Vol. 32, No. 2, 2019

CHANCE

Using Data to Advance Science, Education, and Society

Fair and Efficient by Chance

Including...

Demonstrating the Consequences of Quota Sampling in Graduate Student Admissions

Replicating Published Data Tables to Assess Sensitivity in Subsequent Analyses and Mapping

09332480(2019)32(2)

Taylor & Francis
Taylor & Francis Group

ASA

EXCLUSIVE BENEFITS
FOR ALL ASA MEMBERS!

SAVE 30% on Book Purchases with discount code **ASA18**.
Visit the new ASA Membership page to unlock savings on the latest books, access exclusive content and review our latest journal articles!

With a growing recognition of the importance of statistical reasoning across many different aspects of everyday life and in our data-rich world, the American Statistical Society and CRC Press have partnered to develop the **ASA-CRC Series on Statistical Reasoning in Science and Society**. This exciting book series features:

- Concepts presented while assuming minimal background in Mathematics and Statistics.
- A broad audience including professionals across many fields, the general public and courses in high schools and colleges.
- Topics include Statistics in wide-ranging aspects of professional and everyday life, including the media, science, health, society, politics, law, education, sports, finance, climate, and national security.

DATA VISUALIZATION
Charts, Maps, and Interactive Graphs

Robert Grant, BayersCamp

This book provides an introduction to the general principles of data visualization, with a focus on practical considerations for people who want to understand them or start making their own. It does not cover tools, which are varied and constantly changing, but focusses on the thought process of choosing the right format and design to best serve the data and the message.

September 2018 • 210 pp • Pb: 9781138707603: ~~$29.95~~ $23.96 • www.crcpress.com/9781138707603

VISUALIZING BASEBALL
Jim Albert, Bowling Green State University, Ohio, USA

A collection of graphs will be used to explore the game of baseball. Graphical displays are used to show how measures of batting and pitching performance have changed over time, to explore the career trajectories of players, to understand the importance of the pitch count, and to see the patterns of speed, movement, and location of different types of pitches.

August 2017 • 142 pp • Pb: 9781498782753: ~~$29.95~~ $23.96 • www.crcpress.com/9781498782753

ERRORS, BLUNDERS, AND LIES
How to Tell the Difference

David S. Salsburg, Emeritus, Yale University, New Haven, CT, USA

In this follow-up to the author's bestselling classic, "The Lady Tasting Tea", David Salsburg takes a fresh and insightful look at the history of statistical development by examining errors, blunders and outright lies in many different models taken from a variety of fields.

April 2017 • 154 pp • Pb: 9781498795784: ~~$29.95~~ $23.96 • www.crcpress.com/9781498795784

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
Vol 112, 2017

THE AMERICAN STATISTICIAN
Vol 72, 2018

STATISTICS AND PUBLIC POLICY
Vol 5, 2018

Taylor & Francis Group
an **informa** business

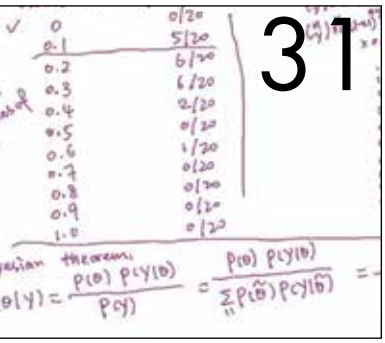
ASA

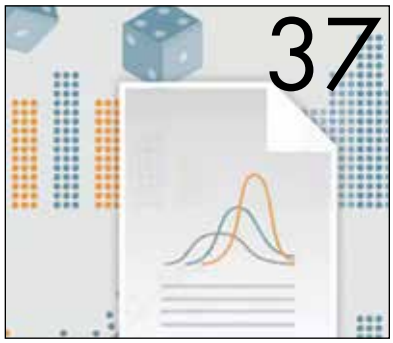
http://bit.ly/CRCASA2018

CHANCE

Using Data to Advance Science, Education, and Society

http://chance.amstat.org





ARTICLES

COLUMNS

DEPARTMENTS

Abstracted/indexed in Academic OneFile, Academic Search, ASFA, CSA/Proquest, Current Abstracts, Current Index to Statistics, Gale, Google Scholar, MathEDUC, Mathematical Reviews, OCLC, Summon by Serial Solutions, TOC Premier, Zentralblatt Math.

Cover image: Melissa Gotherman

WEBSITE

http://chance.amstat.org

AIMS AND SCOPE

CHANCE is designed for anyone who has an interest in using data to advance science, education, and society. *CHANCE* is a non-technical magazine highlighting applications that demonstrate sound statistical practice. *CHANCE* represents a cultural record of an evolving field, intended to entertain as well as inform.

SUBSCRIPTION INFORMATION

CHANCE (ISSN: 0933-2480) is co-published quarterly in February, April, September, and November for a total of four issues per year by the American Statistical Association, 732 North Washington Street, Alexandria, VA 22314, USA, and Taylor & Francis Group, LLC, 530 Walnut Street, Suite 850, Philadelphia, PA 19106, USA.

U.S. Postmaster: Please send address changes to *CHANCE*, Taylor & Francis Group, LLC, 530 Walnut Street, Suite 850, Philadelphia, PA 19106, USA.

ASA MEMBER SUBSCRIPTION RATES

ASA members who wish to subscribe to *CHANCE* should go to ASA Members Only, *www.amstat.org/membersonly* and select the "My Account" tab and then "Add a Publication." ASA members' publications period will correspond with their membership cycle.

SUBSCRIPTION OFFICES

USA/North America: Taylor & Francis Group, LLC, 530 Walnut Street, Suite 850, Philadelphia, PA 19106, USA. Telephone: 215-625-8900; Fax: 215-207-0050. UK/Europe: Taylor & Francis Customer Service, Sheepen Place, Colchester, Essex, CO3 3LP, United Kingdom. Telephone: +44-(0)-20-7017-5544; fax: +44-(0)-20-7017-5198.

For information and subscription rates please email s*ubscriptions@tandf.co.uk* or visit *www.tandfonline.com/pricing/journal/ucha*.

OFFICE OF PUBLICATION

American Statistical Association, 732 North Washington Street, Alexandria, VA 22314, USA. Telephone: 703-684-1221. Editorial Production: Megan Murphy, Communications Manager; Valerie Nirala, Publications Coordinator; Ruth E. Thaler-Carter, Copyeditor; Melissa Gotherman, Graphic Designer.
Taylor & Francis Group, LLC, 530 Walnut Street, Suite 850, Philadelphia, PA 19106, USA. Telephone: 215-625-8900; Fax: 215-207-0047.

RESPONSIBLE FOR ADVERTISEMENTS

Send inquiries and space reservations to: *advertising@taylorandfrancis.com*.
Printed in the United States on acid-free paper.

EDITOR'S LETTER

Scott Evans

Dear CHANCE *Colleagues,*

We celebrate the coming of spring with a new issue of *CHANCE*.

The United States has perhaps never been so entrenched in political gridlock. Negotiations over issues such as healthcare and immigration are at a stalemate. Might we look to a statistical concept to play a role in addressing these issues? Could randomness be part of the solution? **Leonard Wapner** discusses whether flipping a coin can help address the country's problems in "Fair and Efficient by Chance."

Graduate admission to STEM-related programs at universities in the United States often involves a quota system to ensure that some applicants from the United States are accepted when competing against a very talented international pool of applicants. **Paul Kvan** discusses the implications of this in "Demonstrating the Consequences of Quota Sampling in Student Admissions."

Statisticians understand that point estimates should be accompanied by standard errors or the precision associated with such estimates, but in many cases, tabular summaries are published without this information. **Tom Krenzke** and **JianZhu Li** describe an approach of "replicating tables" that allows sampling errors from a table based on the margins of error associated with numbers in the table, to carry into analyses that use this information. An R function and Excel file can be used to implement this approach.

Thank you to **Sarjinder Singh** for letting us have fun visualizing the connection between a box plot and the human face!

Be sure to reserve Saturday, September 28, 2019, for the next installment of the New England Symposium on Statistics in Sports (NESSIS), one of the preeminent research conferences on statistics in sport. NESSIS is the brainchild of **Mark Glickman** and yours truly (well—mostly Mark). For more information, visit *www.nessis.org*.

In our columns, **Monika Hu** shares a rewarding experience about teaching an upper-level undergraduate statistics course through a shared/hybrid model in "Taking a Chance in the Classroom." **Mine Cetinkaya-Rundel** revisits teaching massive open online courses (MOOCs) and what we have learned from them in "Teaching Statistics in the Health Sciences." **Nicole Lazar** discusses crowdsourcing in the Big Data era in "The Big Picture." **Christian Robert** reviews *The Beauty of Mathematics in Computer Science*, *IAQ*, *Let the Evidence Speak*, *Surprises in Probability—Seventeen Short Stories*, and *Is that a Big Number?* in our book reviews column. Finally, **Howard Wainer** examines the work of Devah Pager, who died recently at an all-too-young age. Howard discusses her work on the relative effects of race and crime on employment in "Visual Revelations."

Scott Evans

LETTER TO THE EDITOR

Another Look at Meta-analysis

David C. Hoaglin


Meta-analysis: Can We Extract Gold from the Biomedical Literature?

Deborah Dawson and Derek Blanchette ("Meta-analysis: Can We Extract Gold from the Biomedical Literature?" *CHANCE* 31(4):45–50) give an accessible overview of meta-analysis, its potential benefits, and some of its challenges. The metaphor in their title, however, suggests a need for vigilance. Many of the meta-analyses in the biomedical literature do less to enrich our understanding than their authors intend.

The difficulty lies in the statistical analyses, which are often not so straightforward or minor. It is tempting to think that sophisticated software, such as Comprehensive Meta-Analysis and the Cochrane Collaboration's Review Manager, will do the proper analysis with little need for intervention. Unfortunately, the statistical analysis often has traps for the unwary; it needs supervision by an experienced professional analyst who keeps up with the meta-analysis literature. Such a source of guidance can warn analysts of the gaps and problems in meta-analysis methods. Its absence is evident in numerous meta-analyses published in major journals.

For example, the customary fixed-effect approach estimates the overall effect by taking a weighted average of the study-level estimates with inverse-variance weights (i.e., the weight for each study's estimate is the reciprocal of the within-study variance of that estimate). Those variances are almost always unknown, and simply replacing them with their estimates can yield poor results (a shortcoming well-documented in the literature). The most-popular method for random-effects meta-analysis, in which the weights include an estimate of the between-study variance, has been described as producing biased estimates with falsely high precision, and its confidence intervals have below-nominal coverage and are inferior to those produced by another method. Analysts who, trustingly, use meta-analysis software without expert guidance generally receive no warning of these difficulties.

Formally testing for study-to-study heterogeneity has risks and little benefit. The standard test uses an incorrect null distribution. Even with the correct null distribution, [though], it is not appropriate to use the result for choosing between a fixed-effect analysis and a random-effects analysis.

It is often difficult to evaluate the quality of published meta-analyses, because reports seldom give adequate details of the analysis. In the interest of transparency and reproducibility, authors should follow the long-standing advice of the International Committee of Medical Journal Editors: Describe statistical methods with enough detail to enable a knowledgeable reader with access to the original data to verify the reported results.

Disappointingly, the PRISMA checklist does not contain a corresponding item.

I had not heard of Seymour Glass. Gene V. Glass is generally credited with introducing the term *meta-analysis*, in 1976.

In other ways that I have not mentioned, meta-analysis is currently in a less-than-pretty state. Analysts and readers need a well-informed appreciation of where the problems lie, so they can advocate for improvements and help meta-analysis reach its full potential. ◘

Further Reading

Cornell, J.E., Mulrow, C.D., Localio, R., et al. 2014. Random-effects meta-analysis of inconsistent effects: a time for change. *Annals of Internal Medicine* 160:267–70.

Glass, G.V. 1976. Primary, secondary, and meta-analysis. *Educational Researcher* 5(10):3–8.

Hoaglin, D.C. 2016. Misunderstandings about Q and "Cochran's Q Test" in meta-analysis. *Statistics in Medicine* 35:485–95.

IntHout, J., Ioannidis, J.P.A., and Borm, G.F. 2014. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Medical Research Methodology* 14:25.

Kulinskaya, E., Morgenthaler, S., and Staudte, R.G. 2014. Combining statistical evidence. *International Statistical Review* 82:214–42.

About the Author

David C. Hoaglin is an independent consultant in Sudbury, Massachusetts.

AUTHORS' RESPONSE

We very much appreciate the opportunity to correct our misstatement of Dr. Glass's given name; our thanks to Dr. Hoaglin for bringing it to our attention, and our abject apologies to Dr. Glass.

We also appreciate the many thoughtful comments contained in Dr. Hoaglin's letter. We would certainly agree that meta-analysis has potential shortcomings; clearly, the techniques of meta-analysis cannot reasonably be expected to overcome all the shortcomings of the existing scientific literature. We tried to carefully convey these ideas in our article while pointing out the potential advantages of meta-analysis. Meta-analysis is, after all, placed near or at the peak of most "evidence pyramids," denoting the recognition of its potential contributions to our understanding of the state of current science and, as noted by Berlin and Golub (2014), its potential as an important source of evidence to inform decisionmaking. And of course, sometimes the contribution of meta-analysis is its revelation of the deficiencies of the extant literature and the paucity of our current understanding.

We are unclear about whether there is some concern regarding our inclusion of explanations of fixed-effects models. Certainly, few experts would advocate choosing the fixed-effects model as a starting point, or based on the results of a heterogeneity assessment; rather, most advocate a measured consideration of the literature and the potential for heterogeneity (Borenstein, et al., 2009, 2010). We hope that we made these points clear. However, we believe it is important that the student of meta-analysis have an understanding of the underlying assumptions and conceptual framework of both the fixed and random effects models, because they are critical to the proper interpretation of results.

Further, fixed-effects models appear in the literature, and fixed-effects options are available in various software packages. For all of these reasons, we feel some understanding of these alternative approaches is useful. Dr. Hoaglin does point out a number of relevant methodologic concerns and suitable references for exploring them.

Berlin, J.A., and Golub, R.M., 2014. Meta-analysis as evidence: building a better pyramid, *Journal of the American Medical Association*, 312:603–606.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., and Rothstein, H.R, 2010. A basic introduction to fixed-effect and random-effects models for meta-analysis, *Research Synthesis Methods* 1:97–111.

Borenstein, M., Hedges, L., Higgins, J., and Rothstein, H. 2009. *Introduction to Meta-Analysis*. Chichester: Wiley & Sons, Ltd.

Deborah V. Dawson
Derek R. Blanchette

Fair and Efficient by Chance

Leonard M. Wapner

When some systems are stuck in a dangerous impasse, randomness and only randomness can unlock them and set them free.

