, the answer for “Mississippi” is only 34,650. Okay, now you try one. How many permutations are possible for the word “statistics”? Did you get the same answer as the one on Page 19?

Did you know permutation calculations also can be used to determine a benchmark for the uniqueness or originality of a melody—a crucial calculation in copyright infringement court cases? Let’s consider the pitches (i.e., note frequencies, ignoring note durations) for the opening of the song “Home on the Range.” (Hey! Range is a statistics word.) You do not have to be a musician to recognize there are duplicate notes in this melody excerpt.

In how many ways can the first five notes be arranged? In how many ways can the first 11 notes be arranged? Try the calculations yourself and check them against the answers on Page 19. Could this lead

Syllable:	Oh	give	me	a	home	where	the	buf-	fa-	lo	roam
Note:	G	G	C	D	E	C	B	A	F	F	F



to ‘repeated measures’ analysis? Does this give you insight into how much of a song’s melody might be ‘sampled’ by someone else before suspicions would be raised?

Songs

Speaking of song, let’s close with more fun statistics song parodies. Lyrics and even sound files for many statistics songs are available in the CAUSEWeb fun resources collection (see the reference section for the web site).

Here are some of my favorites from the songs I have parodied recently:

MLE. Sung to John Lennon and Paul McCartney’s classic hit “Let it Be.”

Hit Me with Your Best Plot. Sung to the tune of Eddie Schwartz’s 1980 “Hit Me with Your Best Shot” – a #9 hit for Pat Benatar

What’s the Average. Sung to the tune of Sammy Cahn and Jimmy Van Heusen’s “Love and Marriage” – a #5 hit for Frank Sinatra in 1955, one of his dozen million-sellers, and also used as the theme song for the 1987–1997 hit TV series “Married ... with Children”

Y Hat Dance. Sung to the tune of “The Mexican Hat Dance”

Chi-Square for Us. Sung to the tune of James Rado, Gerome Ragni, and Galt MacDermot’s Grammy-winning 1969 #1 hit song “Aquarius” from the musical “Hair”

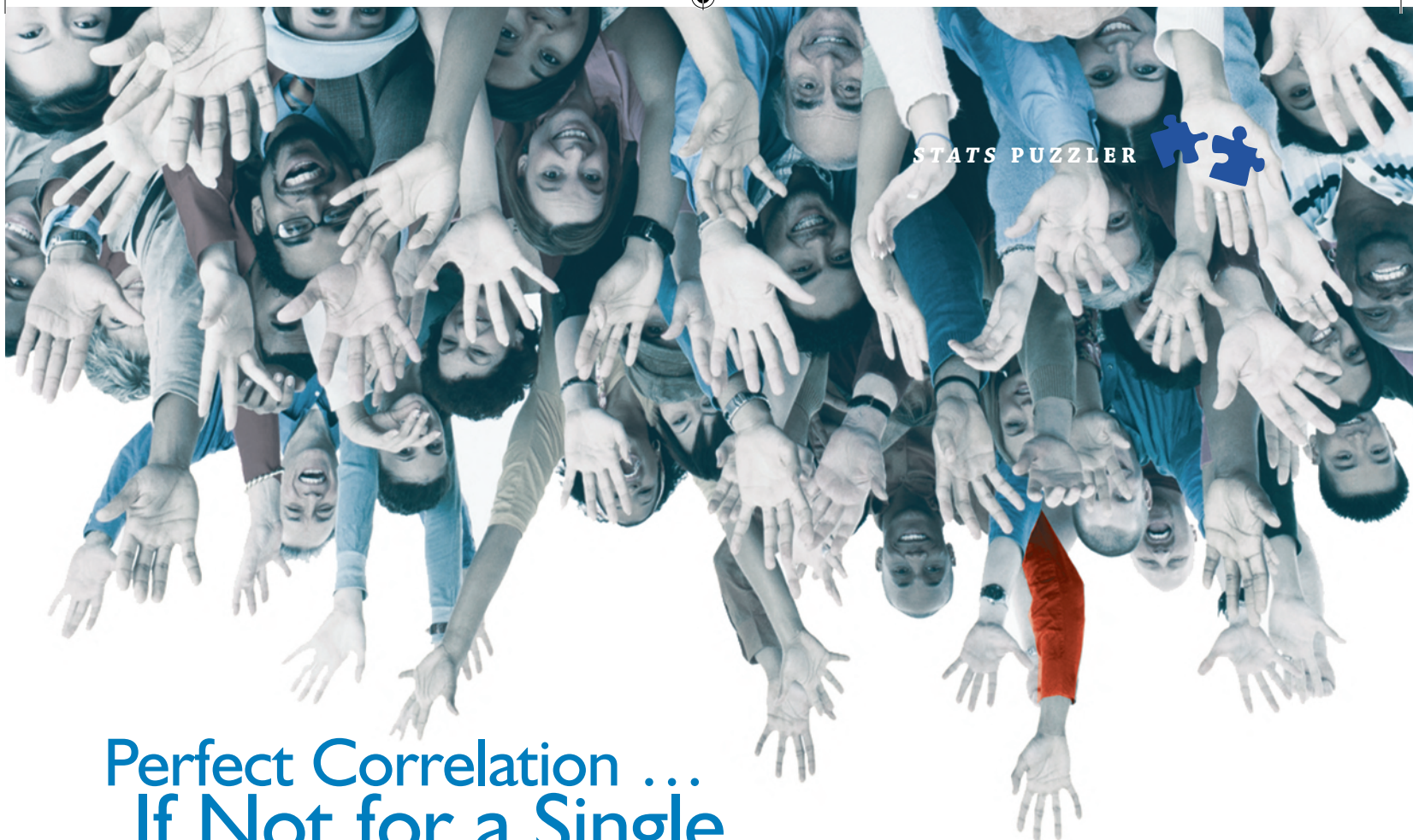
It’s Only Just a Poll (But I Like It). Sung to the tune of Mick Jagger and Keith Richards’ #16 hit “It’s Only Rock’n Roll (But I Like It)” from the same-titled 1974 album

Go to a local music store or an online karaoke music store and get karaoke discs with the original songs I have parodied. Most of them were big hits, and so should be available and in stock. You also can do a Google search for any song you want, adding “+midi” to the search field to find a free MIDI file of the music.

Check out Lynda Williams’ how-to tutorial (see the reference section) for using PowerPoint to make your own sing-along karaoke. Now you can throw a statistics karaoke sing-along party, where you and your friends sing the (new) statistics lyrics over the musical accompaniment. Maybe you have some musicians in your department who could provide live accompaniment.

The Fun Never Stops

I hope you have been able to think of even more ways to make learning statistics fun from reading this article. If you are willing to share some of your creative ideas, contact me at lesser@utep.edu. I would be delighted to hear from you. The fun of learning, and the fun of learning statistics, never stops. ●

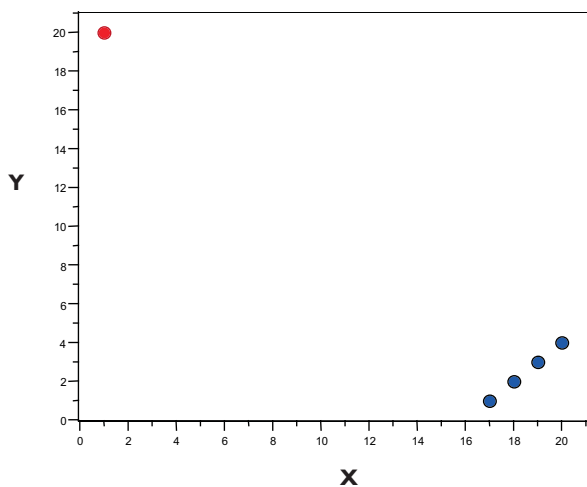


Perfect Correlation ... If Not for a Single

Outlier

This puzzle involves a plot of the data for a group of people measured on two variables, X and Y . Let's consider the group to be a sample from a population, rather than the entire population. Your job is to look at the data and then guess the size of Pearson's product-moment correlation for the scores in the scatter plot. Your guess should be a two-digit number between -1.00 and $+1.00$, the range of possible values for Pearson's r .

Although the scatter plot shown here appears to contain just five data points, there actually are 101 pairs of X, Y scores in the data set. That is because



each of the four dots in the lower right-hand portion of the scatter plot represents 25 individuals. In other words, 25 people have a score of 17 on X and a score of 1 on Y , another 25 people have a score of 18 on X and a score of 2 on Y , etc. There is only one person, however, who has a score of 1 on X and 20 on Y .

If the data set had included only the 100 scores located in the lower right-hand portion of the scatter plot, Pearson's r would be $+1.00$. For those 100 scores, there is a perfect positive correlation between X and Y because all those scores lie on a straight line that tilts from lower-left to upper-right.

The single point at $(1, 20)$ obviously makes it impossible for Pearson's r to equal $+1.00$ for all 101 pairs of scores. In a sense, the addition of that single 'nonconformist' data point causes r to move away from $+1.00$ toward -1.00 . But, how far in that direction does r move? What is your guess as to r 's numerical value for all 101 pairs of scores? ●

See Page 23 for the puzzle's solution.



SCHUYLER W. HUCK

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



We asked JESSICA UTTS, a leader in statistics education, if she would be our statistical myth buster and identify and clear up the most common misconceptions people seem to have about statistics. Here are our questions and her answers.



JESSICA UTTS

BUSTING STATISTICAL MYTHS

by Jackie Miller



JACKIE MILLER 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. When not at school, Miller enjoys a regular life (despite what her students might think), including keeping up with her many dogs!

What are some common misconceptions in statistics?

First, let's separate practical misconceptions from technical misconceptions. A practical misconception is one that could lead someone to make a bad decision in life. A technical misconception is a statistical misunderstanding, but one that does not really affect how someone will make a decision. Although the technical misconceptions can drive statistics teachers crazy, they typically are of little practical importance to students or anyone else.

Here are some common misconceptions. Look over the list and decide for each one whether you think it is a practical or technical misconception. Remember, none of these are true statements; they are all misconceptions.

MISCONCEPTION #1: Statistics is a boring subject and has little relevance in daily life, so it does not matter if you remember anything you learn about it.

MISCONCEPTION #2: What is most important in a study is whether the results were statistically significant. If so, then an important difference or relationship exists. If not, then no difference or relationship exists.

MISCONCEPTION #3: It is okay if a study looks at many relationships, but only reports those that are statistically significant.

MISCONCEPTION #4: If an observational study finds a relationship between the explanatory and response variables, then it is probably a cause-and-effect relationship, especially if it makes sense in the context of the study.

MISCONCEPTION #5: The p -value is the probability that the null hypothesis is true.

MISCONCEPTION #6: When doing an analysis, we can assume all variables have normal distributions.

MISCONCEPTION #7: After you compute a 95% confidence interval for a mean, you can say the probability is 95% that it contains the population mean.

MISCONCEPTION #8: The sampling distribution for the sample mean is normal, whether or not the original population is normal.

If you said the practical misconceptions are the first four, then I agree with you. The remaining four misconceptions are serious, but they probably would not affect any decision you need to make, so they fit into the category of technical misconceptions.

What are the corrected versions of these misconceptions?

Let's clear up these misconceptions. We will focus on the practical ones first.

Misconception #1: Statistics is a boring subject and has little relevance in daily life, so it does not matter if you remember anything you learn about it.

Statistics affects our lives in ways most people never realize. Here are some examples of common elements in life for which statistics most likely plays a role:

- The cost of insurance for your car
- When new movie you want to see will be released
- Whether the store has your size in stock when you want to buy a certain pair of shoes
- The best way to grow, process, ship, and sell the food you eat
- Whether you are a good risk for a credit card company
- How to design a calling plan for your phone that will appeal to you and make money for the company
- Whether an email message from your friends is identified as spam by the filter in your email program
- Which students are most likely to succeed at your college, based on information from their high-school records, standardized exam scores, and other admissions materials

Statistics is literally everywhere in all aspects of life today.

Misconception #2: What is most important in a study is whether the results were statistically significant. If so, then an important difference or relationship exists. If not, then no difference or relationship exists.

There are three misconceptions rolled into one here; each of the three statements is wrong. First, statistical significance alone tells you nothing about how important the result is, even if the p -value is extremely small. Importance of the

results is based on their impact on decisions that will be made. Second, if the sample size is large enough, a minor difference or relationship could produce a small p -value, and some people interpret a small p -value to mean an important relationship exists. You need to know the magnitude of the difference or strength of the relationship that was found to assess how important it is. Third, if the sample size is too small, a very important difference or relationship could be missed. Lack of statistical significance means there is insufficient evidence of a difference or a relationship. When using samples to test hypothesis, size matters.

Let's consider an example. Suppose you are interested in selling a new pill that you claim raises peoples' IQ for 12 hours after they take it. It would be perfect for students who have to take that important final exam or college entrance exam. You know you can sell hundreds of thousands of these pills if you show they raise IQ on average by a 'significant' amount. So, it is worth it to you to invest heavily in doing this study. You recruit 6,000 people to participate, randomly assign half (3,000) to take the pill and half (the other 3,000) to take a placebo. Then, you measure the IQ scores for all 6,000 participants. You find a sample mean difference of two IQ points, as the mean was slightly higher for the group that took your new pill. The standard deviation was 15 points in each group. Voila! A two-sample t -test gives $t = 2.58$, p -value = .005. Now you can advertise, "Pill found that significantly increases IQ!" Or, you might say "Increase in IQ after taking new pill is highly significant!" The problem is that the sample means differed by only two IQ points, and most people would not think of that as having any practical importance.



Now suppose you are in a different situation with your miracle pill. It is very expensive to produce the pills, but you do want to conduct a study to show the pills work. You can afford to assign only 10 people to each group. In your samples, the group that took the real pills had a mean IQ of 110, while the placebo pill group had a mean IQ of 102. A difference of eight points—four times larger than in the first situation. That could be important to people. However, there is a problem: The difference is not statistically significant. A two-sample t -test gives $t = 1.19$, p -value = .124. You naively report that your pill did not work. Oh well, back to the drawing board.

But wait, if we look at the confidence intervals, it appears your pill might have worked much better in the second version than in the first version of the experiment. A 95% confidence interval for the difference in the population means for the first experiment was 1.2 to 2.8 IQ points, while for the second experiment it was -6 to +22 IQ points. In other words, for the first experiment, the difference is conclusively above zero, but is clearly very small and of little practical importance. In the second study, the difference could be zero, but it also could be as high as 22 IQ points, and most of the confidence interval is above 0. Which study indicates there might be an important difference? The study that was not statistically significant.

One antidote to this problem is to produce a confidence interval to accompany the results of hypothesis testing whenever it is feasible. A confidence interval provides information about the possible magnitude of the effect, and the width of the interval provides information about how accurate the results are. Most people can easily understand that the width of a confidence interval depends on the size of the sample, but have a harder time understanding how the size of the sample affects the p -value. In reality, sample size plays a major role in both.

Okay, now let's dispel the technical misconceptions. I will not go into great detail about them because you can read about them in most introductory statistics textbooks.

This misconception is an example of "confusion of the inverse," in which a conditional probability in one direction is confused with a conditional probability

Misconception #5: The p -value is the probability that the null hypothesis is true.

in the other direction. A common example of confusion of the inverse is that physicians often confuse the conditional probability of having a positive test for a disease, given that you have the disease, with the conditional probability that you have the disease, given that you have a positive test. For a rare disease, the first probability can be very high, but the second probability can still be very low. Needless worry can result from this confusion.

The same confusion leads to this misconception of the meaning of a p -value. In truth, a p -value is the conditional probability of observing the data observed (or something more extreme), given that the null hypothesis is true. Instead, many people think it is the probability that the null hypothesis is true, given the data observed.

This misconception has probably come about because, for many years, statistics textbooks wrote about "checking the assumptions," instead

Misconception #6: When doing an analysis, we can assume all variables have normal distributions.

of "checking whether the conditions are met," for statistical inference. For many procedures, one of the necessary conditions is that the original population is normal, or approximately so. That does not mean we can assume the original population is normal. We need to check, as much as possible, by using appropriate graphical techniques, whether this condition seems to be reasonably met. Too often, textbooks (and teachers) gloss over this distinction because it is much more time-consuming and tedious to "check" rather than "assume." A useful guide is "assume nothing; check everything."

A confidence interval is the outcome of a random circumstance, just like observing a head or a tail when a coin is tossed is the outcome of a random circumstance. Once a coin has been tossed, and shows a head

Misconception #7: After you compute a 95% confidence interval for a mean, you can say that the probability is 95% that it contains the population mean.

facing up, it no longer makes sense to talk about the probability of getting a head. You did get a head.



Misconception #3: It is okay if a study looks at many relationships but only reports those that are statistically significant.

Unfortunately it is common practice to test lots of hypotheses in the same study and report only those that are statistically significant, but it is not

good statistical practice. If 20 independent hypothesis tests are done at an $\alpha = .05$ level, then by chance alone, one will be statistically significant on average. There are statistical methods that adjust for this problem, called "multiple comparison methods," but they are often overlooked by researchers. If you read about a study that makes headlines because of one or two statistically significant findings, be sure to find out how many other relationships were tested as part of the study.

Misconception #4: If an observational study finds a relationship between the explanatory and response variables, then it is probably a cause-and-effect relationship, especially if it makes sense in the context of the study.

This misconception is probably the most common and most dangerous of them all. It is very easy to fall into this trap. For example, if a study were to find that vegetarians have lower blood pressure than nonvegetarians, a natural inclination would be to think that

if you were to become vegetarian, your blood pressure will go down. But, that is not a valid conclusion from such a study. There may be all kinds of additional differences between vegetarians and nonvegetarians, any of which could be responsible for the difference. Examples include alcohol consumption, exercise, other health problems, eating organic versus nonorganic food, water consumption, and so on.

Similarly, once a confidence interval has been computed, it does not make sense to talk about the probability that it contains the population mean. It does or it does not. Of course we do not know if it does or does not, unlike the coin-toss example, but that does not change the fact that the outcome is now determined. The probability interpretation goes with the confidence level of 95% and the procedure used to compute the confidence intervals. We can say that the procedure produces a confidence interval that covers the true population mean with probability .95. Similarly, we can say that tossing a coin produces a head with probability .5. Once the toss is over, the head is either there or not. Similarly, any one confidence interval either covers the true mean or it does not.

Misconception #8: The sampling distribution for the sample mean is normal, whether or not the original population is normal.