—Nassim Nicholas Taleb

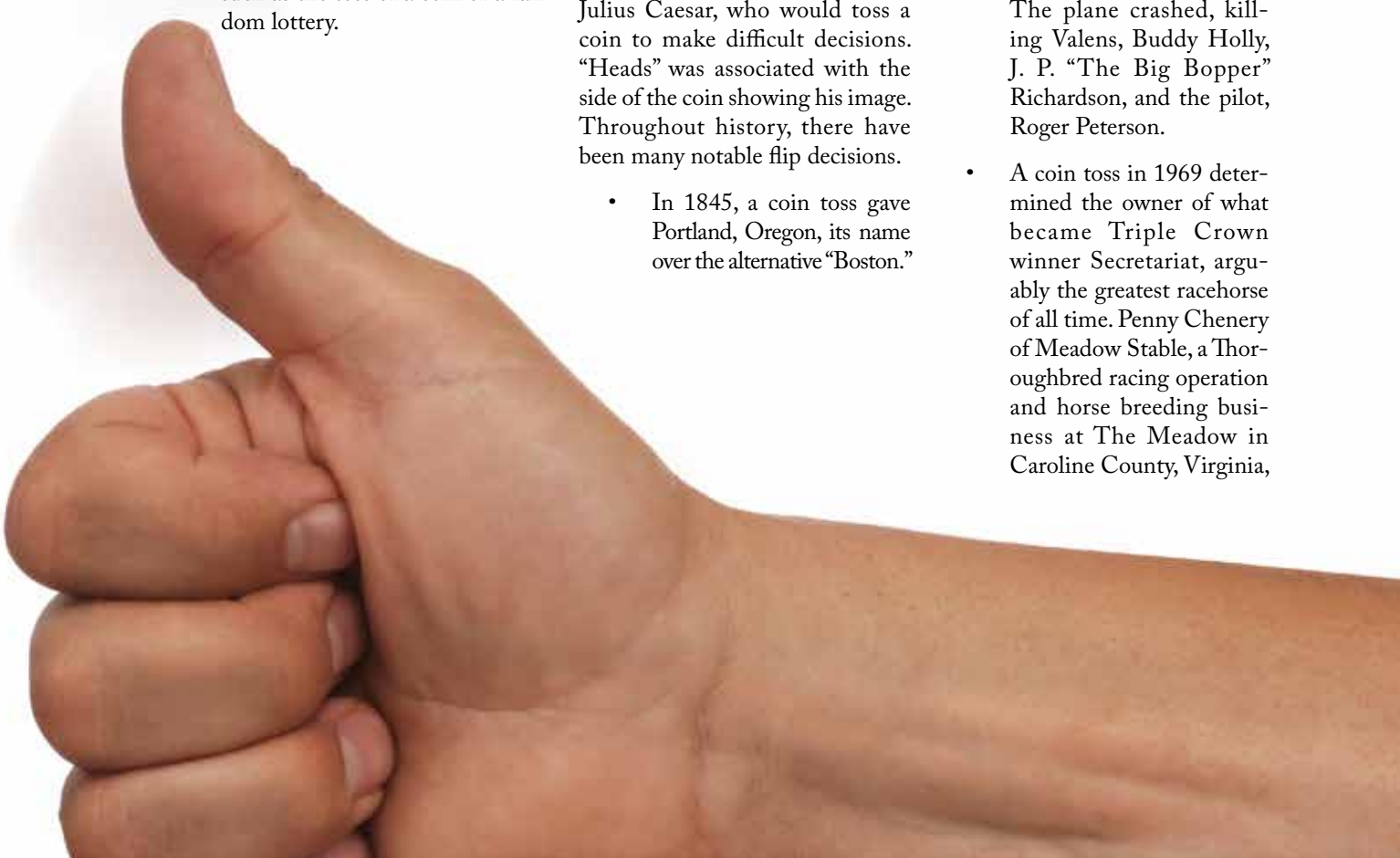
Negotiation deadlocks in business, politics, and social interactions can be costly in more ways than one. An arbitration scheme developed by Nobel Prize-winning mathematician John Nash provides a way to break budget impasses and similar negotiation deadlocks such as those leading to the U.S. government shutdowns in 2018–19. The scheme is elegant, fair, and efficient for all parties, providing a unique resolution. Surprisingly, it also relies on a randomizing device, such as the toss of a coin or a random lottery.

Flip Decisions

How do you celebrate June 1—U.S. National Flip a Coin Day? You don't, and neither does anyone else. But there is something to be said for "letting the dime decide" when people are unable to make a decision.

At an early age, children learn that "flipping for it" affords a quick, simple, and fair way to settle disputes. By letting chance decide, they can't blame themselves or the others for the outcome. The tradition is believed to date back to Julius Caesar, who would toss a coin to make difficult decisions. "Heads" was associated with the side of the coin showing his image. Throughout history, there have been many notable flip decisions.

- In 1845, a coin toss gave Portland, Oregon, its name over the alternative "Boston."
- Wilbur and Orville Wright tossed a coin in 1903 to determine which of the two would pilot the first flight. Wilbur won the toss, but crashed on his attempt. Orville later made the first successful flight.
- A coin toss on February 3, 1959 ("The Day the Music Died"), gave Latin star Ritchie Valens the last seat on the ill-fated flight from Mason City, Iowa, to Fargo, North Dakota. The plane crashed, killing Valens, Buddy Holly, J. P. "The Big Bopper" Richardson, and the pilot, Roger Peterson.
- A coin toss in 1969 determined the owner of what became Triple Crown winner Secretariat, arguably the greatest racehorse of all time. Penny Chenery of Meadow Stable, a Thoroughbred racing operation and horse breeding business at The Meadow in Caroline County, Virginia,



lost the toss, leaving her with the yet-unborn foal that she called Big Red but whose official name was Secretariat.

- In the Donald Duck episode "Flip Decision," Professor Batty converts Donald to the pseudo-philosophy of *flipism*, where all life's major decisions are made by the flip of a coin. The Great Society of Flippists proclaims, "Life is but a gamble! Let flipism chart your ramble!"

Coin tossing is almost always associated with an inability to decide caused by indifference or conflicting preferences. Beyond this, *randomization in the form of a coin toss or lottery may provide the only means of achieving a fair and efficient resolution as defined by Nash.*

The Potential Cost of a Budget Impasse

The U.S. Antideficiency Act, enacted in 1870 and amended over the years, is intended to prevent government expenditures in excess of available funds. As a consequence, a government shutdown occurs when Congress and the president fail to meet the associated deadline. Such an impasse occurred most recently in February 2019 when the U.S. Senate failed to get the required 60% of its members to pass a government funding law.

In the January 2018 shutdown, Democrats insisted the bill include provisions, including a path to citizenship, to protect roughly 2 million undocumented young immigrants brought to the United States illegally, known as DREAMers. United States President Donald Trump and Republican senators rejected this, favoring instead funding for stronger border protection, including construction of a wall on the United States/Mexico border.

These options were rejected by Democrats. At the time, Republicans controlled 51 of the 100 Senate seats. In a highly polarized atmosphere with senators voting close to party lines, a vote to advance the proposed bill split, falling far short of the 60% supermajority required to pass. A brief government shutdown followed. The deadlock was caused by legitimate differences combined with polarization and political brinksmanship, with both parties prepared to go over the cliff into a government shutdown if no agreement was reached.

We recognize this as classical "chicken," the theoretical dilemma game arising in the behavioral sciences. The 2019 shutdown created a similar impasse and associated shutdown.

Shutting down the government costs U.S. taxpayers billions of dollars each week the shutdown is in effect. Federal workers may be furloughed or have their pay delayed. Government services are curtailed. Additional costs are associated with lost productivity and in preparing for the possibility of future shutdowns.

Neither political party wants such an outcome, but such is the nature of brinkmanship, politics, and the game of chicken.

A Nash Model of a Simple Arbitration

Mathematically, resolving such a deadlock can be thought of as negotiating a settlement in a non-zero-sum game. This is best illustrated with a highly simplified model of such a negotiation and then applying the Nash arbitration scheme to arrive at a settlement.

Assume Democrats and Republicans are deadlocked over a negotiation involving three possible settlement outcomes. Democrats prefer outcome A: assistance for DREAMers, over outcome B: building a border wall. Republicans prefer outcome B over outcome A. Both parties prefer either outcome to outcome SQ: the status quo, which will certainly lead to a government shutdown if no agreement is reached.

In general, the SQ outcome to a negotiation need not be the least-favored option for all parties, but this would be the case for the 2018–19 U.S. budget negotiations modeled here, where a status quo deadlock is associated with a government shutdown.

To summarize:

A: Preferred by Democrats over options B and SQ

B: Preferred by Republicans over options A and SQ

SQ: Negotiations remain deadlocked and government shuts down

Ideally, a settlement should be both *fair* and *efficient*. A fair settlement is one without an *a priori* bias in reaching it. An efficient settlement is one for which no other settlement would be preferred by both parties. Nash provides a scheme that derives a unique solution, both fair and efficient.

The Nash scheme begins with each party assigning a numerical utility to each of the possible settlement outcomes. An ordered pair of the form (d, r) would then be associated with each outcome, where d and r denote the worth of the outcomes to Democrats and Republicans, respectively. If Democrats favor outcome A with coordinates (d_1, r_1) over outcome B with coordinates (d_2, r_2) and Republicans favor the opposite, it would follow that $d_1 > d_2$ and $r_1 < r_2$.

We may also assume that if Democrats are indifferent to payoff

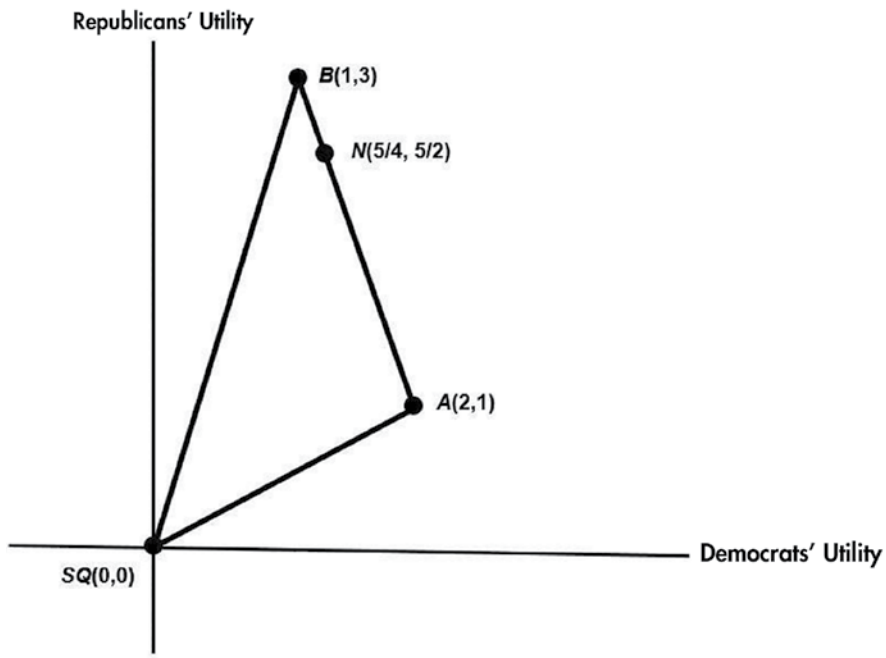


Figure 1. Payoff polygon.

(d, r) and a lottery with probability p of payoff (d_1, r_1) and probability $1 - p$ of payoff (d_2, r_2), then $d = pd_1 + (1 - p)d_2$. A similar assumption applies to Republicans. A possible agreement or resolution would be one of A, B, SQ, or a random lottery combination of the three.

The utility function used by a party to assign value to outcomes is to reflect the party's relative preferences for the outcomes. In this sense, the functions used and utilities assigned are not unique and fall into equivalence classes. For A, B, and SQ as above, say Democrats assign utilities 1, 0, –1 respectively, indicating the order of preference A > B > SQ as well as Democrats' indifference to outcome B and a lottery with equiprobable outcomes A and SQ.

However, assigning 5, 1, –3 to A, B, and SQ, respectively, conveys identical information. In general, two utility functions U_1 and U_2 are considered equivalent by a utility transformation if there constants $\alpha > 0$ and β such that $U_2 = \alpha U_1 + \beta$ exist. In this example, multiplying the first set of utilities by 4 ($\alpha = 4$) and then adding 1 ($\beta = 1$) yields the second set.

We could assume that 1, 0, –1 are utilities assigned by Democrats to A, B, and SQ, respectively, indicating Democrats' order of preference A > B > SQ. Say Republicans assign –1, 1, –2, respectively, indicating Republicans' order of preference B > A > SQ. We can characterize each of the three possible outcomes to the negotiation by the ordered pairs $A(1, -1)$, $B(0, 1)$, and $SQ(-1, -2)$. For convenience, transform all utilities by adding 1 to all Democrats' utilities and 2 to Republicans' utilities, translating all utilities so the status quo point SQ is now located at (0, 0), as shown in Figure 1. No information is lost with respect to the parties' relative preferences.

In addition to the three outcomes A, B, and SQ, additional *expected outcomes* can be achieved by a random lottery, with probabilities assigned to each actual outcome. For the preceding example, assume we assign to A, B, and SQ the probabilities 1/2, 1/3, and 1/6 respectively. One outcome is then selected by random lottery. The expected outcome (or payoff) to the parties would be $E = (1/2)(2, 1) + (1/3)(1, 3) + (1/6)(0, 0) = (4/3, 3/2)$, a point in the interior of the Figure 1 triangle. This triangle, including its interior is called the *payoff polygon*. It is defined as the smallest convex set containing the actual outcomes. It represents all outcomes graphically, including expected outcomes associated with random lotteries.

Nash's Arbitration Solution

Nash believes the fair and efficient agreement (solution, resolution), represented by a point in the payoff polygon, should satisfy these five conditions, or axioms. We call this the Nash point and denote it such as $N(d^*, r^*)$. From here one, we assume the status quo point SQ is (0, 0), since a transformation of the utilities can always achieve this result.

Axiom 1 (rationality): The Nash point $N(d^*, r^*)$ should be such that $d^* \geq 0$ and $r^* \geq 0$. That

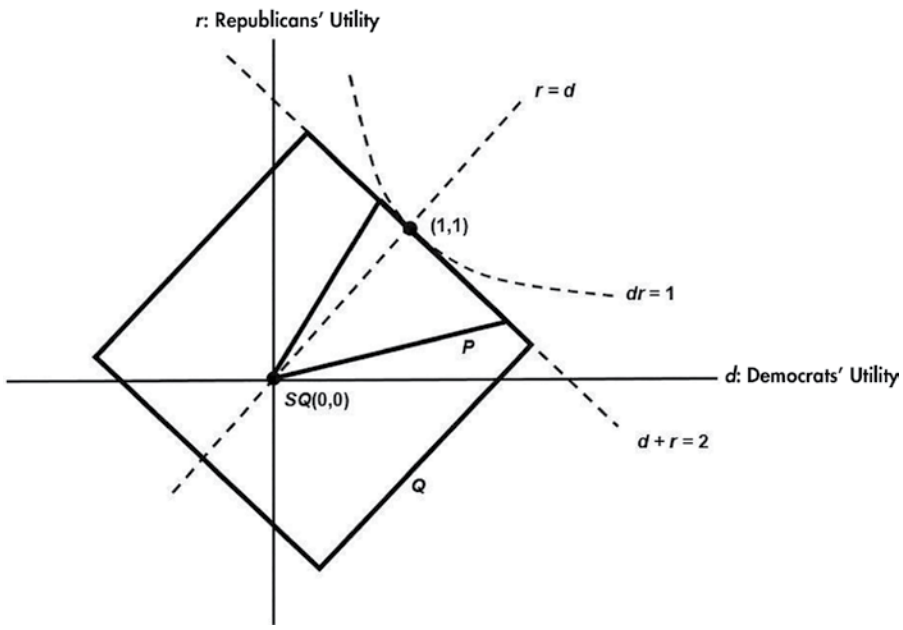


Figure 2. Proof of Nash's theorem.

is, neither party should accept a negotiated settlement less than what they would receive if negotiations were to fail and result in an impasse (SQ).