Statistics teachers love to quibble over wording in situations such as this. The truth is that the sampling distribution

is *approximately* normal, if the sample size is large, whether or not the original population is normal. Ho hum. One situation for which this misconception may matter is for anyone conducting a study using a small sample for which there are extreme outliers or skewness. If you ever find yourself in that situation, you can get out your old stats book and read about what to do, such as using nonparametric procedures.

What is the best way to make sure we do not fall prey to any of these misconceptions?

I think it is important to recognize these common misconceptions and focus on examples you are likely to remember that illustrate how these misconceptions could lead you astray. Most people have heard of the example used to show that “correlation does not imply causation.” Someone collected some data that showed a strong correlation between the number of babies born and the number of storks nesting in chimneys for towns in Europe. Can we conclude, therefore, that storks bring babies? Well, perhaps you want to think this is true, but it is more likely that a third variable—the size of the town—is responsible for the large numbers of storks and babies in some towns and small numbers of them in others.

When you think about what is important in a statistical study, spend less time focusing on the computational details, and more time on interpreting the results. Think about the entire process of statistical reasoning:

Formulating a question about the world

Creating hypotheses as proposed answers to the question

Designing a study to obtain data relevant to the question

“When you think about what is important in a statistical study, spend less time focusing on the computational details, and more time on interpreting the results.”

Collecting and analyzing the data

Interpreting the results

Seeing that process done right, and interpreted correctly, for a variety of interesting questions is the most logical method for making sure you appreciate the power of statistics, but do not harbor misconceptions about what can be concluded from statistical studies.

Choose some studies that interest you and see if any of the misconceptions we have discussed appear in them. Most universities provide free online access to journals, and you can find all the details you need to see if the researchers perpetuated any misconceptions in their reports.

What is the best way to debunk misconceptions when we find them?

The best way to debunk misconceptions is through actual practice. In one of my recent classes, the students conducted a survey to examine a relationship between two variables they were almost certain would be related. They were disappointed to find that the *p*-value for the chi-square test was only 0.12. They learned the importance of sample size from that experience in a much more direct way than if I had hammered them with the concept. They realized that if they had been able to include more people in their study (there were 86) and found the same difference in proportions for the two groups, the result could have been statistically significant.

While it may not be feasible to conduct studies to debunk each misconception you come across, you can ask questions such as the following:

How many relationships were examined?

If none was statistically significant, would any appear to be solely due to chance?

What confidence intervals can be computed to accompany any significant results that are reported?

Based on the confidence intervals, do the results appear to be of practical importance?

If a *p*-value is larger than .05, what would happen to it if the sample size was 10 times larger, but the sample statistics—such as the mean and standard deviation—were the same?

Statistics is all about learning things about the world—hopefully things useful to us. Keep the misconceptions in mind, and see what you discover as you examine the world around you. That way you can be a statistical myth buster, too. ●

JESSICA UTTS is a professor of statistics at the University of California, Davis. She has published extensively, including her popular books *Seeing Through Statistics* and *Mind on Statistics*. She is a Fellow of the American Statistical Association, the Institute of Mathematical Statistics, and the American Association for the Advancement of Science. Her research interests include applied statistics, parapsychology, and statistics education and literacy.

How Many Fish in the Sea?

Simulating Capture-Recapture Sampling in Excel

by Bruno C. de Sousa

If we want to estimate the abundance of a certain population of wildlife, maybe the number of fish of a certain species in the sea, we cannot just take a simple random sample to find out. Why is that? Because we would need to know the size of the population, N , from which we drew the sample. The parameter, N , is exactly what we do not know in our study and what we want to estimate. How can we do this without taking a census—a complete count of the entire population? Capture-recapture sampling is the answer.

Capture-Recapture Sampling

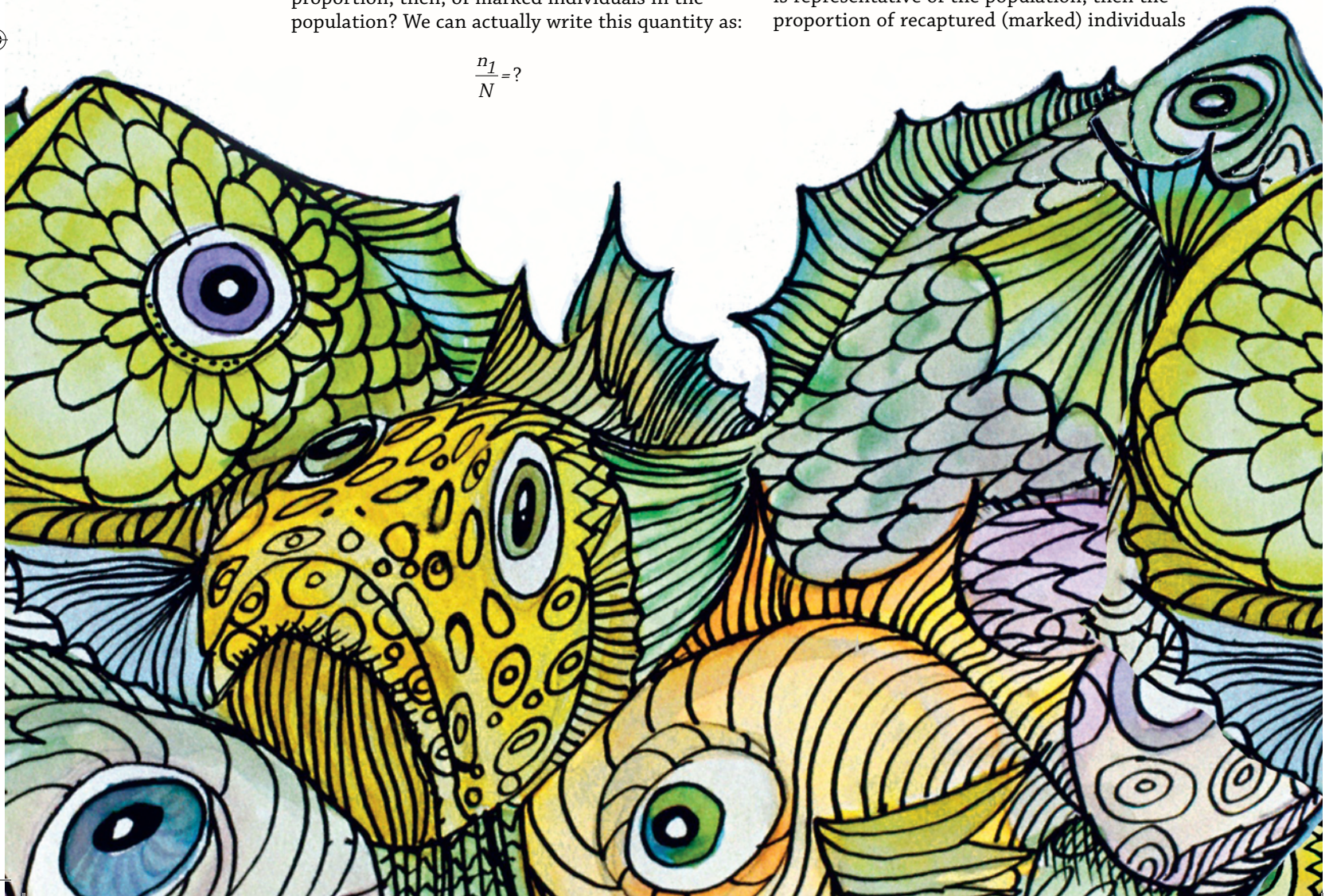
Imagine we first capture a sample of size n_1 from a population of interest and mark (or tag) these individuals and let them go. What will be the proportion, then, of marked individuals in the population? We can actually write this quantity as:

$$\frac{n_1}{N} = ?$$

But, unfortunately, its value is unknown, as we do not know the value of N . To solve this problem, after the marked individuals have dispersed evenly throughout the population, we take a second sample of size n_2 from the population— independent from the initial sample—and count the number of marked individuals in it. Let r be the number of individuals from the first sample who were recaptured in the second sample. Now, the proportion of marked individuals in the second capture is:

$$\frac{r}{n_2}$$

and it is a known quantity. What can we do with these quantities? If the second sample is representative of the population, then the proportion of recaptured (marked) individuals



in the second sample should be the same as the proportion of marked individuals in the population. So, we can set the quantities equal to each other:

$$\frac{n_1}{N} = \frac{r}{n_2}$$

Solving this equation in terms of N , we can easily propose an estimator for the number of individuals in a population. The estimator is then equal to:

$$\hat{N} = \frac{n_1 \times n_2}{r}$$

Can you see any problem with this estimator? Imagine there were no marked individuals in the second capture. If $r = 0$, that would be a problem. To deal with this possibility, consider an alternative estimator for N :

$$\tilde{N} = \frac{(n_1 + 1) \times (n_2 + 1)}{r + 1} - 1$$

This modified estimator, although less intuitive, solves the problem of $r = 0$ and reduces the bias in the population estimate.

Practical Issues

In practice, when N , n_1 , and n_2 are large, the two estimators, \hat{N} and \tilde{N} , will be essentially equal, and $r = 0$ can be very unlikely. Therefore, we would expect \tilde{N} to be particularly appropriate for small populations.

The entire process in capture-recapture sampling can be simplified if the individuals in the population of interest can be clearly identified

through unique natural markings, for example on whales' tails or fins. In such cases, there is no need to add new marks to the animals, and, consequently, we simply register these initially captured individuals and proceed with the method.