Axiom 2 (Pareto optimality): There should be no point (d, r) in the payoff polygon where $d > d^*$ and $r \geq r^*$ or $d \geq d^*$ and $r > r^*$. This guarantees the efficiency of N.

Axiom 3 (invariance under utility transformations): Consider payoff polygons P and Q related by a utility transformation. The Nash points of each should be related by the same utility transformation.

Axiom 4 (symmetry): If the payoff polygon is symmetric about the line $r = d$, then N should be on this line. This is a requirement of fairness, eliminating discrimination.

Axiom 5 (independence of irrelevant alternatives): Suppose that N is the Nash point solution for payoff polygon Q. Let P be a payoff polygon that is completely contained in Q which contains both $(0, 0)$ and N. Then N should also be a Nash point solution for P.

Theorem (Nash): One and only one point $N(d^*, r^*)$ in the payoff polygon satisfies the five axioms. It is given for each of three possible cases:

1. If there are no points in the payoff polygon where $d > 0$ or $r > 0$, let $N = SQ(0, 0)$.
2. If $d = 0$ and $r > 0$ throughout the payoff polygon let $N = (0, r_{max})$.

Similarly, if $d > 0$ and $r = 0$ throughout the payoff polygon let $N = (d_{max}, 0)$.

3. If there are points in the payoff polygon where both d and r are positive, let N be the point where the product dr is maximized.

Proof

The first two cases are self-evident. For Case 3, we note the payoff polygon is closed, bounded, and convex. That means there is a unique point (λ_1, λ_2) in the payoff polygon where dr is maximized. Ultimately, this point would be $N(d^*, r^*)$, the Nash solution point for Case 3, because it is the only point in the payoff polygon that satisfies all five axioms.

For Case 3, use Axiom 3 to change the utility scales for the parties so (λ_1, λ_2) is transformed to $(1, 1)$. This is done by multiplying d by $1/\lambda_1$ and r by $1/\lambda_2$. Let P denote the transformed polygon where dr is now maximized at $(1, 1)$. By Axiom 3 and the product maximizing nature of $(1, 1)$, polygon P lies entirely on or below the upper branch of the hyperbola $dr = 1$, as shown in Figure 2. The vertex of the hyperbola is $(1, 1)$ and the equation of the tangent line at this point is $d + r = 2$. Being convex, P must lie on or below this tangent line, for if not, there would be points in P such that $dr > 1 \cdot 1 = 1$.

Now enclose P in a large square Q with one edge of Q on the tangent line so Q is symmetric about the line $r = d$. See Figure 2.

According to Axioms 1, 2, and 4, $(1, 1)$ is the only point satisfying the Nash axioms for the large square region Q and must therefore be the unique Nash point solution for Q. Axiom 5 requires that $(1, 1)$ also be the Nash solution point for the embedded payoff polygon P. From Axiom 3

we conclude for the original payoff polygon that $N(d^*, r^*)$ is the point where dr is maximized. It is the unique, fair, and efficient solution because no other payoff point on or within the original payoff polygon satisfies the five Nash axioms. This completes the proof.

When applied to the example given (Figure 1), it is easily shown that dr is maximized at $d^* = 5/4$, $r^* = 5/2$ yielding the Nash solution point $N(5/4, 5/2)$. This point lays on line segment AB, three-fourths of the way from A to B, as shown in Figure 1. It is achievable by a lottery of outcomes A and B, having probabilities 1/4 and 3/4 respectively, or by tossing a biased coin where the coin's side associated with A has a probability of 1/4.

Limitations and Other Caveats

A technical objection relates to the seemingly innocuous Axiom 5: Independence of Irrelevant Alternatives (IIR).

> *While dining at your favorite restaurant, you see the menu offers beef, chicken, or fish as a main course. You decide on beef. When your waiter arrives, you are apologetically told that fish is not available this evening.*

Will beef remain your preference? Of course it will! You prefer beef to chicken, with fish being irrelevant.

Admittedly, one might argue that levels of aspiration and other psychological considerations can play a role in determining preferences. Under these circumstances, the addition or deletion of available options may indeed alter relative preferences.

Another theoretical limitation of the process involves that of either party providing false information about their utilities to finesse the outcome in their favor and corrupt the entire process. However, lying can produce an inferior (Pareto) outcome to the lying party, with no guarantee of improving the party's position. (A full discussion appears in Philip D. Straffin's *Game Theory and Strategy*.)

Lastly, a coin toss, lottery, or any form of random outcome generation may not reflect the will of the people when democratic processes are involved, such as U.S. government budget negotiations. Should the flip of a coin do the thinking for a matter of great consequence, such as adoption of the U.S. federal budget? In defense of randomizing, note that the mathematical expectation of the lottery to both parties exceeds that of the status quo. Why not choose by a lottery or coin toss and move on?

Sometimes an arbitrary decision is far better than no decision at all. Analogies can be made to sports, where the toss of a coin or the spin of a tennis racket is used to determine how the game will start. Close calls made by referees are effectively arbitrary due to visual limits. Jump balls in basketball and face-offs in hockey have virtually random outcomes, although some skill by the players is required. In all cases, players and fans understand the decision might go either way, but above all, a decision is required for the game to proceed.

Drawing lots is not new to politics, nor something that only happened in the past. In 2018, a lottery was held to determine the winner of a tied Virginia House of Delegates election. Virginians used a similar lottery to break an election tie in 1971.

Discussion

Can politics be mathematized? It would be naïve to suggest that the simple scheme of Nash resolves all negotiation deadlocks such as the 2018 U.S. government shutdown. Despite the noted objections, the Nash solution is compelling and should be considered. It is efficient and fair, costs almost nothing to implement, and eliminates the concern of a party having to back down on one or more of the issues being negotiated. In addition, there is a surprising additional psychological benefit to using a randomizing device to make decisions.

> *Faced with a tough decision? Flip a coin to decide. Even before the coin lands, during that brief moment, the coin is spinning in the air, your subconscious will tell you what you really want. When rationality fails, trust your heart.*

If Democrats and Republicans were to assign numerical utilities to their preferences and agree to accept an outcome as determined by chance lottery, both parties would achieve a better understanding of the relative importance of their objectives.

In an address to the nation on March 8, 1982, United States President Ronald Reagan urged Congress to "…get off the dime and adopt the deficit-reducing budget." Once off the dime, Congress might consider tossing it to move things along. ◘

Further Reading

Luce, R.D., and Raiffa, H. 1985. *Games and Decisions*, New York: Dover Publications, Inc.

Polgreen, P. 1992. Nash's arbitration scheme applied to a labor dispute. *UMAP Journal* 13: 25–35.

Straffin, P. 1993. *Game Theory and Strategy*, Washington, DC: Mathematical Association of America.

About the Author

Leonard Wapner has taught mathematics at El Camino College in Torrance, California, since 1973. He received his BA and MAT degrees in mathematics from the University of California, Los Angeles. He is the author of *The Pea and the Sun: A Mathematical Paradox* and *Unexpected Expectations: The Curiosities of a Mathematical Crystal Ball*. His writings about mathematics education have appeared in the *Mathematics Teacher* (National Council of Teachers of Mathematics) and *AMATYC Review* (American Mathematical Association of Two-Year Colleges).

Demonstrating the Consequences of Quota Sampling in Graduate Student Admissions

Paul Kvam

In graduate programs across the USA, admissions committees for science, technology, engineering, and math (STEM)-related programs are faced with an imbalance of qualified applicants due to a large majority of top student applications coming from outside the United States. As a consequence, admissions programs may apply a quota system to ensure that at least a minimum number of U.S. applicants are accepted into a graduate program.

As long as admission standards are strictly enforced for all applicants, this quota system will not elevate an unqualified applicant unfairly over a more-deserving student applicant from abroad. However, once students are enrolled in the graduate program, this lopsided admission procedure can create intellectual strata within a department, with international students academically outperforming U.S. students, on average.

This bias can be subtle, especially if the admissions data are unknown to the faculty and students. Suppose, for example, that 150 applications to a graduate program are determined to be "qualified," but 100 are from outside the USA. If the program accepts a small number of students from within the USA that is equal to the number accepted from outside the USA, it is probable that the international students will be more successful, on average, than the U.S. students.

This admissions problem can be seen in terms of quota sampling, and can demonstrate the inequality between the two groups using order statistics.

For examples featured in this article, we greatly simplify the admissions process to illustrate the deleterious effects of the biased sampling. Specifically, we reduce student outcomes down to a single random variable that refers to a final success score for the admitted applicant.

For clearness, we assume that while these success outcomes are not known until the student finishes the program, the admission committee has the ability to select students from the applicant pool according to how well they will score. That assumption is convenient albeit unrealistic, and it will help us make an important point about admission bias from quota sampling without addressing the obvious differences between a student's potential and actual success rate in a graduate program.

Why Is This a Problem?

Over the past 50 years, the demand for graduate education has grown considerably in most academic fields. The Council of Graduate Schools published an evaluation in 2010 that shows the increase in U.S. students applying to graduate programs is steady, but dwarfed by the explosion of applications from abroad, such as China, where these kinds of educational opportunities were relatively nonexistent a generation ago. In fact, the percentage of doctorates awarded to international students has doubled in the 30 years.

Our example is based on applications to a respected graduate program in the United States and uses the applicants' GRE scores as a stand-in to determine the subset of qualified applicants for the PhD program. Graduate admissions do not rely solely on GRE scores as an effective tool for finding the most-qualified applicants, and any savvy admissions committee would consider GRE results (especially for non-subject tests) only marginally.

Despite the overwhelming number of qualified applicants from abroad, many top graduate programs show a measurable preference toward U.S. applicants. Without creating official policies, many graduate programs in the USA use implicit quota systems to guarantee that a sufficient number of qualified U.S. students are enrolled in any given academic year.

In theory, this can work in a university's favor by bolstering some facets of diversity. If the policy

Suppose we are sampling from two distinct but equivalent populations: $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_m$, where all $n + m$ are independently generated from the same distribution $F(x)$. Suppose we select the largest r values from the first sample and the largest s values from the second sample. If $X_{r:n}$ and $X_{s:m}$ are order statistics from independent samples of sizes n and m from a continuous distribution F, then

$$P(X_{r:n} \leq X_{s:m}) = \sum_{i=r}^{n} \frac{s}{s+i} \frac{\binom{n}{i}\binom{m}{s}}{\binom{n+m}{i+s}}.$$

is misused, however, second-tier programs may resort to admitting unqualified U.S. applicants who are at high risk of failing in graduate classes.

A more-pernicious problem can occur if graduate programs exploit international applications to achieve gender diversity. For example, a graduate program may admit a higher proportion of female international students to achieve diversity goals, thus reducing opportunities for U.S. women to gain admittance.

The potential problems in applying quotas to graduate admissions can be argued based on fairness and the potential outcomes created by biased admission preferences, but we can also show direct effects on how quota systems affect student success rates using rules of probability. Our results can be illustrated with simple examples.

Quota Sampling

In contrast to random sampling, *quota sampling* requires that observations from specific subgroups are represented in the sample, whether or not they are sampled in proportion to their frequency in the population. Quota sampling is still a prevailing issue in many research fields, including business, medicine, and the social sciences.

In this problem, quota sampling requires a disproportionate amount of sample representation from the application pool of U.S. students. If we consider the 150 qualified students who applied in our previous example, we can model their performance outcomes as a random variable, governed by some distribution.

These outcomes can be treated as a set of independent observations, with large values indicating successful outcomes.

We assume each of the 150 applicants, no matter where they are from, has an equal chance of success in the graduate program. Although we cannot measure the student's program performance during the application process, it will help our illustration if we assume the top students from each group were selected without error according to how they were ranked during the admissions process. Rankings may be based on information from GRE test scores, letters from the applicant and references, school transcripts, and other helpful information from the applicant's vita.

From the 150 qualified applicants, suppose the quota sampling rule requires that at least three out of six students selected for admission are from the U.S. The performance scores of the U.S. students are denoted $X_1, X_2, \cdots, X_{50}$, which are independent and identical scores from some distribution $F(x)$. From this lot of 50, the admission committee will select the three highest scores, which can be represented as order statistics $X_{48:50} \leq X_{49:50} \leq X_{50:50}$. That is, $X_{k:n}$ represents the kth lowest score out of the group of n. We will assume $F(x)$ is a continuous function, so there are no tied scores. If we designate the sample of international student scores as $Y_1, X_2, \cdots, Y_{100}$, then the top students selected for the program are denoted $Y_{98:100} \leq Y_{99:100} \leq Y_{100:100}$.

While it might be clear how the top international student is superior to the second-best international student, it is not as easy to see how that student compares to the top U.S. student. Although it is not determined that $Y_{100:100}$ will be higher than $X_{50:50}$, we can determine how probable this event will be. For this case, it is relatively easy to explain: If the top student is randomly contained in a population of 150, then there is a 2/3 chance that student will be from the larger group of 100 international applicants, and only a 1/3 chance the student is from the 50 U.S. applicants, so $P(Y_{100:100} \geq X_{50:50}) = 2/3$.

Comparing other students between these two groups is not as simple. To show how students selected from the larger population have higher scores, on average, than the students from the smaller population, we need a more-general result for comparing order statistics.

Let's return to our example where three students from each applicant pool are admitted, but the international applicant pool (100) is twice as large as the U.S. applicant pool (50). Table 1 lists the probability the international student will outperform the U.S. student for every possible coupling of students from the two groups. Notice the asymmetry where the top international student is 96.5% likely to outperform the third-ranked U.S. student, but the top U.S. student is only 70.6% likely to outperform the third-ranked international student.

Graduate Admissions Data

The data we analyze are from a particular graduate program in the United States. The program and the

Table 1—Probability the Ranked International Student Performance is Better than the Ranked U.S. Student Performance)

	International #1	International #2	International #3
U.S. #1	0.667	0.443	0.294
U.S. #2	0.890	0.742	0.593
U.S. #3	0.965	0.892	0.793

Table 2—Probability the ith Best International Student Score ($X_{39-i+1:39}$) is Better than the jth Best Domestic Student Score ($Y_{8-j+1:8}$)

	$Y_{8:8}$	$Y_{7:8}$	$Y_{6:8}$	$Y_{5:8}$
$X_{39:39}$	0.8298	0.9741	0.9965	0.9996
$X_{38:39}$	0.6855	0.9292	0.9874	0.9982
$X_{37:39}$	0.5636	0.8710	0.9711	0.9950
$X_{36:39}$	0.4611	0.8043	0.9473	0.9892
$X_{35:39}$	0.3753	0.7328	0.9159	0.9800
$X_{34:39}$	0.3038	0.6595	0.8775	0.9669

year are concealed to help ensure the confidentiality of the school and the students who applied to this program.

In the selected year, 47 applicants from the hundreds of applications to this program achieved a minimum undergraduate GPA of 3.4, GRE verbal scores greater than 500 (this is the old scale for the graduate record exam, so the application year is before 2013), and GRE quantitative math scores over 760. For the purpose of this exploration, we will use these minimum thresholds to determine qualification for this PhD program. These artificial standards were not directly related to actual admissions decisions, so the GRE scores for any students who were selected are in no way identifiable from this analysis.

Of those final 47 students identified as qualified applicants, eight were U.S. citizens (the remaining 39 are labeled "international"). In a typical recruiting year, the program accepts eight to 12 applicants, with the constraint that at least 1/3 of them must be from the USA. This constraint was not strictly required, but followed as a suggestion of the school advisors. This demonstration chooses a final group of students consisting of the top six international students and the top four domestic students, according to their ranked scores within each subgroup.

From the group of $n = 39$ international students $(X_1, \cdots, X_{39})$, the best $r = 6$ students selected for admission are denoted using the order statistics $X_{34:39}, X_{35:39}, \cdots, X_{39:39}$. From the $m = 8$ U.S. students, the best $s = 4$ students are admitted: $Y_{5:8}$, $Y_{6:8}$, $Y_{7:8}$, $Y_{8:8}$. Then $X_{39:39}$ represents the top scoring student from the international group and $Y_{8:8}$ represents the top scoring student from the U.S. group. Because the U.S. students were accepted in disproportionate numbers, the distribution for the scores of the top four students (out of 8) is going to be different from the top six international students (out of 39). To further simplify, we will assume all 10 accepted applicants will choose to enroll (in actuality, 85% of students accepted into these programs will choose to enroll, on average).

Table 2 shows the probability a ranked international student outperforms a ranked U.S. student for every possible coupling. Although students chosen randomly from

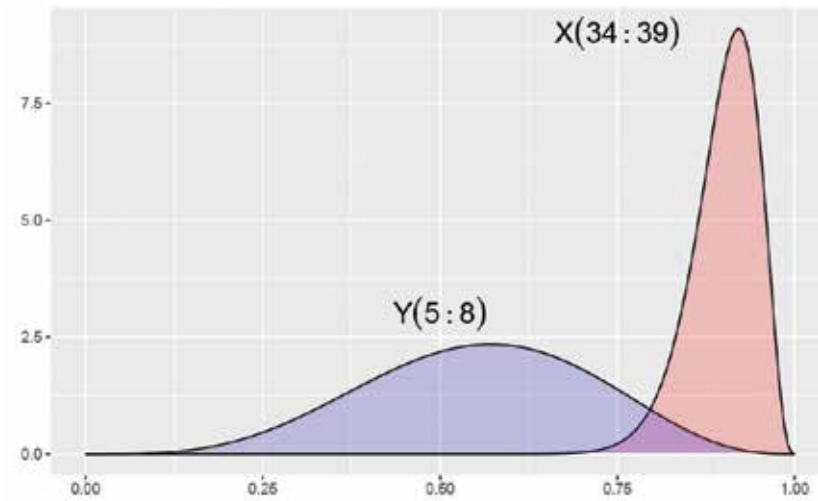


Figure 1. Student performance distributions for the lowest-ranked students in each of the two groups: $Y_{5:8}$ from the U.S. pool, and $X_{34:39}$ from the international pool.

the 47 qualifiers have the same performance distribution, those distributions are no longer identical once the students are ranked according to that measure. While any randomly chosen student from the eight U.S. qualifiers has a half-chance of outperforming any randomly chosen student from the 39 international qualifiers, the fourth-best U.S. student has only a 3.3% chance of performing better than the sixth-best international student, according to the table.

Figure 1 compares the performance density for the fourth-best U.S. student (blue) and the fourth-best international student (red). In this case, we modeled student performance with the Uniform distribution, so $F(x) = x$, where $0 < x < 1$, and the distribution of any order statistic $X_{k:n}$ has a beta distribution with parameters $\alpha = k$ and $\beta = n-k+1$. Because there are so many more international applicants, the expected performance of the fourth-best international student far exceeds the expected performance for the fourth-best U.S. student. The probability a randomly selected observation from the blue population is larger than a randomly selected observation from the red population is 0.0331.

Effects of Bias on Student Attrition

If all accepted students are sure to succeed in the program and complete graduation, the ramifications of the bias demonstrated in this article will be dampened, but if the attrition rate for the program is significant, the effects of the inequality will be magnified. The Council for Graduate Schools PhD Completion Project shows that attrition rates vary dramatically, depending on factors such as the school and the field of study, but a 50% attrition rate for a PhD program is not unexpected.

For example, suppose that, in general, 10% of students who are qualified for the program featured in our example and take the required classes are expected to eventually earn the PhD. Because the 10 accepted students are the top applicants from their pool, we can expect more than 10% of them will graduate. In fact, we can show that from these 10 selected students, we would expect half of them to eventually graduate.

However, for the four U.S. students, we can easily show that the probability of success is much lower compared to the international students. Figure 2 shows the probability that each of the eight ranked students score in the top decile (and thus graduate), and only one of the four US.. students has a graduation probability of over 0.5. Of the top five students in the program, four are international students.

This highlights a serious quandary for admissions committees that are inclined to admit a higher proportion of U.S. students than is reflected in the application pool—but these results also provide a fresh rationalization for why international students might be outperforming U.S. students in their courses. Typical excuses

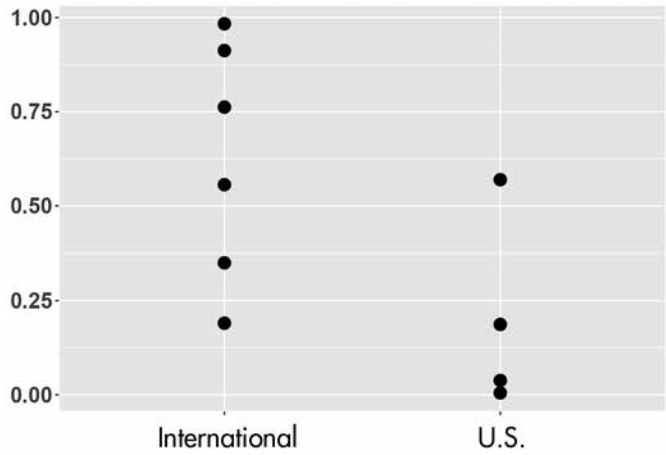


Figure 2. Student graduation probability for U.S. and international students based on the assumption that 10% of qualified students will graduate, on average.

might include the opinion that some international students come into the program better-prepared, or that some U.S. students do not spend as much time studying class material.

Without discounting them, these possible reasons for a performance disparity do not have to be necessary to explain the differences in academic outcomes. In fact, they do not have to be true. The bias created by quota sampling may represent a more-germane explanation, and we have demonstrated a way it can be characterized and measured based on enrollment data available to the admissions committee.

Discussion

Quota sampling is an implicit tool used in graduate school admissions to ensure that students from the USA maintain a certain level of representation in their graduate programs. This preference in admissions may help maintain aspects of diversity while also helping public schools show state legislators that the school considers education of its citizens as a primary goal, but this study shows that even a subtle bias toward U.S. students can create an adverse hierarchy in which international students routinely outperform their U.S. counterparts.

In programs that have an ample number of qualified applicants, this problem may be assuaged if graduation rates for both populations are sufficiently high. For programs with a smaller number of applicants, though, performance measures may manifest effects of bias on student attrition. This article provides a simple model that assumes that top students are accepted into the program and they all enroll at the same rate. The effects for quota sampling on more complicated and realistic admissions problems could be addressed with a more-empirical approach if sufficient data are available, but this order statistics model provides a foundation to quantify its effects.

We provide a simple program to assist researchers who want to consider specific admission problems of interest. Users can compute probabilities to compare potential student performances by entering their own admissions data into the author's code, which is written in the R programming language. Along with matrix results similar to Tables 1 and 2, users can generate a discrete contour graph to show how probabilities change as a function of student rank.

As an example, Figure 3 shows a sequence of these graphs for a hypothetical admission in which four international students and four U.S. students will be accepted into a graduate program.

The six graphs represent how U.S. students fare against international students as the proportion of international students admitted into the program increases. In each case, there are 500 applications from international students and a decreasing number of applications from U.S. students.

Specifically, the graph shows how student comparisons changes as the number of U.S. students changes from 500 (50%), 300 (62%), 200 (71%), 150 (77%), 100

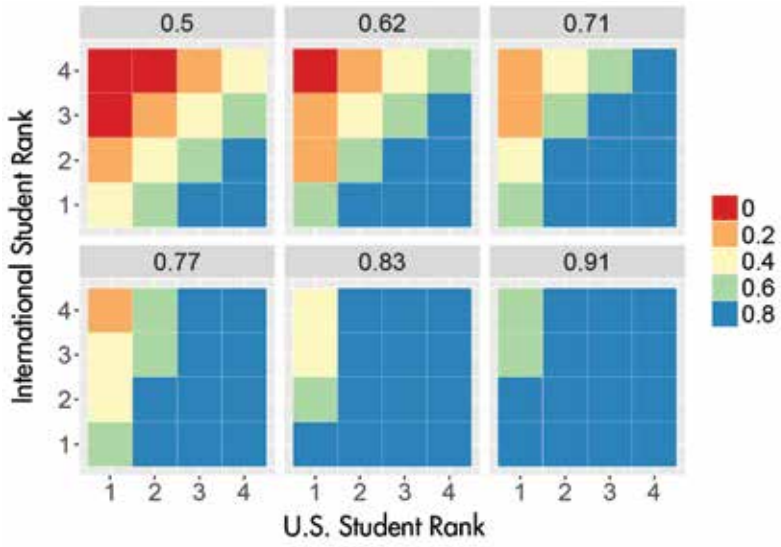


Figure 3. Contours display the probability any particularly ranked international student outperforms a ranked U.S. student: red (0 to 0.20), orange (0.2 to 0.4), yellow (0.4 to 0.6), green (0.6 to 0.8), and blue (0.8 to 1.0). The proportion of international students increases from 50% (upper left) to 91% (lower right) in each plot.

(83%), and finally 50 (91%) in the lower right corner. The blue squares indicate where the international student has between an 80% and 100% chance of outperforming the U.S. student. This dominance is absent in the first graph (where there are 500 applicants of both populations), but is plainly evident as the number of U.S. applicants decreases.

In the final plot, on the lower right, there are 500 international applicants and only 50 U.S. applicants (so the proportion is 50/550 = 0.91) and except for the top-ranked U.S. student, the international students have over an 80% chance of outperforming all of the other U.S. students.

In that scenario, the top U.S. student has a less than 40% chance of outperforming the lowest- and the second-lowest-ranked international students.

About the Author

Paul Kvam is a professor of statistics at the University of Richmond in the Department of Mathematics and Computer Science. He is a Fellow of the American Statistical Association with more than 70 peer-reviewed publications. He was previously at Georgia Institute of Technology (1995–2014) and Los Alamos National Laboratory (1991–1995). He received his PhD in statistics from the University of California, Davis in 1991.

Further Reading

Arcones, M., Kvam, P.H., and Samaniego, F.J. 2002. Nonparametric estimation of a distribution subject to stochastic precedence. *Journal of the American Statistical Association* 97:170–182.

Attiyeh, G., and Attiyeh, R. 1997. Testing for Bias in Graduate School Admissions. *Journal of Human Resources* 32:524–48.

Council of Graduate Schools. 2010. *PhD Completion and Attrition: Policies and Practices to Promote Student Success*. Washington, DC: Council of Graduate Schools.

Özturgut, O. 2011. Standardized testing in the case of China and the lessons to be learned for the U.S. *Journal of International Education Research* 7.

Posselt, J.R. 2016. *Inside Graduate Admissions: Merit, Diversity, and Faculty Gatekeeping*. Cambridge, MA: Harvard University Press.

Replicating Published Data Tables to Assess Sensitivity in Subsequent Analyses and Mapping

Tom Krenzke and Jianzhu Li

Estimates in published tabular results from surveys sometimes display, and should display, an associated standard error (measure of variation due to sampling). Using the published point estimates in a subsequent analysis should bring with it the associated variation that propagates through so the user knows how much to trust the subsequent analysis results. Failure to do so can mislead researchers and policymakers due to the inaccurate precision estimates in the results.

For a couple of examples, estimates and margins of error (MOEs) are produced in a Special Tabulation on Aging from the American Community Survey (ACS) data by the U.S. Census Bureau for the Administration for Community Living, and in the Census Transportation and Planning Products (CTPP) that is produced for the American Association of State Highway Officials. In each case, the users do not have access to the underlying data set and cannot readily reflect the MOEs in subsequent analyses. The MOEs in published static tables may capture different sources of error, such as from sampling, modeling, imputation, or perturbation, and therefore contain good information about the precision of the estimates. The MOEs should be taken into account in subsequent use of the estimates.

Sensitivity Analysis

A sensitivity analysis can provide a sense of robustness in analytic models due to fluctuations in the input data values or assumptions. Lehana Thabane and co-authors provided a thorough overview in 2013 of sensitivity analyses as applied to clinical trials, including ways to report on sensitivity analysis results. They described a wide range of types of sensitivity analyses, including study of the impact from outliers, missing data, clustering, and distributional assumptions.

A sensitivity analysis can provide credibility to a model result by showing the model outcomes stemming from strategically different sets of input data with varying values. Here, we make use of MOEs shown in static tables in subsequent analysis through generating "replicated tables." The right amount of variation in the replicated tables must be generated to align with the reported sampling error associated with the cell estimates in the static table. The replicated tables can then be used as a diagnostic tool that provides a sensitivity assessment, or a simulation study, that takes into account the impact of the sampling variance component. Graphs, such as maps, scaled bar chart, bar charts, and pie charts, can be produced to get a picture of the variation of the results across the replicated tables.

REPLICATED TABLES HIGHLIGHTS

- Allowing sources of error to propagate into subsequent analyses can provide results closer to the true overall picture of a situation.
- Practical statistical methods can replicate original data tables to provide a usable reflection of the variation reported in the original data table.
- Sensitivity analysis can provide an overall picture of the possible results that could occur.

Distinction from Variance Replicated Tables from the American Community Survey

The U.S. Census Bureau has published Variance Replicate Tables for select tables from the American Communities Survey (ACS) five-year (2011–15) so that users can combine subgroups and arrive at the exact MOEs. The successive

differences replication (SDR) variance estimation methodology is used to derive the MOEs in ACS tables. The SDR variance is calculated using the ACS estimate and the 80 variance replicate estimates.

The variance is the sum of the squared differences between the ACS estimate and each of the 80 variance replicate estimates, multiplied by 4/80. The MOE is calculated by multiplying the standard error (the square root of the variance) by the factor 1.645, which is associated with a 90% confidence level.

In addition to the original published MOEs, 80 variance replicate estimates are provided for a select set of tables. Although this set of tables includes a wide range of topics and is helpful to users, it does not cover all tables, including special tabulations. Applying the ACS Variance Replicate Tables *directly* in subsequent analysis is a different application than it is intended for, and does not give correct variance estimates. Using the replicate tables directly means conducting a sensitivity analysis by using a number of the replicated tables as input in a model to see the impact of sampling error in the model results.