Now the question becomes, "Can we see what is going on in capture-recapture sampling without 'going fishing'?" For that, we can use a simple procedure to simulate capture-recapture sampling using Microsoft Excel.

Simulating with Excel

Let's say we have a population of size 250 individuals, $N = 250$, and that we randomly mark 50 individuals on the first capture, $n_1 = 50$. For the second capture, we will randomly sample 25 individuals from the population, $n_2 = 25$, and count the ones that were marked in the first capture, r . To illustrate this process, we will build an Excel spreadsheet.

First, open a blank Excel spreadsheet. So that the spreadsheet recalculates values manually, instead of automatically, go to the "Tools" menu and choose "Options." On the "Calculations" tab, check "Manual." Now, the spreadsheet will recalculate values only when function key F9 is pressed. In the Excel spreadsheet, take the following steps:

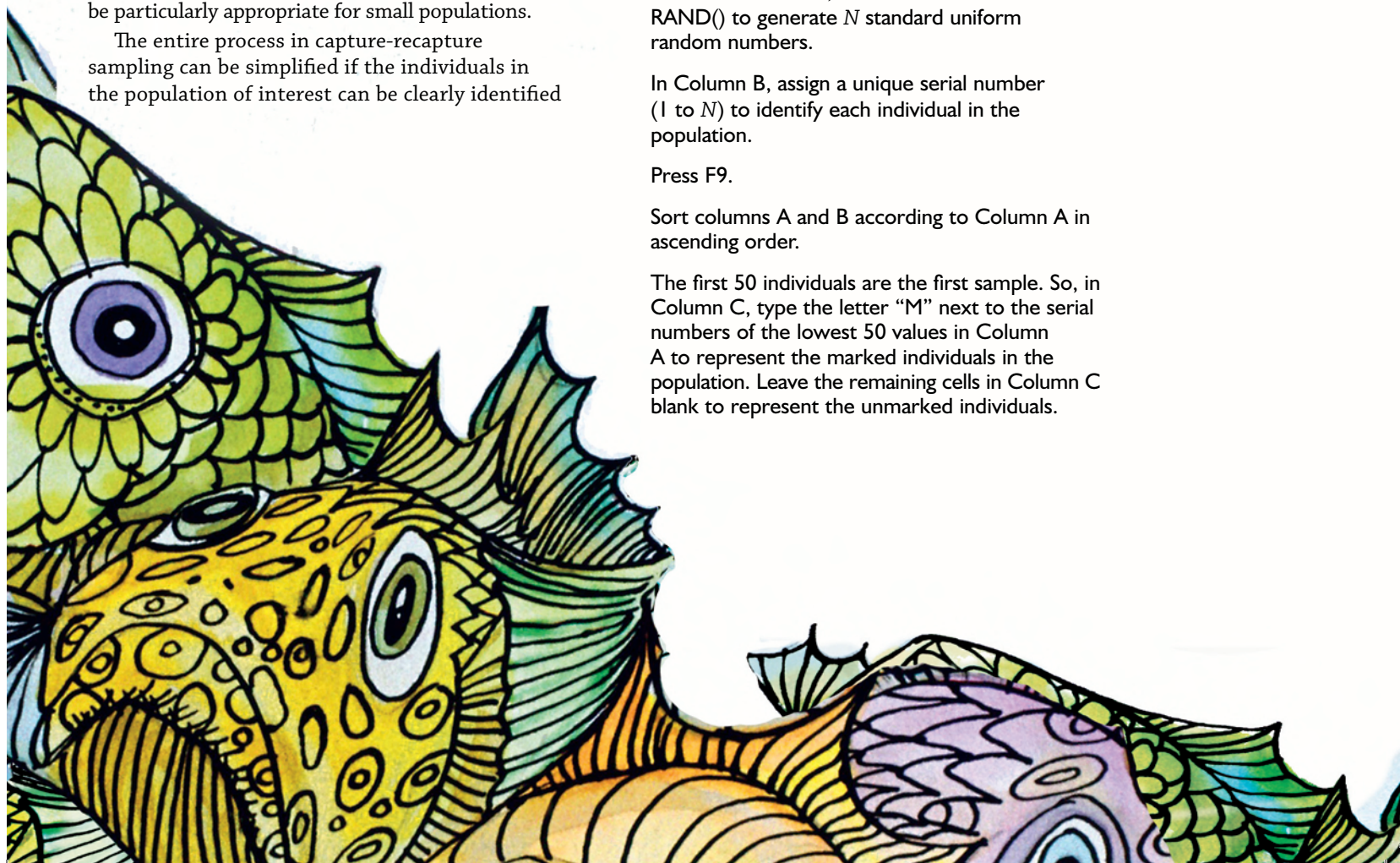
In columns A and E, insert Excel's function `RAND()` to generate N standard uniform random numbers.

In Column B, assign a unique serial number (1 to N) to identify each individual in the population.

Press F9.

Sort columns A and B according to Column A in ascending order.

The first 50 individuals are the first sample. So, in Column C, type the letter "M" next to the serial numbers of the lowest 50 values in Column A to represent the marked individuals in the population. Leave the remaining cells in Column C blank to represent the unmarked individuals.



Consider Table 1, which shows how the procedure is implemented in an Excel spreadsheet.

	A	B	C
1	Random Number	Serial Number	First Capture
2			
3	0.00278	156	M
4	0.00528	223	M
5	0.00811	98	M
6	0.01381	225	M
7	0.01403	159	M
8	0.01822	186	M
9	0.01844	241	M
...
52	0.23231	9	M
53	0.98719	50	
54	0.99526	195	
...
252	0.99632	33	

TABLE 1. Example simulation of the first capture of 50 individuals out of a population of 250. The 50 individuals are marked with “M.”

For the second capture, do the following:

Copy columns B and C to columns F and G.

Press F9.

Sort columns E, F, and G according to Column E in ascending order.

The first 25 individuals are the ones that belong to the second capture. Count how many of these 25 individuals are marked with an “M” in Column G.

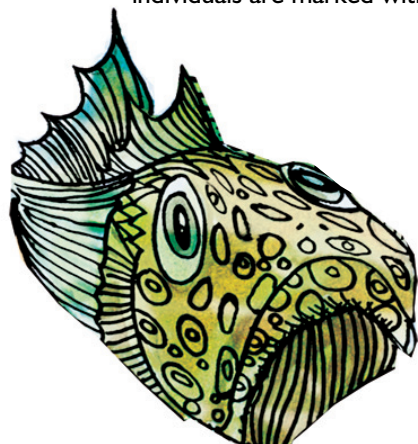


Table 2 shows an example of the results of one simulated recapture sample.

	E	F	G
1	Random Number	Serial Number	Second Capture
2			
3	0.00270	197	
4	0.00524	37	
5	0.00987	40	
6	0.01187	9	M
7	0.01352	136	
8	0.01558	151	M
9	0.01848	241	M
10	0.02030	33	
11	0.02067	16	
12	0.02112	107	
13	0.02145	26	M
...
27	0.07648	60	

TABLE 2. Results of the first iteration simulating sampling 25 individuals. The recaptured individuals are identified by “M.”

Use Excel’s function COUNTIF(range, “M”) to count the number of individuals recaptured (the “Ms” in Column G). Record the number of recaptures in a table such as Table 3. Repeat the process 100 times. This will give 100 iterations of the simulation. Of course, the more iterations you do, the better the results should be.

Table 3 shows the estimated values for \hat{N} and \tilde{N} obtained for the first 20 repetitions of the procedure. In the first run of this experiment, we see we have four individuals who were recaptured, giving the respective estimates for N of 312.5 using the \hat{N} estimator and 264.2 using the \tilde{N} estimator.

Iteration	Number Recaptured	\hat{N}	\tilde{N}
1	4	312.5	264.2
2	8	156.3	146.3
3	5	250.0	220.0
4	3	416.7	330.5
5	6	208.3	188.4
6	9	138.9	131.6
7	7	178.6	164.8
8	3	416.7	330.5
9	5	250.0	220.0
10	2	625.0	441.0
11	8	156.3	146.3
12	2	625.0	441.0
13	5	250.0	220.0
14	7	178.6	164.8
15	8	156.3	146.3
16	8	156.3	146.3
17	5	250.00	220.00
18	4	312.50	264.20
19	5	250.00	220.00
20	3	416.67	330.50
...