The ACS Variance Replicate Tables would yield results with not enough variation. This can be seen by the factor used in the SDR variance estimator of 4/80. That is, using the ACS Replicated Tables directly in the sensitivity assessment is conceptually applying a factor of 1/80, not 4/80, when reviewing the results. Therefore, for a sensitivity assessment on a subsequent analysis, the variation seen in the results for this purpose is not enough because there should be four times more variation.

Methods to Generate Replicated Tables for Sensitivity Analysis

A goal of using the replicated tables for sensitivity analysis is to arrive at a set of M tables so the variation across the tables is approximately the same as the MOE in the single published table. The challenge is to incorporate the correlation between the cell estimates—the cell estimates are not independent of one another. The Dirichlet distribution is useful for simulating contingency tables since it takes the relationship between the table cells into account.

For example, the Dirichlet distribution was useful in synthesizing results for the Longitudinal Employment Household Dynamics (LEHD) OnTheMap system. Suppose that the static table has K interior table cells, with weighted cell frequency X_k, and associated MOEs. The cell proportion in cell k can be expressed as $p_k = \frac{X_k}{X}$, where we assume the overall weighted frequency X for the table follows a normal distribution, $N(\mu,\sigma^2)$.

Assume that the cell proportions $(p_1,p_2,\ldots p_k,\ldots,p_K)$ are the realizations of a set of random variables that follow a Dirichlet distribution model with parameters $(\alpha_1,\alpha_2,\ldots \alpha_k,\ldots,\alpha_K)$. The means and variances of the cell proportions are $E(p_k) = \frac{\alpha_k}{\alpha_0}$ and $Var(p_k) = \frac{\alpha_k(\alpha_0-\alpha_k)}{\alpha_0{}^2(\alpha_0+1)}$, where $\alpha_0 = \alpha_1 + \alpha_2 + \cdots + \alpha_K$. A basic algorithm for the replicated tables approach is as follows.

First, randomly draw the overall weighted frequency, X, from a normal distribution. Next, randomly draw a set of cell proportions, p_k's, from the Dirichlet distribution. Then derive the cell counts of a replicated table as $X_k = Xp_k$. Repeat the above steps M times to generate M replicated tables. After producing M replicated tables, the resulting variation among the replicated tables is checked.

Generating the replicated requires setting up the parameters of the normal distribution and the Dirichlet distribution. The parameters of the normal distribution μ and σ^2 can be estimated by the overall weighted total and its associated variance (MOE^2 divided by 1.645^2 under a 0.10 significance level) for a published table. Two methods are proposed to estimate the parameters $(\alpha_1,\alpha_2,\ldots\alpha_k,\ldots,\alpha_K)$ of the Dirichlet distribution: the generalized variance function (GVF) method and the distance function method.

GVF Method (Method 1)

GVFs have been used in large national surveys mainly as a way to easily estimate the variance associated with a resulting point estimate. Static tables are reduced to publishing the point estimate, and then publishing a model for users to compute the MOEs. Using a GVF to compute the MOE is simple once the model has been established. For example, a subset of estimated weighted frequencies for each subgroup or table cell and the associated directly estimated relative variances ($V^2_{X_k}$) are selected.

The GVF is a curve of the form $V^2_{X_k} = a + \frac{b}{X_k}$, where a and b are parameters to be estimated by an iterative weighted least squares process. The parameters (a and b) are published so the user can estimate the relative variance, which results in an estimated MOE for a 90 percent confidence interval through algebra $(MOE = 1.645 \times \sqrt{aX_k^2 + bX_k})$. The linear regression provides a smoothed set of variances, which is largely considered as an approach to stabilize the variance estimates. For a proportion, $p_k = \frac{X_k}{X}$, where X represents an estimate for a

certain subpopulation, the estimated relative variance is approximated as $V_{pk}^2 \approx V_{X_k}^2 - V_X^2$.

This leads to $var(p_k) = p_k^2\left(\frac{var(X_k)}{X_k^2} - \frac{var(X)}{X^2}\right)$. A special case results when using the same GVF function for both X_g and X, which results in $var(p_k) = \frac{b}{X} p_k(1-p_k)$. Using this formula assumes independence between the numerator X_k and the denominator X. The parameters of the Dirichlet distribution can be estimated from the observed cell proportions $(p_1, p_2, \ldots p_k, \ldots, p_K)$. Set $\hat{\alpha}_k = \left(\frac{X}{b} - 1\right) p_k$. As a result, $var(p_k) = \frac{\hat{\alpha}_k(\hat{\alpha}_0 - \hat{\alpha}_k)}{\hat{\alpha}_0{}^2(\hat{\alpha}_0+1)} = \frac{b}{X} p_k(1-p_k)$, where $\hat{\alpha}_0 = \hat{\alpha}_1 + \cdots + \hat{\alpha}_K$. This result is identical to the formula that uses the GVF functions to estimate the variance of a cell proportion mentioned above. If the GVF-based variances are not good approximations of the observed variances, this method will not perform well, since the variation among the cell estimates in the replicated tables will reflect the variances computed from the GVF method, not the variances in the published table.

Distance Function Method (Method 2)

As an alternative to the GVF method, the parameters of the Dirichlet distribution can be estimated through a distance function method so the variation among the replicated tables may be a better approximation of the published variances. The goal is to find a factor f that minimizes the distance function $\sum_k \left(\frac{\hat{\alpha}_k(\hat{\alpha}_0 - \hat{\alpha}_k)}{\hat{\alpha}_0{}^2(\hat{\alpha}_0+1)} - var(p_k)\right)^2$ where $\hat{\alpha}_k = f p_k$ and $var(p_k)$ is the variance of a cell proportion for cell k computed from the method given by Kirk Wolter's 1985 book *Variance Estimation*: $var(p_k) = p_k{}^2\left(\frac{var(X_k)}{X_k{}^2} - \frac{var(X)}{X^2}\right)$, where $var(X_k)$ and $var(X)$ are the published variances ($MOE^2/1.645^2$).

The GVF-based replicated tables method is a special case of the distance function method where $f = \frac{X}{b} - 1$, and the $E(p_k) = \frac{\alpha_k}{\alpha_0}$ holds for each approach. The distance function method may improve the GVF method in the sense that it finds an f that minimizes the distance between the variances from the Dirichlet distribution, $\frac{\hat{\alpha}_k(\hat{\alpha}_0 - \hat{\alpha}_k)}{\hat{\alpha}_0{}^2(\hat{\alpha}_0+1)}$, and the variances of the cell proportions derived from the observed variances. If the GVF-based variances are very close to the observed variances, the Dirichlet parameters estimated from the GVF method and the distance function method would be very similar.

Evaluation Results

The replicated table methodology was evaluated with CTPP data. The CTPP tables are produced to meet the needs of transportation planners in understanding local journey-to-work patterns. The tables relate worker and household characteristics to travel mode based on the worker's residence, workplace, and travel from residence to workplace.

The residence-based, workplace-based, and residence-to-workplace flow tables involve dozens of variables and provide cell aggregates, means, medians, and estimated MOEs for small geographic units such as census tracts and Traffic Analysis Zones (TAZs) that are roughly the size of census blocks or block groups. The 2006–10 CTPP are based on five years of American Community Survey (ACS) data.

A key challenge for CTPP data users is to account for the MOEs appropriately when developing long-range plans, validate traffic models, and provide information so decision-makers will deploy necessary funds. Incorporating

GLOSSARY

Dirichlet distribution. The Dirichlet distribution has two parameters, a scale and a base measure. Given the components of the base measure and scale, the Dirichlet provides the distribution over multinomials. It is an extension of the Beta distribution, which provides the distribution over binomials.

Generalized variance function. The generalized variance function (GVF) typically relates the relative variance to an estimated total through a regression model. The model parameters are published to allow data users to predict the variance associated with the estimate in-hand.

Imputation error. Imputation error is a component of the total variance when imputation is present in the data. Imputation is a statistical approach used to assign data values from available information when missing values are originally present in the dataset.

Interquartile range. The interquartile range is the range (or difference) between the values for the 25th and 75th percentiles.

Margin of error. The margin of error is the half-width of a confidence interval.

Metropolitan planning organization. A metropolitan planning organization is a policy board created by the governor and local governments that carries out the transportation planning for the metropolitan area.

Perturbation error. Perturbation error is a component of the total variance when perturbation is present in the data. Perturbation is a statistical approach that is sometimes used to replace data values from available information to maintain data confidentiality and reduce the risk of unintended data disclosure.

Relative variance. Relative variance is equal to the variance divided by the estimate squared.

Traffic analysis district. Traffic analysis districts (TADs) are basic aggregates of traffic analysis zones (TAZs) to allow transportation planners in various forecasting efforts.

Traffic analysis zone. Traffic analysis zones (TAZs) are aggregates of Census blocks with similar commuter travel.

Table 1—Age by Means of Transportation (MOT)—Original Table (in MPO 34198200 and TAD 00000063), CTPP 2006-2010

Cell	AGE	MOT	Weighted Frequency	Cell Percent p_k	Original MOE
1	18–24	Car, truck, van, drove alone	350	0.024	126
2	25–44	Car, truck, van, drove alone	2,755	0.193	444
3	45–59	Car, truck, van, drove alone	1,585	0.111	258
4	60–64	Car, truck, van, drove alone	290	0.020	125
5	65–74	Car, truck, van, drove alone	160	0.011	80
6	75+	Car, truck, van, drove alone	25	0.002	38
7	18–24	Carpool	175	0.012	88
8	25–44	Carpool	705	0.049	221
9	45–59	Carpool	475	0.033	164
10	60–64	Carpool	70	0.005	62
11	65–74	Carpool	15	0.001	21
12	16–17	Other	40	0.003	44
13	18–24	Other	1,115	0.078	223
14	25–44	Other	4,180	0.292	453
15	45–59	Other	1,730	0.121	288
16	60–64	Other	210	0.015	100
17	65–74	Other	365	0.026	111
18	75+	Other	55	0.004	53
Total			14,300	1.000	820

Source of original data table: U.S. Census Bureau, American Community Survey 2006-2010 Five-year estimates.
Special Tabulation: Census Transportation Planning.

the sampling error associated with CTPP estimates in travel demand modeling is difficult to do. At the very least, accurate measures of precision are needed when conveying information for decision-making.

Table 1 shows the weighted frequencies, cell percents, and original MOEs for a published CTPP table "Age by Means of Transportation (MOT)" in a Traffic Analysis District (TAD) in Newark, New Jersey (Metropolitan Planning Organization (MPO) ID is 34198200 and TAD ID is 00000063). This table has 18 non-zero internal cells (cells with a weighted frequency of zero become undefined). The parameters for the Dirichlet distribution $\hat{\alpha}_k$ were estimated using both methods described above.

Table 2 shows the results from two sets of 1,000 replicated tables generated by the GVF-based and distance function methods. To evaluate the GVF-based replicated tables method, we created a GVF function using a sample of the tables, which is a representative sample of geographic areas.

The resulting estimated parameters were $a = -0.00023$, and $b = 24.8988$. As a check, the average of the weighted frequencies across the 1,000 tables for each cell is the same for both methods, and are very close to the original weighted frequencies. For both methods, the average interquartile range across the 1,000 replicates is 919, and it is 920 for the original data. The parameters of the Dirichlet distribution, $\hat{\alpha}_k$, are slightly different between the two methods.

The MOE, based on the standard deviation among the replicated tables for each cell estimate, was computed. The results from the GVF-based replicate estimates are close to the traditional GVF-based MOEs. The GVF-based replicate estimates are only as good as

Table 2—Age by Means of Transportation (MOT)—MOE Results from 1,000 Replicated Tables (in MPO 34198200 and TAD 00000063), CTPP 2006–2010

Cell	Original Weighted Frequency	Mean Computed Across Replicated Tables	$\hat{a}_k$ (Method 1)	$\hat{a}_k$ (Method 2)	Original MOE	MOE Computed from GVF	MOE Computed from GVF Replicated Table Method	MOE Computed from Distance Function Method
1	350	349	14.03	16.97	126	153	153	140
2	2,755	2,756	110.46	133.58	444	426	427	394
3	1,585	1,584	63.55	76.85	258	324	314	288
4	290	292	11.63	14.06	125	140	141	129
5	160	161	6.41	7.76	80	104	103	94
6	25	25	1.00	1.21	38	41	44	40
7	175	176	7.02	8.48	88	109	108	98
8	705	706	28.27	34.18	221	217	219	200
9	475	474	19.04	23.03	164	179	179	163
10	70	69	2.81	3.39	62	69	68	62
11	15	16	0.60	0.72	21	31	32	29
12	40	40	1.60	1.94	44	53	47	43
13	1,115	1,116	44.70	54.06	223	273	262	239
14	4,180	4,182	167.59	202.67	453	520	516	480
15	1,730	1,730	69.36	83.88	288	339	332	305
16	210	212	8.42	10.18	100	118	116	106
17	365	368	14.63	17.70	111	156	156	142
18	55	54	2.21	2.67	53	61	60	55
Total	14,300	14,310			820	915	835	835

the GVF model. That being said, for the CTPP table used in this evaluation, the distance function method is closer to the original MOE in most cases. If the traditional GVF variances do not approximate the true CTPP variances well, the distance function method would be able to generate a set of replicated tables that works better to reflect the CTPP variances.

A user of the two approaches is encouraged to do both approaches and evaluate the closeness to the original MOE. In practice, a set of five replicated tables can be fed into the sophisticated traffic analysis software one table at a time to help understand the uncertainty about the predictions of traffic flows and amount of traffic.

Discussion Through Demonstrations

There are several ways to generate the replicated tables for sensitivity analysis. The applications show 1) use of GVFs from a survey in combination with the GVF-based replicated table generator, 2) use of distance-function method with application to linear regression, and 3) sensitivity assessment via heat maps from the distance function approach.

Demonstration 1: Use of GVF-based Method

The Current Population Survey Annual Social and Economic Supplement (APECS) 2016 data uses GVFs to provide flexibility for users to compute variances for a wide range of queries. A query tool is available at *https://www.census.gov/cps/data/cpstablecreator.html*. The table of interest (Table 3) is the estimated population of health

Table 3—Replicated Tables for Estimated Population of Health Insurance and Disability Status for Asians in the West

Insurance Status	Disability Status	X (000s)	MOE (000s)	Table 1 X (000s)	Table 2 X (000s)	Table 3 X (000s)	Table 4 X (000s)	Table 5 X (000s)
Insured	Disability	414	76	387	396	441	422	365
Insured	No Disability	5,932	278	5,871	5,930	6,046	5,813	5,892
Insured	Not in Universe	1,241	131	1,324	1,325	1,201	1,229	1,202
Uninsured	Disability	18	16	8	13	17	14	15
Uninsured	No Disability	459	80	352	432	473	420	519
Uninsured	Not in Universe	39	23	26	36	30	17	46

Source of original data table: U.S. Census Bureau, Current Population Survey, APECS 2016.

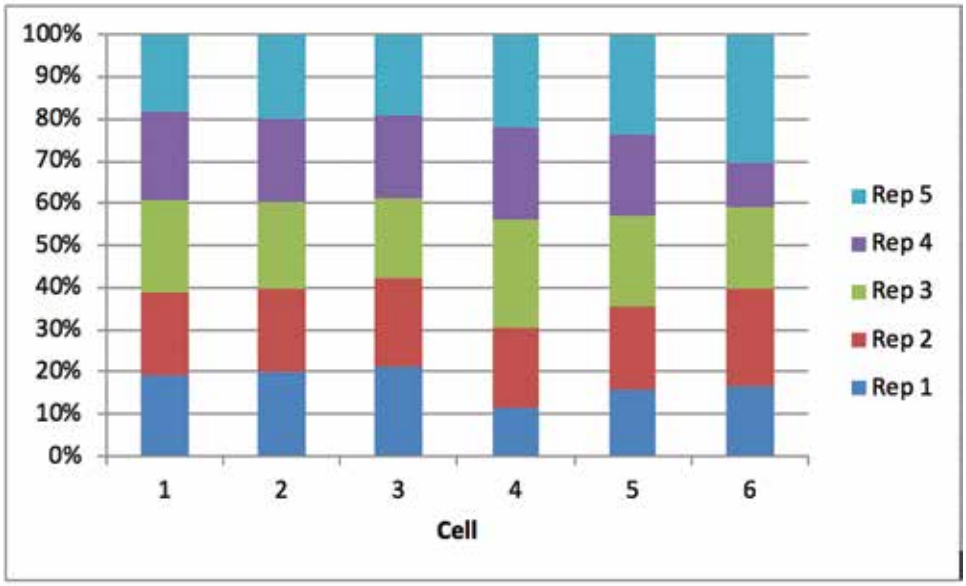


Figure 1. Health Insurance and Disability Status for Asians in the West—Scaled Bar Chart by Cell for the Five Replicated Tables.

insurance and disability status for Asians in the West.

The table shows the original estimates and MOEs, with MOEs computed following the instruction for using GVF *b* parameter value of 4,653 from Table 3 in the Source and Accuracy appendix in the Current Population Survey technical documentation for tables on health insurance. The GVF-based replicated tables method generated five replicated tables displayed in Table 3, with variation displayed visually in the scaled bar chart in Figure 1.

Demonstration 2—Use of Distance-function-based Method and Subsequent Regression Analysis

Continuing with the CPS APECS 2016 data, the table of interest, as shown in Table 4, is the number of non-citizen males by state. The

Table 4—Replicated Tables for Estimated Population of Non-citizen Males, by State

STATE	X in 000s	MOE in 000s	Table 1 X (000s)	Table 2 X (000s)	Table 3 X (000s)	Table 4 X (000s)	Table 5 X (000s)
AL	43	22.1	63	49	45	28	23
	12	4.7	7	21	10	12	6
AZ	303	61.2	318	281	300	275	331
AR	56	20.4	55	53	46	53	42
CA	2,521	176.9	2,398	2,489	2,561	2,473	2,394
CO	183	47.3	206	151	228	147	189
CT	104	30.0	106	95	149	87	94
DE	22	7.0	14	10	8	12	18
DC	33	7.4	49	27	37	23	25
FL	939	102.9	977	896	927	931	970
GA	357	64.1	321	363	361	387	305
HI	52	12.9	54	57	34	50	35
ID	45	13.6	34	33	35	31	37
IL	451	72.6	465	410	458	354	488
IN	120	36.8	143	115	104	102	107
IA	52	20.2	53	51	31	56	45
KS	58	22.0	70	59	70	66	45
KY	75	29.4	93	85	106	79	78
LA	58	24.5	69	54	85	58	75
ME	6	4.9	2	1	10	3	3
MD	257	54.1	318	273	214	217	261
MA	301	57.3	297	262	298	262	293
MI	216	49.3	239	243	224	241	224
MN	126	38.0	154	109	116	93	102
MS	34	15.5	18	27	13	32	18
MO	55	25.3	65	65	61	63	67
MT	5	3.3	0	7	5	10	4
NE	40	14.5	58	36	37	33	56
NV	129	31.4	162	103	122	154	144
NH	16	7.4	26	21	22	8	18
NJ	539	77.7	491	557	506	594	537
NM	63	18.0	47	45	61	45	65
NY	910	103.5	865	964	1,039	990	978
NC	257	55.3	277	251	262	340	264
ND	14	4.9	14	18	37	22	20
OH	157	42.0	142	168	152	165	149
OK	88	30.7	111	70	71	86	102
OR	148	39.8	94	153	172	173	105
PA	185	45.8	148	207	163	175	184
RI	46	11.3	54	41	63	43	51
SC	61	25.8	93	72	79	66	40
SD	7	4.0	12	1	12	1	3

continued on p. 24

Table 4—Replicated Tables for Estimated Population of Non-citizen Males, by State (continued)

STATE	X in 000s	MOE in 000s	Table 1 X (000s)	Table 2 X (000s)	Table 3 X (000s)	Table 4 X (000s)	Table 5 X (000s)
TN	147	40.5	156	134	135	189	153
TX	1,621	144.4	1,613	1,617	1,549	1,592	1,570
UT	71	19.5	60	63	90	81	57
VT	3	2.4	0	2	0	7	1
VA	263	56.0	259	315	313	306	299
WA	320	61.1	315	288	284	273	267
WV	5	5.0	0	0	4	19	0
WI	85	31.3	87	69	110	104	63
WY	4	2.6	1	5	2	1	0
Total	11662	325.7	11674	11484	11824	11614	11408

Source of original data table: U.S. Census Bureau, Current Population Survey APECS 2016

Table 5—Regression Coefficients and Adjusted R^2 Values from the Original Data and Replicated Tables

	Intercept		Proportion Hispanic		Log of Annual Wage per Employee		
Sample	Estimate	Standard Error	Estimate	Standard Error	Estimate	Standard Error	Adjust R^2
Original	-0.509	0.0753	0.092	0.0112	0.049	0.0070	0.756
Rep 1	-0.634	0.0984	0.083	0.0146	0.060	0.0091	0.668
Rep 2	-0.529	0.0874	0.079	0.0130	0.050	0.0081	0.667
Rep 3	-0.590	0.0936	0.086	0.0139	0.056	0.0087	0.681
Rep 4	-0.412	0.1006	0.087	0.0150	0.040	0.0094	0.572
Rep 5	-0.471	0.0873	0.101	0.0130	0.045	0.0081	0.707

table shows the original estimates and MOEs. Replicating the table by the distance-function-based replicated tables method generated five replicated tables displayed in Table 4. To illustrate a subsequent analysis using the replicated tables, we include a multiple regression such as $Y^{(m)}=X\beta^{(m)}+\varepsilon$, for m = 1, 2, 3, 4, 5, where $Y^{(m)}$ is vector of estimates for the mth replicate table, X is a matrix of census counts for various characteristics, $\beta^{(m)}$ are the regression coefficients, and ε is the error term.

For illustration purposes, the regression estimated the proportion of non-citizen males in each state using two state-level covariates as independent variables: the proportion of Hispanic males from the Census Bureau 2016 Population Estimates program, and the log of annual wages per employee from the Bureau of Labor Statistics 2015 "Employment and Wages, Annual Averages" program. The

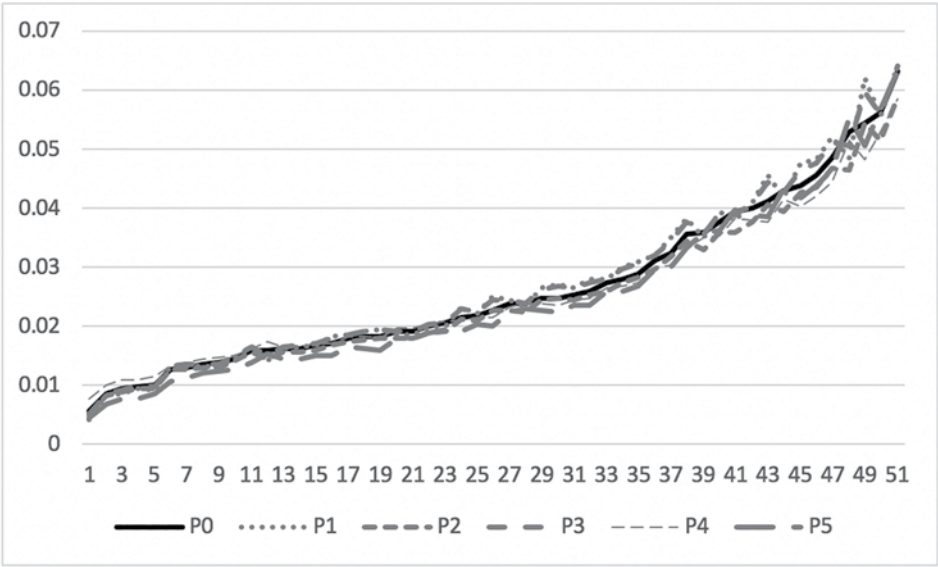


Figure 2. Graph of predicted values of the proportion of male non-citizens, by state.

resulting regression coefficients and adjusted R^2 values are shown in Table 5.

It is interesting to note that the standard errors for each replicate are wider than from the original, and the adjusted R^2 values from the replicated tables are lower than for the original. The plot in Figure 2 shows the predicted values of the proportion of male non-citizens, by state, from the original data (sorted from smallest to largest) and each of the five replicated tables. The original data result is always within the results of the replicated tables.

Demonstration 3—Use of Heat Maps

The impact of sampling error on a heat map can be demonstrated by replicating the original map five times. The demonstration uses 2006–2010 CTPP data on the estimated number of workers 16 and over who arrive at one of the 95 workplace tracts in Dakota County, Minnesota, between 9 and 10 in the morning. We applied the distance function methodology to produce variation across the five estimates to align with the margin of error.

As shown in Figure 3 on the following page, we created six heat maps for the 95 tracts, one for the original estimates and one for each of the 5 replicated estimates. The heat maps show quite a bit of traffic heading into workplace area tracts in the northwest section of the county, with some of the more-stable estimates not changing its shade, while most tracts in other areas of the county are subject to a change in the shading, signifying less-stable estimates.

Conclusions

Static tables continue to be published and made available online. Users of the tabular estimates should take into account the MOEs in some manner before making decisions based on analysis results that use the tabular data. To address this issue, a mechanism we refer to as the replicated tables approach has been developed to allow the sampling errors from an upstream table (based on the MOEs associated with the numbers in the table) to permeate into downstream analyses.

Two tools are available to users to generate replicated tables for sensitivity analysis. An R function *rep.tab* generates replicated tables based on the algorithms described above (see Appendix B in our 2017 report for NCHRP Project 8-36c, Task 135, for the R code). The R code was developed to allow users to prepare the replicated tables for a given CTPP table. An Excel file called the CTPP MOE ToolKit is an alternative to R. The Excel macro program for replicated tables approaches is available at *https://bit.ly/2Hv1RKG* with a tutorial available at *https://bit.ly/2Jd5mI8*.

The GVF method in the Excel file uses the *b* parameter derived from the CTPP application and

Note: The replicated tables methodology was developed with support from the RAND Corporation for the National Cooperative Highway Research Program Project 08-36.

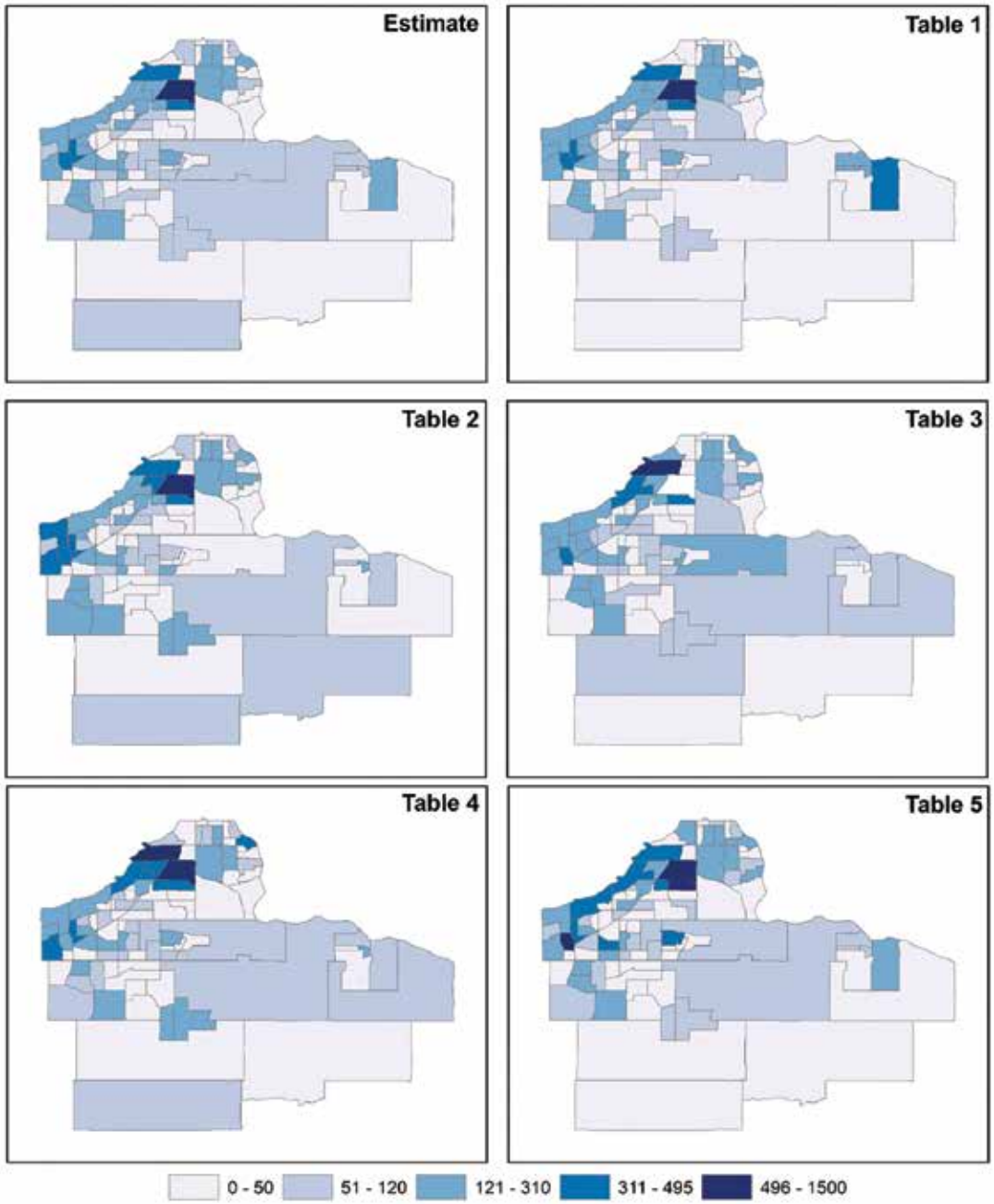


Figure 3. Replicated heat maps on the estimated number of workers 16 and over who arrive at one of the 95 workplace tracts in Dakota County, Minnesota, between 9 and 10 a.m., 2006–10 Census Transportation Planning Products.

may not be appropriate for other applications. In practice, as demonstrated in this article, researchers can use these tools for static tables to generate the replicated tables multiple times and fit the simulated values to their desired subsequent analysis, one replicated table at a time. Various graphs (such as maps, bar charts, and pie charts) can be produced to get a picture of the variation across the replicated tables, which helps the user visualize the precision of the estimates, given the MOE. C

Further Reading

Connor, R.J., and Mosimann, J.E. 1969. Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution. *Journal of the American Statistical Association* 64 (325):194–206

CPS. 2016. *Current Population Survey, 2016 ASEC Technical Documentation*. Available at *https://bit.ly/2T6H7Lw*.