TABLE 3. Results from the first 20 iterations out of 100, estimating N with \hat{N} and \tilde{N}

We see immediately that some of the values are very close to, or even the same as, the true value of $N = 250$, but the values vary greatly between iterations. Were all the runs done correctly? Absolutely, but unfortunately, we do not always obtain the optimal value of 5 recaptured individuals, leading to an estimate of $\tilde{N} = 250$, the true value of N . But, which of the estimators, \hat{N} or \tilde{N} , is performing better? To answer this

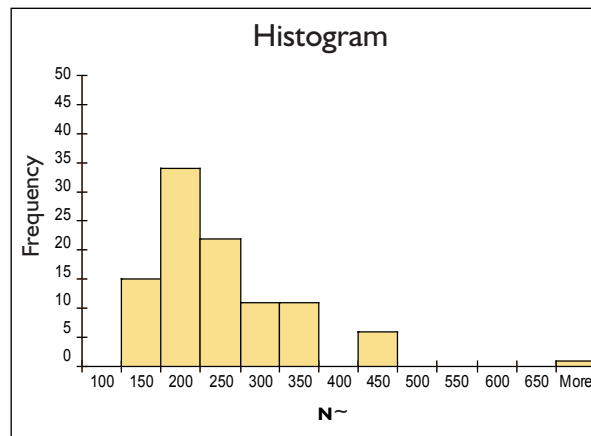
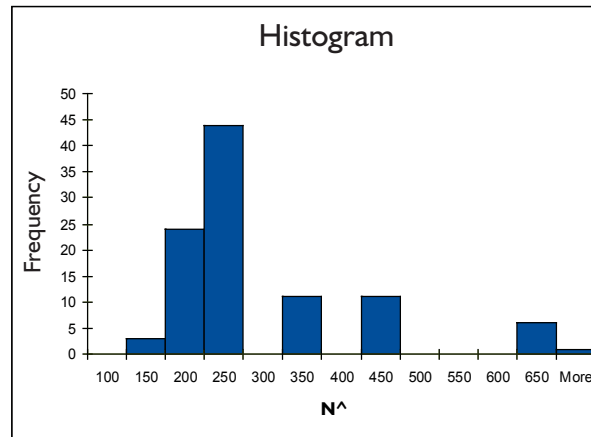


FIGURE 1. Population estimates based on \hat{N} and \tilde{N} using simulation for a population with $N = 250$

question, we can compare histograms of the estimated values for both estimators. In Figure 1, we see the simulated sampling distribution of \hat{N} and of \tilde{N} .

We observe that the simulated sampling distribution for \tilde{N} seems to be more consistently around the true value of the parameter, $N = 250$, with somewhat more extreme values when using the estimator \hat{N} than when using \tilde{N} . This also can be seen in the graph in Figure 2, which displays the estimates for N showing the estimators \tilde{N} in yellow and \hat{N} in blue. Although they are close most of the time, we see that when the estimates are large (look at the peaks), the values given by \hat{N} are much higher. We must be cautious in drawing conclusions too strongly, as this simulation only used 100 iterations. These results should be considered as rough approximations.



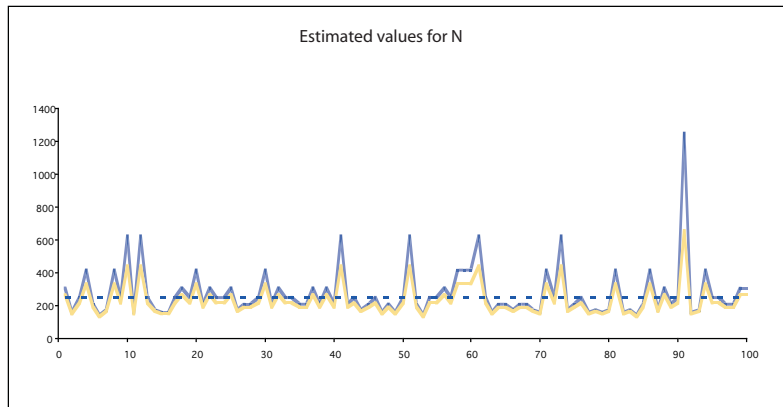


FIGURE 2. Population estimates based on \hat{N} (blue) and \tilde{N} (yellow) and 100 simulation iterations for a population with $N = 250$ (blue)

However, comparing the estimates with the true value of N (the blue line) in Figure 2, it appears both estimators produce estimates that are not normally distributed around the true value. This could complicate calculating confidence intervals. It is often desirable to construct confidence intervals for the estimates and, although we are not going to discuss them here, maximum likelihood methods or bootstrap approaches would be appropriate to use.

Does the Behavior of the Captured Individuals Affect the Estimations?

A basic assumption in abundance estimation using capture-recapture sampling is that the capture and recapture events are independent. One common problem that can lead to biased results when estimating the size of a population comes from what is known as “trap-happy” or “trap-shy” behavior. For example, some animals might tend to remain around the area where the captures take place because it has greener pastures or they become accustomed to human contact. Or, maybe after the first capture, they are drawn more to the bait in the trap than their natural food. These situations would imply that their probability of recapture would be higher than the probability of capture for the rest of the population.

For trap-shy behavior, after the first capture, the animal might learn to be cautious of the trap and the probability of recapture would go down. So, the animal becomes conditioned, and the capture and recapture events are no longer independent.

What would be the effect of trap-happy behavior in the estimated values of the size of the population? In such cases, the number of recaptured individuals is too high, and thus causes both estimators—to underestimate the size of the population.

To illustrate trap-happy behavior, imagine that, in our example, once a particular individual is marked, it has a probability of being captured greater than

1/250. That individual would be more likely to be recaptured compared to everyone else, and so its frequency of recapture would be higher, thereby inflating the recapture counts and lowering the population estimates. Although it would be difficult to know the new conditioned probability of recapture, we can simulate this situation to see its impact. Example results of the estimates obtained using Excel are represented in Figure 3.

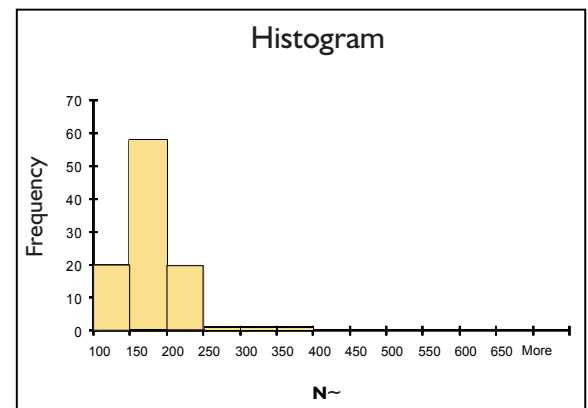
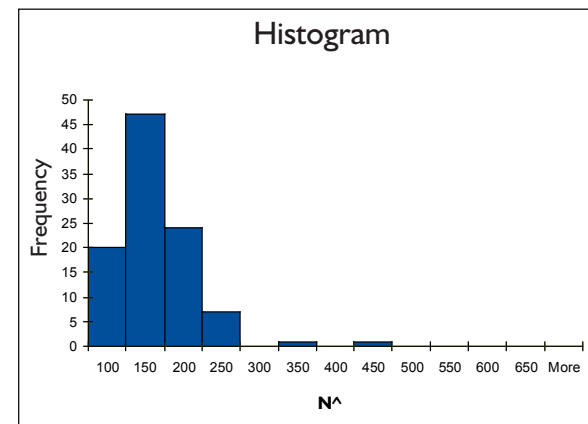


FIGURE 3. Population size estimates based on \hat{N} and \tilde{N} using simulation for a population with $N = 250$, but the individuals become trap-happy, so their individual probability of recapture increases

As predicted, most of the estimated values are lower than the true size of the population, $N = 250$.

We can perform a similar simulation in Excel to see the effect of trap-shy behavior on capture-recapture sampling in our population of 250. The population will be overestimated.

Final Thoughts

Using Excel is a good way to get your feet wet (so to speak) in learning about the mechanics of capture-recapture analysis. Try the Excel simulation we have looked at here and then experiment with the concepts we have discussed. Many questions can be asked, from how the

sample sizes affect the estimation of the size of the population to how we can improve the estimation process, or even how bootstrap methods can be applied in capture-recapture sampling. These are fun questions to address, and I will be happy to discuss them with you, so please contact me and we can talk about them.

Capture-recapture sampling can be used in many real-life situations. It has been used to not only estimate the abundance of animals, but also the number of minority individuals in human populations, such as the number of homeless people in a city. It has even been used to estimate the number of errors in large, complex software programs. It is an exciting subject, and the more you know about it, the greater the challenges you will be able to conquer. Watch for more articles on capture-recapture sampling in upcoming issues of *STATS*. ●

Answers from **Learning Stats Is FUN**

Answers to Riddles: ① Lucky Charms and Total; from the bi-cereal correlation

② ANOVA (a.k.a an egg)

Answers to Poisson Questions:

$\lambda = (0*299+1*211+2*93+3*35+4*7 + 7*1)/576 = .9323$; expected frequencies are 226.74, 211.39, 98.54, 30.62, 8.71; chi-square test statistic is 1.0176, whose one-tailed *p*-value (for 4 degrees of freedom) is .91—nowhere near significant!

Answers to Anagrams: mode, median, residual, range, bell curve, sample, time series, variance, outlier, correlation, experiments, kurtosis, least squares, confidence, logistic, binomial, Poisson, Pearson, Student's *t*, probability

Answers to Permutations

Questions: The word “statistics” has $10!/(2!3!3!) = 50,400$ permutations; the first five notes have only $5!/2! = 60$ permutations, but the first 11 notes have $11!/(2!2!3!) = 1,663,200$ permutations!

Graduate Certificate in Applied Statistical Strategies via Distance Education

This quality e-learning opportunity is available as a part-time program for working professionals. No residency required, but you must be admitted to the University of Tennessee. Semester-based – Summer (10 weeks), Fall & Spring (15 weeks)

THE UNIVERSITY of TENNESSEE **UT**

Visit www.anywhere.tennessee.edu/de or call 1-800-670-8657 for more information.

University of Northern Iowa **Do You Want to Use Mathematics in a World of Industry?**

A Professional Science Master's (PSM) degree from the University of Northern Iowa will help you succeed.

Developed in consultation with industry leaders, University of Northern Iowa's Professional Science Master's will help graduates excel in today's competitive climate.

The degree features:

- A mathematical focus on either continuous quality improvement or computing and modeling.
- Integrated business and experiential components.
- An internship or project on a problem of interest to a host industry.

For further information, including specific requirements and course information, visit www.uni.edu/math or contact Dr. Syed Kirmani at kirmani@math.uni.edu

PSM
PROFESSIONAL
SCIENCE MASTER'S

A Hands-On Capture-Recapture Study Without Going Outside

by Jonathan Chipman

Estimating the size of a population can be a fascinating statistical challenge. As Bruno de Sousa asks, “How many fish are in the sea?” We also could ask, “How many birds are in the air?” Or, “How many bears are in the woods?” The list is endless of populations that could be interesting to estimate. Capture-recapture analysis is an excellent way to estimate an unknown (and often unknowable) population size. But, the capture-recapture method seems almost too simple to be effective. Can it really be as straightforward as de Sousa describes?



JONATHAN CHIPMAN is a senior in statistics at Brigham Young University. He did an internship at the National Institutes of Health this past summer, and he plans to attend a graduate program in biostatistics next year.

Let’s see if we can simulate a capture-recapture study without the expense—and risk—of capturing, tagging, and recapturing fish, birds, or bears. We will use popcorn instead of wildlife. This study follows all the steps of a basic capture-recapture analysis, as follows:

- Drawing and marking an initial sample
- Releasing the marked individuals back into the environment
- Collecting additional independent samples
- Estimating the population size based on the proportion of recaptured individuals

Simulating Capture-Recapture

Our question of interest is how many ‘fish’ are in the ‘sea.’ For the simulation, we need a bowl to represent the sea and 1,000 popcorn kernels to represent the fish. Now, follow these steps:



STEP 1: Create the sea and fish.

To begin, place 1,000 popcorn kernels in a sealable bowl. These will be the fish in the sea. Set 50 colored kernels aside for the next step and seal the bowl.



STEP 2: Capture and tag

Shake the bowl so each fish has an equal probability of being caught. Then, open the bowl and remove a sample of 50 popcorn kernels. This is the capture sample. In a real study, we would place a mark or tag on these fish for identification in case they are caught again in later samples. For our simulation, replace the 50 regular popcorn kernels in the capture sample with 50 colored kernels and close the lid. The 50 colored kernels are the tagged fish.



STEP 3: Shake and sample again.

After tagging the fish, seal and shake the bowl again to allow the tagged fish to disperse back into the population. In a real study, we would wait for the tagged individuals to assimilate back into their environment. The species' normal movement patterns would determine how long to wait before sampling again. After shaking the bowl, take another sample. Of course, do not look into the bowl while sampling, and do not just sample from the top.



STEP 4: Record the fish recaptured.

Count the colored kernels in the second sample and calculate the proportion of recaptured fish in the sample. Knowing the proportion of fish in the second sample and the number of tagged fish we released back into the population, we can estimate the size of the population using the modified estimator de Sousa describes on Page 15.



STEP 5: Repeat, and repeat, and repeat.

Repeat the sampling process over and over to better estimate the true proportion of fish in the sea.



STEP 6: Estimate the fish population.

The last step is to calculate the grand mean of the estimates of the fish population from all the samples taken. We also could calculate a confidence interval around our estimate of the total population of fish in the sea.

TABLE 1. Results of 20 recapture samples showing the number of individuals recaptured in each sample, along with the population estimate based on each sample and the grand mean updated with each sample

RECAPTURE SAMPLE	NUMBER RECAPTURED	POPULATION ESTIMATE	GRAND MEAN
1	1	662	662
2	0	1,325	994
3	1	662	883
4	0	1,325	994
5	0	1,325	1,060
6	2	441	957
7	0	1,325	1,009
8	0	1,325	1,049
9	1	662	1,006
10	0	1,325	1,038
11	0	1,325	1,064
12	0	1,325	1,086
13	1	662	1,053
14	1	662	1,025
15	0	1,325	1,045
16	1	662	1,021
17	1	662	1,000
18	1	662	981
19	0	1,325	999
20	0	1,325	1,016

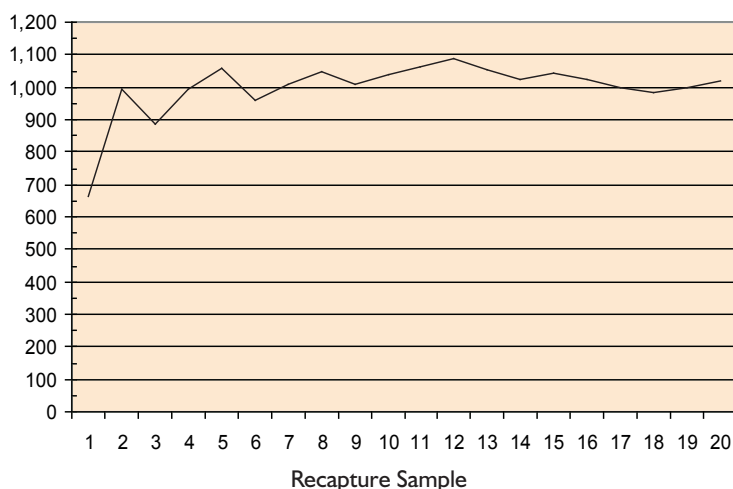


FIGURE 1. Population size estimated by the grand mean of a sequence of 20 recapture samples

Results

Table 1 shows the results of 20 recapture samples of size 25 when the population contains 1,000 individuals of which 50 were tagged. Figure 1 shows the grand mean of sequential samples converging to the actual size of the population.

Key Points

There are several key points. In this study, we simulated taking independent samples from a geographically bounded population of identical individuals that does not change in size during the study due to births, deaths, or migration.

In designing a capture-recapture study, we should ask ourselves the following questions:

Who exactly is in the population of interest?

Is the population restricted within geographic bounds?

Are some members of the population more difficult to catch than others?

Does the population size increase or decrease over time?

What steps are needed to ensure independent samples?

Although the objective of a capture-recapture study is often to estimate the size of a closed population, we could extend the principles of capture-recapture we have described here to estimate the size of an open population. I suggest studying *Handbook of Capture-Recapture Analysis* to learn more about both simple and advanced capture-recapture methods. (See the references section.)

Estimate How Many Fish Are in Your Sea

This simulation provides a hands-on way to explore more about capture-recapture analysis and shows how straightforward (and fun) it can be to estimate a population's size. Try the simulation yourself and see if your results are similar to those above. Then, think about how to simulate an open population with a birth or death rate. We would like to hear about your study and what you learned. Send your results to the *STATS* editor at pjfields@byu.edu. ●



SOLUTION

Perfect Correlation ... If Not for a Single **Outlier**

Over the past 30 years, I have shown the scatter plot included in this puzzle to many, many people. When I have done this, I have asked each one to write down his or her guess as to the value of r for the 101 data points. On each occasion, the vast majority of the guesses lie between +.90 and +.99. The modal guess, over the years, has been +.99.

The actual value of r for the 101 data points is -.42!

People who guess that the solitary point will not cause r to move very far away from +1.00 are underestimating the influence of a single data point that is an 'exception to the rule.'

Archimedes supposedly said he could move the world if you gave him a long enough lever and the freedom to stand far away from our celestial planet. In a similar fashion, give me the freedom to position a single outlier anywhere I wish and I can cause the r based on the other $n - 1$ data points to change from ± 1.00 to 0.00 when based on all n pairs of scores. Or, my single outlier can transform an r of 0.00 into a large r that suggests the X and Y variables have an exceedingly strong relationship.

To understand why this puzzle's 101 data points produce an r of -.42, consider this formula (of the

many that exist) for Pearson's sample correlation:

$$r = \frac{\sum z_x z_y}{n - 1}.$$

In this formula, the value $z_x z_y$ is the cross-product of the standard scores of X and Y for each data point. In other words, each person's contribution to the numerator is equal to

$$\left(\frac{x - \bar{x}}{s_x} \right) \left(\frac{y - \bar{y}}{s_y} \right)$$

where \bar{x} and \bar{y} are the means calculated for the two variables and s_x and s_y are the two sample standard deviations.

If the above formula for r does not provide full insight into why an outlier can have a big impact on r , take some time to apply this formula to this puzzle's data. You will discover two things: (1) the sum of the z-score cross-products for the 100 data points located in the lower right-hand portion of the scatter plot is equal to 28.5 and (2) the size of $z_x z_y$ for the single outlier is -70.1. These two components of the numerator for r show that the solitary point, by itself, had more influence on r than did all the other 100 data points combined. This occurred because that point was so far away from \bar{x} and \bar{y} . So, like Archimedes, you can apply enough leverage with a single influential point to 'move the world'—or at least change r to any value you choose. ●



Diogenes: The Sampler from Sinope

Diogenes' bathtub domicile and diet of onions are portrayed in the J.W. Waterhouse painting "Diogenes" (d.c. 320 BC) 1882 (oil on canvas).

Reprinted with permission from The Bridgeman Art Library International



by Chris Olsen

CHRIS OLSEN teaches mathematics and statistics at Thomas Jefferson 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.