Fuller, S. 2010. Analyzing Generalized Variances for the American Community Survey 2005 Public Use Microdata Sample. Final Report, April 20, 2010. Decennial Statistical Studies Division. U.S. Bureau of the Census. Available at *https://bit.ly/2ucocEJ*.

Li, J., and Krenzke, T. 2017. Addressing Margins of Error in Small Areas of Data Delivered through the American Factfinder or the Census Transportation Planning Products Program. NCHRP Project 8-36c, Task 135. Available at *https://bit.ly/2HzR4is*.

Thabane, L., Mbuagbaw, L., Zhang, S., Samaan, Z., Marcucci, M., Ye, C., Thabane, M., Giangregorio, L., Dennis, B., Kosa, D., Debono, V., Dillenburg, R., Fruci, V., Bawor, M., Lee, J., Wells, G., and Goldsmith, C. 2013. A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. *BMC Medical Research Methodology* 13:92. *https://doi.org/10.1186/1471-2288-13-92*.

Wolter, K. 1985. *Introduction to Variance Estimation*. New York, NY: Springer-Verlag, New York, Inc.

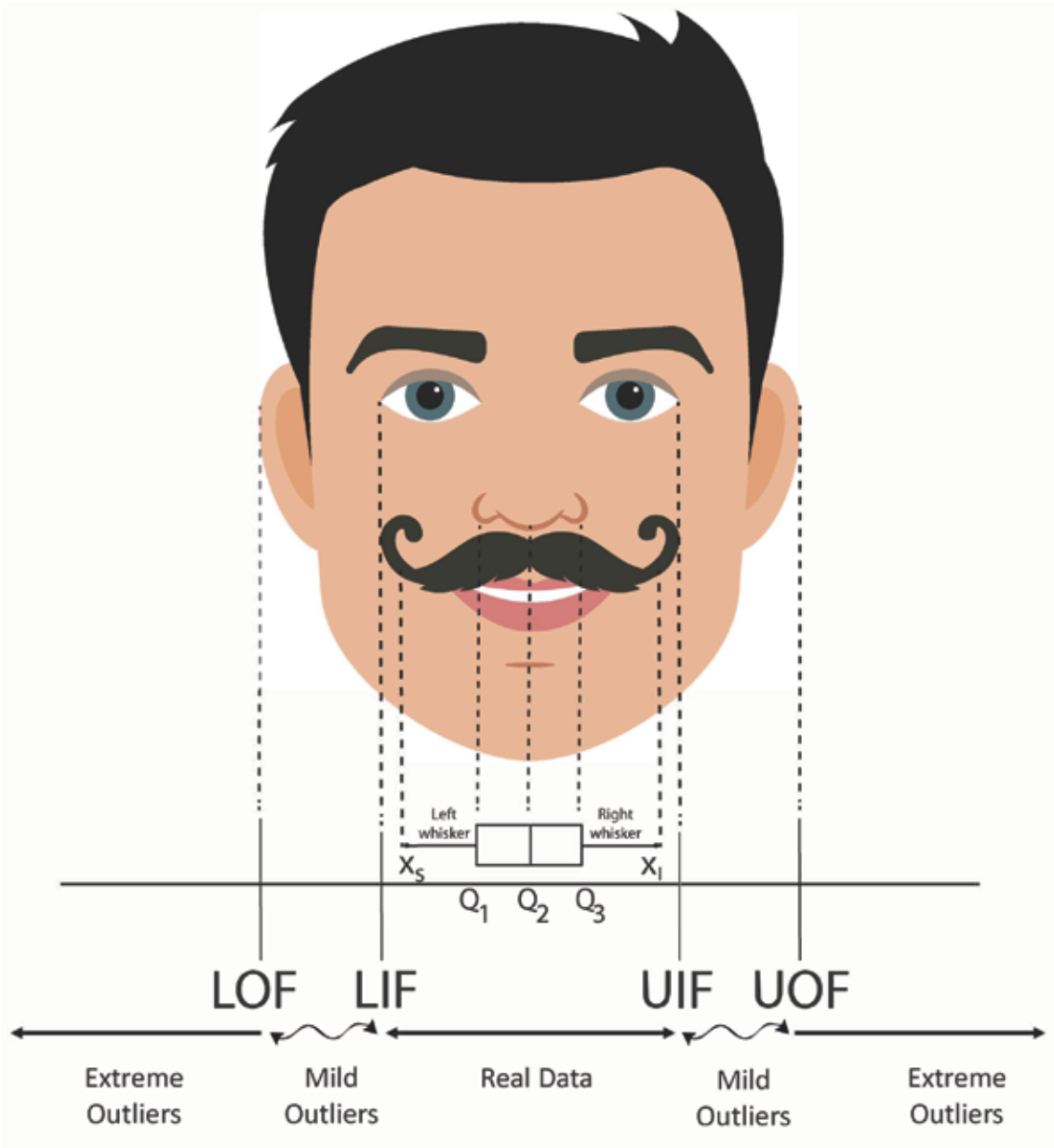
About the Authors

Tom Krenzke is an associate director in Westat's Statistical and Evaluation Sciences Unit. He has led research on statistical topics such as sample design, data confidentiality methods, and small area estimation, working toward solutions for several federal agencies. He is an ASA Fellow and president of the Washington Statistical Society, and serves on the ASA Privacy and Confidentiality Committee and Westat Institutional Review Board.

Jianzhu Li is a senior statistician in Westat's Statistical and Evaluation Sciences Unit with more than 15 years of experience in survey research, including sample design, non-response adjustment, imputation, data confidentiality and disclosure protection, and small area estimation. She has worked on several statistical confidentiality and sampling-related projects over the past few years for government agencies such as the NSF, NCHS, NCES, Census Bureau, and USDA.

Box Plot Versus Human Face

Sarjinder Singh



Notations

Q_1 : First quartile—left ala of the nose*
Q_2 : Median—apex of the nose
Q_3 : Third quartile—right ala of the nose

LIF—Lower inner fence; left corner of left eye
UIF—Upper inner fence; right corner of right eye

X_s: Smallest data value between the LIF and the UIF = Dali of left-side whisker/mustache
X_l: Largest data value between the LIF and the UIF = Dali of right-side whisker/mustache

LOF—Lower outer fence; left edge of left ear
UOF—Upper outer fence; right edge of right ear

Explanation

- **Real Data:** Data between the LIF and the UIF = data observed by both eyes

- **Mild Outliers**: Data either between LIF and LOF or between UIF and UOF
  —Not observed, but heard
  —Data between eyes and ears

  Note: A mild outlier could be part of a real data set; must verify before discarding it

- **Extreme Outliers:** Data below LOF or above UOF
  —Never heard something

  Note: Extreme outliers, in general, are not part of data set

  Q_2—Q_1 = Length of left-side box = size of left nostril
  Q_3—Q_2 = Length of right-side box = size of right nostril

  Note: If the left nostril is the same size as the right nostril, the human face is symmetrical and looks beautiful. Otherwise, it will look ugly—either skewed to the right or skewed to the left.

  Q_1—X_S = Length of the left side of whisker/mustache
  X_l—Q_3 = Length of the right side whisker/mustache

  Note: If left whisker/mustache is the same length as right whisker/mustache, the human face looks attractive and powerful.

*The ala nasi—wing—of the nose is the lower lateral surface of the external nose. It is cartilaginous and flares out to form a rounded eminence around the nostril.

About the Author

Sarjinder Singh is a tenured professor in the Department of Mathematics at Texas A&M University–Kingsville. He earned his PhD from Punjab Agricultural University in India. His research interests include survey sampling and applied statistics.

2019 NEW ENGLAND SYMPOSIUM ON STATISTICS IN SPORTS

Harvard University (Science Center), Cambridge, Massachusetts

Saturday, September 28, 2019

The **2019 New England Symposium on Statistics in Sports** (NESSIS) will be a meeting of statisticians, statistical researchers, and quantitative analysts connected with sports teams, sports media, and universities to discuss common problems of interest in statistical modeling and analyzing sports data. The symposium format will be a mixture of invited talks, a poster session, and a panel discussion. Students in particular are encouraged to submit abstracts; a prize will be awarded to the best student poster as decided by a panel of judges.

Registration is now open.

Registration information, including fees and methods of payment, can be found at *http://www.nessis.org/registration.html*.

Abstracts for talks and posters are now requested, and should be submitted via the online abstract submission form.

Submission deadline for abstracts: June 15, 2019.

Decisions on accepting abstracts will be made by June 30, 2019.

Further details of the 2019 NESSIS are forthcoming. Complete up-to-date information will be posted at the 2019 NESSIS website.

Contact symposium co-organizers Mark Glickman (*glickman@fas.harvard.edu*) or Scott Evans (*sevans@bsc.gwu.edu*) with any questions.

We would appreciate your forwarding this announcement to anyone who might be interested.

Teaching Upper-level Undergraduate Statistics through a Shared/ Hybrid Model

Jingchen (Monika) Hu

At small liberal arts colleges like Vassar College, statistics courses attract a large and growing number of students across campus, yet course offerings are limited due to faculty size. With two statisticians on campus, Vassar offers sections of six courses each year: Introduction to Statistics (algebra- and calculus-based), Probability, Applied Statistical Modeling (Stats II), Statistical Inference (classical), and Bayesian Statistics. These offerings stretch faculty resources as far as possible, but we were determined to establish a statistics major and provide a rich curriculum. To do so, we need more intermediate and advanced courses, but are constrained by faculty size and growing demand for current courses.

This challenge is not unique to Vassar College. Almost everyone is trying to establish and grow their statistics curricula. Some small schools have already made progress in offering a statistics minor and/or major, or even data science courses.

Further growth in demand for statistics and data science courses is projected. Statisticians are expected to eventually make up a third of the faculty in a mathematics and statistics department at a small liberal arts college offering courses in pure mathematics, applied mathematics, and statistics, and statistics courses could constitute a third of the curriculum of such a department.

Before we arrive at this world, how can we encourage and enable interested students to immerse themselves in more statistical thinking, analysis, and practice within the constraints of faculty size?

The Upper Level Math/Stats Project

In search for possible answers to this question, I joined the Upper Level Math/Stats Project (*http://lacol.net/summary-upper-level-math-stats/*) of the Liberal Arts Collaborative for Digital Innovation (LACOL, *http://lacol.net/*). The pilot phase of this project was devoted to exploring ways to share mathematics and statistics classes remotely through a number of small liberal arts colleges. The project targeted upper-level courses, since these courses tend to be offered less regularly due to smaller enrollments.

Upper-level statistics courses have been developed at most LACOL member schools, and the topics are usually related to the expertise of the faculty instructors. If such courses can be shared efficiently and effectively through a consortium, using hybrid/online delivery modes, students would be able to take courses offered by professors at member schools other than their own. If, ultimately, faculty members with various teaching styles were equipped with the right technology and sufficient support, all sharable upper level courses were listed and open to all students, and students were able to receive grades and credit as in any other face-to-face courses, this sharing mechanism could expand the statistics curriculum for participating campuses. Students also would be able to network with faculty members and students from other schools, which would provide more-diverse perspectives and opportunities.

Teaching Bayesian Statistics through a Shared/Hybrid Model

To test the shared/hybrid model with an upper-level statistics course, I offered Vassar College's Bayesian Statistics course in fall 2017. Two other mathematics courses were offered during the pilot phase (AY 2016–2018) as well.

Bayesian Statistics met twice a week at Vassar College, with each lecture at 75 minutes. The lectures were available to remote students in two ways:

synchronously using a video conference software called Zoom (*https://www.zoom.us/*) or asynchronously via YouTube.

Local students could use regular face-to-face office hours, remote students could join synchronously via Zoom, and there were additional online office hours each week for remote students.

The course page was hosted through Vassar College's learning platform/course management system, Moodle. Authorized accounts were created for remote students. The Moodle course page contained lecture slides, data sets, R scripts, homework, case studies, R resources, etc. We also made heavy use of Google Docs: Surveys, course schedule, group work assignments, and appointment sign-ups were through Google Docs, mainly with Sheets and Forms.

Figure 1. Equipment.

An iPad Pro, Apple Pencil, and directional microphone formed the basic equipment set for every lecture (Figure 1). Class meetings were initialized through Zoom with the desktop in the classroom. I joined the Zoom meeting with the iPad Pro and shared the screen through either Dropbox or Google Drive, where lectures slides were stored. The screen was also projected in the classroom, so local students and remote students had the same viewing experience. Camera views of the classroom and all remote students were included on the screen (Figure 2).

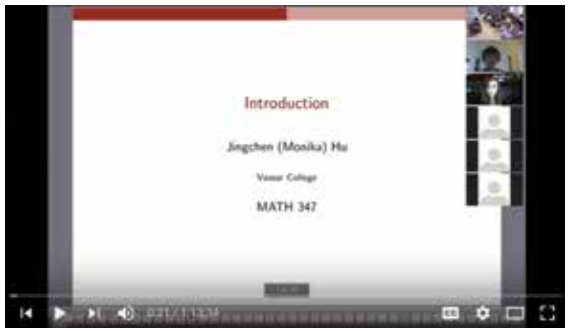


Figure 2. Lecture recording.

Within Zoom, I swiped to turn to previous/next pages or slides. I could also use the Apple Pencil to write on slides with the writing and erasing toolbox options. I usually used the Apple Pencil to highlight key points from the slides, showed formula and derivation, and sketched graphs, etc. (Figure 3).

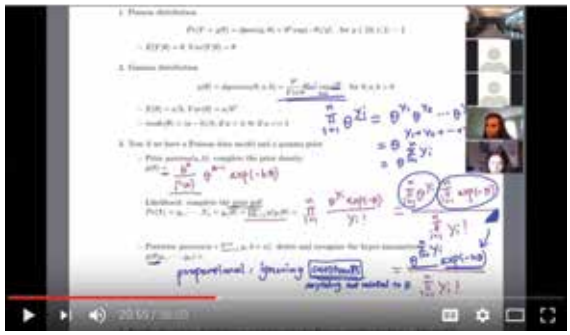


Figure 3. Lecture notes.

I could also share a whiteboard that provides more writing space. I used the whiteboard function to demonstrate problem-solving procedures, go over examples in detail, and set up a model from scratch, etc.

In addition to lecturing, I used the whiteboard function to record shorter videos outside lectures. Some videos aimed at recapping and/or highlighting materials in class that could not be explained thoroughly because of time constraints, while other videos focused on hints for selected homework problems (Figure 4).

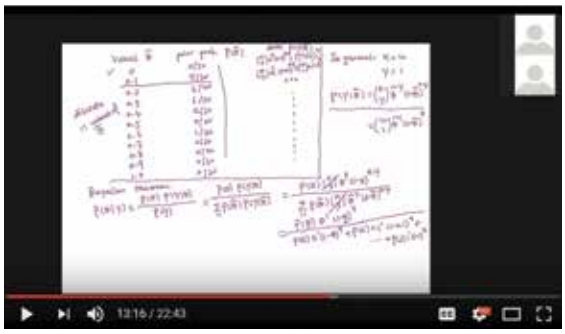


Figure 4. Lecture whiteboard.

When demonstrating R, I joined the Zoom meeting with my laptop, then shared my screen to show my R/RStudio. The demonstration worked similarly to a traditional classroom when projecting the screen in the classroom. I also used R demonstration in shorter videos outside lectures, to go over elements such as programming techniques and script explanation (Figure 5).

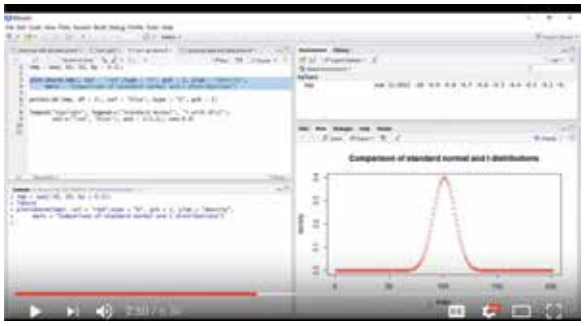


Figure 5. Lecture R demonstration.

This semester-long experiment brought technological, administrative, and pedagogical challenges. Technological and administrative challenges, such as adapting to writing on an iPad, were plentiful at the beginning of the semester, but resolved with practice and by early preparation, such as creating sponsored accounts for remote students to access the course material.

Overcoming some of these challenges requires joint effort within the consortium, such as creating a guideline for students to receive credits from remote courses. Some challenges and solutions were unique to instructors. For example, I used an iPad and slides, while two math colleagues taught with a chalkboard. Their course delivery focused more on how to use a high-definition camera to capture lectures well.

As a consortium-wide project team, we were experimenting individually and trying to provide workable examples and solutions to ultimately create a guideline document. The guideline document aims at helping future faculty members redesign their courses with a shared/hybrid model. Technology and administration challenges and solutions in the Bayesian course have made a meaningful contribution to such a guideline.

Unlike the technology and administration issues, the pedagogy challenges persisted throughout the semester. Some of the challenges arose due to the shared/hybrid nature of the course; for example, what to move online and in what form, and how to create an online learning community.

Some of the challenges arose because of the course material. Including theory and methods, computation, and applications of Bayesian statistics pushed instructors to work out a fine balance between the three at the undergraduate level. This challenge already exists in a traditional face-to-face course, but implementing a shared/hybrid model created an extra layer of difficulty and complexity.

Expanded Thinking

This teaching experience and the associated challenges helped me to think more broadly about pedagogy and statistics education, and how the shared/hybrid model can provide some solutions and opportunities for overcoming the challenges.

In general, four categories of material are worth being moved online: review of pre-requisites, preview of upcoming content, recaps from lectures, and hints for homework problems and/or R programming.

In a traditional face-to-face course, I believe the aforementioned categories of materials are necessary as well. I usually make such material available in more-traditional forms—for example, posting a review document of useful matrix algebra results before covering multivariate normal models (review of prerequisite material), giving a motivating application example before covering the methodology (preview of upcoming material), spending 10 minutes at the beginning of a lecture to recap material from the previous lecture (recap from lecture), and scheduling designated lab time to practice R programming (hints for homework problems and/or R programming).

Such approaches are often useful, but they usually take up lecture time and can take away the flow of the material. I am always searching for alternative media to share and go over these necessary materials.

Thanks to recording and watching lectures with the shared/hybrid instruction model, both the students and I became comfortable with multimedia resources.

The flexibility and accessibility of current tools led me to creating shorter videos targeting specific needs.

Recap from Lecture

Early in the semester, we covered Bayes theorem and a working example of applying it to a discrete prior distribution and a collected data set from our class. The example was given right after Bayes theorem and its intuition, and the material was probably new to a good number of students. Students might understand the application mathematically in class, but its details and implications were difficult to digest. To enhance student understanding, I made a 20-minute video available online that went over the in-class example (Figure 4) more thoroughly. The video walked students through the problem-solving process step-by-step.

This helped clear up confusions, and provided a deeper understanding and appreciation of the problem itself and the bigger topic. Although the video was optional, students found it useful, and some watched it several times.

Hints for Homework Problems

In the Gibbs sampler section, one homework question dealt with a mixture model, which made sense intuitively in context, but proved to be challenging to derive mathematically. I made a 20-minute video to set up the model (writing out the likelihood function and prior distributions, thus providing the joint posterior distribution of all parameters) and gave hints for derivation techniques. Compared with a traditional document containing all information in text form, the video provided the opportunity for students to learn the problem-solving process. Posting such a video helped students solve the homework question and enhanced their understanding of mixture models and the Gibbs sampler. It also reduced the number of times that students came to office hours and saved me from spending time on explaining the same problem multiple times.

Making videos about commonly challenging material was beneficial. This was also true with demonstrations of targeting the uses of R programming specific to problems in this course. Students found these targeted videos more helpful than random YouTube videos. The videos about R demonstration helped overcome the common challenges of teaching R programming in class and/or lab. These short videos were great supplements to the in-class/lab programming instruction and exercise.

Preview of Upcoming Material

As part of my effort to weave applications of Bayesian methods into the curriculum, I collaborated with a colleague in Vassar's Cognitive Science Program to describe his research projects as an introduction to Bayesian hierarchical modeling. In a traditional face-to-face course, such introduction can be a short guest lecture, but it is generally not easy to schedule extra teaching duties during the semester. Instead, I invited my colleague to make a 20-minute overview video for students to watch before our lecture on hierarchical modeling.

The video-making process proved to be both more challenging and more beneficial than I expected. My cognitive science colleague came from an applied background, and understanding his research projects required much background information and explanation. His research used Bayesian hierarchical modeling extensively, and available packages to implement the complicated computation procedures. My students were, overall, mathematically and probabilistically prepared, and they were interested in why the model worked, the way it was set up, and how the computation was realized. My colleague and I went through several iterations before settling on what to cover in the video.

The students found the introduction video very interesting (*https://bit.ly/2TEOvC4*). The applications pushed their thinking about the methodology and the model assumptions and their implications in cognitive modeling. We revisited their questions several times in lectures and online discussion boards. While a guest lecture could achieve similar effects, making a video allowed students to pace their intake and re-watch as much as they like. It also enabled and encouraged them to watch and think about the topic before the class meeting and lecture.

To achieve desirable outcomes, all of the above multimedia tools must work together with creating an online learning community. To create such a community, students made self-introduction posts on Moodle, sharing information about themselves such as name, year, and major(s), along with previous probability and statistics course experience, R programming exposure, and topics of interest for class projects.

Another assignment was a digital project introduction post. We had a poster session at the end of the semester for students to present their project work in an interactive way. A two-minute introduction video to their projects was required when they submitted

their posters. Before the poster session, students were expected to watch each other's videos. In this way, students could get a general idea of their peers' work in advance. This practice helped students plan their poster visits better. It also helped them think about how to present their projects in a concise and appealing way.

To foster an online learning community, encourage students to communicate with each other, and keep students engaged with the course material, I made extensive use of the forum function on Moodle. In addition to the self-introduction purpose, we used the boards for posting general questions and comments about sections, chapters, homework assignments, and R questions; responding to questions from reading/viewing material; case study analyses; and sharing resources.

Responding to Questions from Reading/Viewing Material

I believe undergraduate students in upper-level statistics courses should be exposed to accessible research papers. For this Bayesian Statistics course, I chose "Explaining the Gibbs Sampler" (Casella and George, 1992) from *The American Statistician* for students to read, digest, and discuss.

This paper presented a series of accessible simulation studies that illustrated key ideas of the Gibbs sampler and its practices. Some of practices in the paper are different from what people do nowadays, due to progress made in the field since it was published. I believed discussing those differences can enhance students' understanding.

In addition to sharing this paper, I created a reading guide containing six questions. I asked the students to make one post to respond to any of the questions before we discussed the paper in class. Based on our class discussion, I later added more questions and asked the students to make one more post with their responses.

The discussion questions ranged from understanding the paper content and making comparisons (e.g., "How do Gelfand and Smith (1990) suggest obtaining an approximate sample from f(x)? How is it different from or similar to the approach we talked about in class? What are the advantages and disadvantages of each approach?") to extension and illustration (e.g., "The authors claimed, 'a defining characteristic of the Gibbs sampler is that it always uses the full set of univariate conditionals to define the iteration.' Explain this claim by illustrating how a Gibbs sampler works with k parameters"). I believed a list of questions could help students grasp the paper content. Asking students to post a response to one question from such a wide range gave them the opportunity to express their best understanding based on their strengths.

I found the first round of posts ensured that students had done the required reading and made a reasonable effort to understand the material. The second round of posts encouraged more reflection on the reading and in-class discussion. Moreover, making and sharing posts fostered a learning environment where discussions happen naturally, so students can learn from each other. On multiple occasions, I saw students' responses referring to previous posts, voicing agreement/disagreement, asking further questions, and offering new insights.

I described my collaboration with a colleague from Cognitive Science to create an introduction video about Bayesian hierarchical modeling applications in cognitive modeling. The video served as a preview of our Bayesian hierarchical modeling section.

To ensure that students would watch and react to the introduction video, I asked them to make one post before our class meeting. Their posts touched on various aspects of the video. Some explained the model based on their understanding, while many others asked further questions and created discussions in the posts. I could tell that almost everyone had started thinking about why hierarchical modeling was needed in this application and wondering how it was implemented.

In class, after covering Bayesian hierarchical methods, we were able to revisit the discussion board to discuss some of the posted questions, clarify any misunderstanding, and connect to the methods covered in class. The online posts greatly fostered our in-class discussion, thanks to the common understanding of the material they helped create.

I later responded to students' posts. I sent any questions that I could not answer to my colleague and shared his responses on Moodle. I also was able to share my colleague's Just Another Gibbs Sampler (JAGS) script. Since the JAGS code showed the model specification, interested students were able to deepen their understanding and may be able to implement similar methods in the future.

Discussion boards greatly enhanced students' engagement with the course material and the created learning environment. Although started because of

the shared/hybrid nature of this Bayesian Statistics course, the experience has convinced me to use the discussion boards more extensively in my traditional face-to-face courses.

Looking Ahead

This experiment with a shared/hybrid model for an undergraduate upper-level Bayesian Statistics course had been an extremely rewarding experience. Throughout the semester, there were times when I had to scratch my head to figure out how to make students collaborate across campus, and there were also times when I felt thrilled as students' discussion online and in-class showed deeper understanding and appreciation of the material. Through all the ups and downs, I believe I have become a more-thoughtful statistics educator. I cannot wait for more opportunities to test my multi-faceted instruction methods and develop new ones.

If you are interested in offering an undergraduate upper-level statistics course in a similar shared/hybrid model, I recommend starting with a face-to-face course that you have developed already. A substantial amount of course redesign is involved; if you already have your course material from previous teaching, you can be more fully engaged in efforts on the redesign, such as how to move certain material online and how to create and foster an online learning community.

Other necessary preparation includes learning how to give synchronized/asynchronized access to the remote students. I use slides in my face-to-face statistics courses, so using Zoom to online stream and record video was straightforward. My colleague from Williams College uses a traditional chalkboard, so he had to experiment with a swivel system for the iPad to track his movements while he is teaching and record the board work.

This experiment shows that the shared/hybrid model provides opportunities to expand upper-level statistics course offerings at all colleges by using non-traditional ways of instruction and material delivery, and creating and fostering an online learning community. These successful experiments also shed light on ways to increase students engagement with the material and with the learning community in the traditional face-to-face classroom.

About the Author

Jingchen (Monika) Hu is a statistician and assistant professor at Vassar College. She has been active in exploring ways to expand advanced-level course offerings at small liberal arts colleges. In her research, she focuses on developing Bayesian statistical methodology for applied problems that intersect with the social sciences.

Teaching to, and Learning from, the Masses

Mine Çetinkaya-Rundel

According to the *New York Times*, 2012 was The Year of the MOOC (massive open online course). By 2014, the *Chronicle of Higher Education* had already published a post titled "2014: The Year the Media Stopped Caring About MOOCs." It's 2019, and I still care about MOOCs, and hundreds of thousands of learners still do as well. I don't think MOOCs are going to disrupt and reshape higher education (I never thought that, despite all the hype when they first emerged), but they do provide incredible learning opportunities.

A great deal has been written about MOOCs in the last few years, some praising them and some criticizing their lack of openness (for example, some MOOC providers remain free, some charge for certification, and some even charge for content) and the not-so-massive completion rates. This column is about none of those. It's about the experience of putting together a comprehensive introductory statistics MOOC that later evolved into a series of courses that make up a specialization, and the lessons learned along the way. Most importantly, it's about learning from the massive numbers of learners, in a way that is difficult to do in a brick-and-mortar class, and using that knowledge to improve the instructional materials.

Statistics with R Specialization

Statistics with R is a specialization offered on Coursera of five MOOCs designed and sequenced to help learners master the foundations of data analysis, statistical inference, and modeling (*coursera.org/specializations/statistics*). The specialization also has a significant hands-on computing component. The target audience is learners with no background in statistics or computing.

The specialization was designed with four major goals in mind: (1) modularity, (2) accessibility and relevance, (3) clear learning objectives and expectations, and (4) resource parity with a course taught at Duke University that covers the same content.

The first four courses are Introduction to Probability and Data, Inferential Statistics, Linear Regression and Modeling, and Bayesian Statistics. These courses cover exploratory data analysis, study design, light probability, frequentist and Bayesian statistical inference, and modeling. A major focus of all of these courses is hands-on data analysis in R; each course features computing labs in R where learners create reproducible data analysis reports, as well as fully reproducible data analysis projects demonstrating mastery of the learning goals of each of the courses.

The fifth course is a capstone, where learners complete a data analysis project that answers a specific scientific/business question using a large and complex data set. This course is an opportunity for learners to practice what they learned in the first four courses.

Originally, the content in the first three courses was offered in a single, 12-week-long MOOC for two years. It eventually was split into shorter (four-week-long) courses to be bundled up as a specialization. Course 4, Bayesian Statistics, was added to the sequence at that point, to make this introductory specialization more comprehensive by adding a different point of view for approaching statistical analysis.

Table 1 shows the modules and associated topics for each of the first four courses. Each subsequent course assumes learners have either completed the previous course(s) or have background knowledge equivalent to what those courses covered. Each module is designed to be completed in one week, although learners have the flexibility to extend this if they need to.

Table 1—Modules and Topics for the Five Courses in the Statistics with R Specialization

	**Course 1: Introduction to Probability and Data** starts with a discussion of various sampling methods, and discusses how such methods can affect the scope of inference. It also covers a variety of exploratory data analysis techniques, including using data visualization and summary statistics to explore relationships between two or more variables. Another key learning goal for this course is the use of statistical computing, with R, for hands-on data analysis. The course also features a project on exploratory analysis of data from the Behavioral Risk Factor Surveillance System. The concepts and techniques introduced in this course serve as building blocks for the inference and modeling courses.
	**Course 2: Inferential