While it may be true that what happens in Vegas, stays in Vegas, the same cannot be said for Sinope, Turkey—the birthplace of the Athenian philosopher Diogenes. At some point in his youth, the young Diogenes (along with his father) was implicated in some sort of counterfeiting scheme and subsequently found himself in Athens. There, he attached himself to the local cottage industry: philosophizing. Many stories about and descriptions of Diogenes survive. His bathtub domicile and diet of onions are portrayed in the J. W. Waterhouse painting “Diogenes.” Also in that painting is the item central to the most famous story of Diogenes: a lantern. As the story goes, Diogenes—apparently with no sense of irony—carried a lantern around the city during the day looking for an “honest man.”

Neither history nor legend records whether Diogenes actually found an honest man, and this suggests two statistical questions: Was Diogenes’ Athenian random transect method the best way to find honest men, and, if not, what sampling strategies might have been better? Diogenes’ cynicism would have suggested to him that even if honest men existed in Athens at the time, they would be extremely rare, and that could be a significant problem with or without a lantern.

Using a modern-day lantern of the fluorescent variety, I was able to find a copy of *Sampling Rare or Elusive Species: Concepts, Designs, and Techniques for Estimating Population Parameters (SRES)* by William Thompson to aid in answering my sampling questions. Early in the book, Thompson sharpens the concept of a “rare population” into three sorts



of circumstances. First, a population might be rare because there are not very many individuals in the population. Second, a population might be thought to be rare because individuals are clumped into a somewhat small part of a somewhat large region. Finally, the population might be comprised of secretive or nocturnal individuals.

There also is an overarching problem known as detectability. That is, if there is an individual there, will it be detected? As an example, consider the dugong, an herbivorous marine mammal that looks a lot like turbid water from a distance, according to Helene Marsh and Dennis Sinclair's description in "Correcting for Visibility Bias in Strip Transect Aerial Surveys of Aquatic Fauna." The distance in question is 137 meters, the standard altitude of the plane from which biologists look for the dugong. The dugong, for his or her part, only breaks the water's surface for one or two seconds—about five blinks of an inattentive biologist's eyes. Thus, the probability of detection is pretty low, even when the little fellows are there.

Might it be that honest Athenians, seeing themselves in a teeming population of less-than-honest men, would feel a certain peer pressure? If so, detecting honest men in the teeming streets of Athens could be a significant problem for Diogenes—they may blend into the camouflage of the crowd. In contrast, it does seem pretty likely that those less-than-honest folks (e.g., counterfeiters) might be located around Athens, cleverly disguised as honest folks so as to avoid the Athenian treasury agents, who, for all I know, were running a sting operation and were cleverly disguised as counterfeiters disguising themselves as honest folk.

Be that as it may, let's move on to other problems. If the population is actually rare in the sense of there not being many individuals, it would seem likely that an investigator could go for a long time without an individual actually presenting itself, let alone being detected. Only a man like Mr. Avoid-All-Pleasures Diogenes would tramp hither and yon across the Agora and up and down the Acropolis, in the daylight, with a lantern (i.e., fire), in the heat of the Athenian sun. The rest of the populace, honest or not, would no doubt be at their favorite shady spot quaffing ouzo or hemlock, depending on the circumstances.

This brings us to the other problems mentioned above. Might honest, ouzo-quaffing men clump together somewhere eluding detection? Perhaps in a "thoughtery," as conjured by Aristophanes' in "The

Clouds"? (See Paul Roche's new translations of the plays as cited in the reference section.) An example in *SRES* relates the methodology of a nocturnal snake study in the Snake River (well, duh!) Birds of Prey Area in southern Idaho in the mid-1970s. In year one of the study, the investigator used 'standard' techniques, such as turning over rocks and driving around at night, basically not finding any. In the second year, the investigator set up drift fences and funnel traps and detected lots of snakes. Change of method, change of detection probability.

Well, clearly, if honest men are nocturnal, Diogenes—the marauder of daylight sampling—is out of luck. However, it would seem that, if anything, the honest men would be out and about doing their honest stuff during the daylight hours, while the charlatans of the night would be out and about doing their less-than-honest stuff when they are less likely to be detected. (Perhaps they have read an early manuscript version of an Aristotelian work, now lost, on sampling rare species?) In any case, it looks as if Diogenes may have stumbled on a reasonable time of day to sample. However, let's consider this 'elusive' business. Diogenes' first problem is undoubtedly his diet of onions. If that does not make a population turn elusive, nothing will!

There may be an alternative to Diogenes actually being onsite. Perhaps Diogenes could set a trap for the honest men. As *SRES* relates, it just might be easier to detect signs of individuals, rather than the individuals, themselves. Animal tracks, for example, could be used, but this method has its own problems. Ken Gerow and his colleagues point out in their fascinating discussion, "Monitoring Tiger Prey Abundance in the Russian Far East," that it is not easy to tell how fresh the tracks are. (One hopes they are not as fresh as, say, 20 seconds.) *SRES* also mentions that detecting creatures such as possums can be done by putting out blocks of tasty (to possums) wax and inspecting bite marks. Aha! This sounds like the germ of an idea.

It seems to me that Diogenes could have formulated a sampling plan through the simple artifice of baiting a trap for honest men. He could have used what we would now call a "dropped-letter" technique, dropping little bags of coins (counterfeit, just in case) at different locations in Athens. The bag would be tagged "Property of absent-minded Diogenes. Please return to his bathtub at the corner of α Street and β Avenue." Then, he could just sit back in the sun waiting for the honest men to come to him.

Had Diogenes only thought of this, the sign over Plato's Academy entrance surely would have said "Let no man ignorant of Abundance Estimation enter these doors." ●

STAT•DOKU

The popular Sudoku puzzle uses a special 9 x 9 Latin square, where each of the entries come from the numbers one through nine with no number being repeated in any row, column, or in each of the 3 x 3, non-overlapping sub-squares. STAT•DOKU uses letters instead of numbers, and the nine letters spell a statistical term.

The following two puzzles each use one statistical word. The first word is “histogram” while the second is “deviation.” In the first word, no letter is repeated. In the second word, the letter “i” is repeated with all other letters used only once. Before solving each puzzle, note that somewhere in the puzzle, the word is spelled out in the proper order either from left to right or from top to bottom.

	G		I	A		T		R
R	T	H	S			O		A
A			H	R	T	G		M
G	S		O	I	A	R		H
T		R	G		M		O	
	I		T	H	R	M		S
				O			R	T
	R			G	S	I		
			R	T		S		G

STAT•DOKU PUZZLE 2

N			V	D		T	I	A
			T		A	N		O
			N		O		D	E
D		V				I	O	N
I	A	T		O			E	V
I	O	N				I	A	T
	T			N	I		V	
O			E					
	V	D	A	T	I		N	I

STAT•DOKU PUZZLE 3

After solving the puzzle, consider the following questions:

How does knowing that the letters will all be in a certain order somewhere in the puzzle affect how you solve the puzzle?

Does the constraint that the word must be spelled out readably somewhere in the puzzle affect how the letters should be distributed?

These particular puzzles were constructed using a pattern. Can you find the pattern?

Does having a repeated letter in the second puzzle make it easier or harder to solve than the first puzzle?

If we were to construct all possible STAT•DOKU arrangements of “histogram” and “deviation,” without the additional constraint of the word spelled correctly at some location in the puzzle, would there be more, the same, or fewer arrangements of “histogram” than “deviation”? Why or why not?

Editor’s Note:

Thank you to Amy White, a mathematics education major at Brigham Young University, for contributing the STAT•DOKU puzzles and questions for this issue.



WHITE

STATS readers Chad Bohn, Kate Tranbarger, David Unger, and Na Yang solved the STAT•DOKU puzzle in *STATS* Issue 47 and answered the questions we posed about solving a STAT•DOKU puzzle compared to solving a regular Sudoku puzzle. If you can solve the STAT•DOKU puzzles in this issue and answer the questions, send your answers to *STATS* Editor Paul J. Fields at pjfields@byu.edu. If you are the first to send in the correct answers, you will receive an ASA T-shirt. ●

References and Additional Reading List

The references for each article in this issue of *STATS* are included in the listing below, along with suggestions for additional reading on related topics. The page number for each article are the numbers in [blue](#).

3 Remembering Karl Pearson After 150 Years Read the paper containing Karl Pearson's binomial discovery, "Contributions to the Mathematical Theory of Evolution, II: Skew Variation in Homogeneous Material," *Philosophical Transactions of the Royal Society of London (A)*, 186: 343-414, 1895.

Stephen M. Stigler helps us understand the people, situations, and advances involved in the development of statistics in *The History of Statistics: The Measurement of Uncertainty Before 1900*, Harvard University Press, 1986. See pages 333-334 regarding Karl Pearson's binomial polygon.

5 Even More FUN Learning Stats!

Movies

Descriptions of the statistical movies, "Statistics," directed by Frank Robak, 2006; "The Statistician," directed by Tim Robertson, 2001; and "Statistically Speaking," directed by Nandi Bowe, 1995, can be found using the "Search" function at www.imdb.com.

Another film, titled "Statistically Speaking," directed by Cameron Hatch, 2004, can be found at generallyawesome.com/2004/statistically_speaking.html.

"The Passionate Statistician: Florence Nightingale," distributed by The Cambridge Educational Core Curriculum Video Libraries, 1995, can be found at www.films.com/id/9104/The_Passionate_Statistician_Florence_Nightingale.htm.

Books and Videos

Neil J. Salkind explains statistical ideas in humorous ways in *Statistics for People Who (Think They) Hate Statistics* (3rd ed.), Sage Publications, 2008.

PDQ Statistics, by Geoffrey R. Norman and David L. Streiner, B. C. Decker, 1986, provides another humorous approach to learning statistics.

A humorous, but educational, video series explaining statistics is produced by Standard Deviants, a division of Goldhil Entertainment. The videos can be ordered at www.standarddeviants.com.

For nonfiction, check out *Beat the Dealer: A Winning Strategy for the Game of Twenty-One*, by Edward O. Thorp, Vintage Books, 1966.

Fiction aficionados will enjoy Thomas Pynchon's *Gravity's Rainbow*, Viking Press, 1973.

D. O. Koehler comments on *Gravity's Rainbow* in "Mathematics and Literature," *Mathematics Magazine*, 55(2): 81-95, 1982.

To learn more about the Poisson distribution, see R. D. Clarke's article, "An Application of the Poisson Distribution," *Journal of the Institute of Actuaries*, 72: 481, 1946.

Are You Feeling Lucky?

The web site for Dave's Gourmet, Inc., the makers of "Lucky Nuts," is www.davesgourmet.com.

To correct any misconceptions you, or someone you know, may have about skewness, means, and medians, see Paul T. von Hippel's article, "Mean, Median, and Skew: Correcting a Textbook Rule," *Journal of Statistics Education*, 13(2), 2005,

accessible at www.amstat.org/publications/jse/v13n2/vonhippel.html.

An 'odd' book is *Odds of Virtually Everything*, by Heron House Editors, Kensington Publishing, 1981.

There is also *Odds: On Virtually Everything*, by Richard M. Scammon, Penguin Group (USA), 1980.

Songs

For lyrics and sound files for many of Larry Lesser's statistics songs, visit CAUSEweb (Consortium for the Advancement of Undergraduate Statistics Education) and the CAUSEweb fun resources collection at www.causeweb.org and www.causeweb.org/resources/fun/agreement.php, respectively. The CAUSEweb fun resources collection was recently updated with contributions from the winners of its first humor contest.

Karaoke resources can be found at www.karaoke.com and www.karaokewh.com.

Lynda Williams' tutorial for using PowerPoint to make sing-along karaoke songs is available at www.scientainment.com/karaoke.html.

Larry Lesser's web site is www.math.utep.edu/Faculty/lesser/Mathmusician.html.

9 Perfect Correlation ... If Not for a Single Outlier

To learn more about the Pearson product-moment correlation coefficient and many other statistical procedures refer to David J. Sheskin's *Handbook of Parametric and Nonparametric Statistical Procedures*, (3rd ed.), Chapman & Hall/CRC, 2004.

10 **Busting Statistical Myths** For more details to clear away statistical misconceptions, see Jessica Utts' books, *Seeing Through Statistics*, (2nd ed.), 1999, Brooks-Cole/Duxbury Press, and *Mind on Statistics*, (3rd ed.), 2004, Duxbury Press.

14 **How Many Fish in the Sea?** The basic capture-recapture concept is presented in J. M. Landwehr, J. Swift, and A. E. Watkins, *Exploring Surveys and Information from Samples*, New Jersey: Dale Seymour Publications, 1987.

Joe Duran and John Wiorkowski apply capture-recapture methodology in "Capture-Recapture Sampling for Estimating Software Error Content," *IEEE Transactions on Software Engineering*, Vol. SE-7, No. 1, 147-148, 1981.

Steven K. Thompson presents the modified estimator of population size in *Sampling*, John Wiley & Sons, 2002.

One of the pioneering papers on capture-recapture is by Douglas G. Chapman, "The Estimation of Biological Populations," *The Annals of Mathematical Statistics*, 25(1): 1-15, 1954.

20 **A Hands-On Capture-Recapture Study Without Going Outside** A comprehensive presentation of classic and new methods for capture-recapture is in Steven Amstrup, Bryan Manly Bryan, and Trent McDonald (Editors), *Handbook of Capture-Recapture Analysis*, Princeton University Press, 2005.

24 **Diogenes: The Sampler From Sinope** Steve T. Buckland, et al. provide an overview of surveying techniques for abundance

estimation in *Introduction to Distance Sampling: Estimating Abundance of Biological Populations*, Oxford University Press, 2001.

Ken Gerow, et al., "Monitoring Tiger Prey Abundance in the Russian Far East," Roxy Peck, et al. (editors) in *Statistics: A Guide to the Unknown*, 4th Edition, Duxbury Press, 2006.

Helene March and Dennis Sinclair, "Correcting for Visibility Bias in Strip Transect Aerial Surveys of Aquatic Fauna," *Journal of Wildlife Management*, 53:1017-1024.

Paul Roche (translator), *Aristophanes: The Complete Plays*, Penguin Group, 2005.

William L. Thompson, *Sampling Rare or Elusive Species: Concepts, Designs, and Techniques for Estimating Population Parameters*, Island Press, 2004. ●

SAS® Learning Edition 4.1:
Bringing the Power of Data Analytics to Individuals

As companies recognize the competitive advantage and strategic importance of business intelligence and analytics, the need for talented people capable of leveraging world-class business analytics has never been clearer.

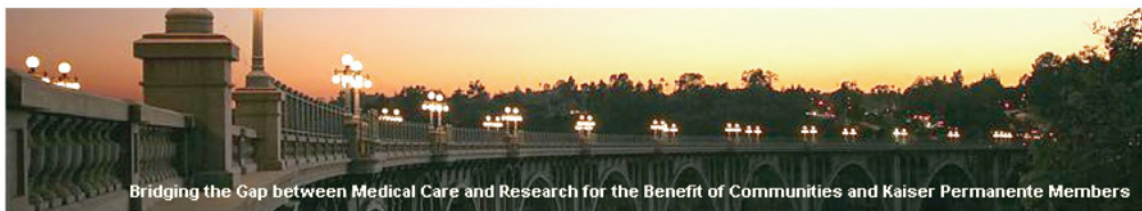
The SAS® Learning Edition educational bundle gives both students and knowledge workers, looking to enhance their SAS® skills, the freedom to learn at their own pace and on their own PCs or laptops—all on low-cost, self-install CD-ROMs.

Get started today and become part of the growing community of SAS users!

Visit: <http://support.sas.com/le>
Preset Die Date: **December 31, 2011**

sas
Publishing

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the U.S. and/or other countries. © indicates USA registration. Other brand and product names are trademarks of their respective companies. Copyright © 2007, SAS Institute Inc. All rights reserved. 46142.0907



Senior Biostatistician Researcher Position

Kaiser Permanente Southern California (KPSC) is searching for a doctorally-prepared Senior Biostatistician to join the Department of Research & Evaluation.

KPSC is a leading managed care organization that provides integrated care for approximately 3.2 million members of diverse race and ethnicity from Southern California. The integrated health care provided to these members is tracked through a paper unit record, which is in transition to a system-wide electronic health record. KPSC has excellent administrative information infrastructure that includes the tracking of the all diagnoses and procedures observed or undertaken in all medically attended visits for health plan members. This provides outstanding passive follow-up of important outcomes and is augmented by detailed pharmacy information.

KPSC is currently conducting a number of epidemiologic, clinical and health services research studies. These include cohort studies in a variety of disease areas that provide outstanding opportunities for applied and theoretical work in longitudinal data analysis. The funding for these studies is derived from Federal, industry and internal resources.

This hard-money position provides a core support package for the successful applicant that can be used to conduct pilot studies that leverage existing infrastructure for an extramurally funded program. This support includes staff for administrative tasks, programming and analytic staff and study assistants, as well as modest funding for non-personnel-related costs.


A description of the Department of Research & Evaluation is available on the web (<http://kp.org/research>). It is the home to 12 doctorally prepared investigators and over 70 support staff. The Department is located in Pasadena, a community of 134,000 residents and the home of the California Institute of Technology, the Rose Bowl, the Jet Propulsion Lab, and other historical and cultural sites. Information about the community can be found on-line at www.pasadenacal.com/visitors.htm. Pasadena is in the San Gabriel Valley 15 minutes north of downtown Los Angeles in sunny southern California.

Interested candidates should submit their CV online at jobs.kaiserpermanente.org (reference number WC.0601307). Inquiries may be directed to Dr. Steven Jacobsen (email: bianca.p.cheung@kp.org). Principals only.

KPSC is an Equal Opportunity/Affirmative Action Employer and offers competitive salary and comprehensive benefit packages.



Join *now*



Become a Student Member
of the **ASA** for only \$10

**Join the more than 4,000 students who
already know what ASA membership means...**

Free subscriptions to *STATS: The Magazine for Students of Statistics* and *Amstat News*, your monthly membership magazine!

Free online access to the *Journal of the American Statistical Association*, *The American Statistician*, and the *Journal of Business & Economic Statistics*.

ASA members-only discounts on ASA publications, meetings, and Continuing Education Courses, PLUS special discounts from publishers.

A network of professional colleagues made up of more than 18,000 ASA members, including 4,000 students.

Free or discounted dues for most regional chapters and special-interest sections.

Career opportunities and information through www.amstat.org, our JSM Career Placement Service, and *Amstat News*.

www.amstat.org/join

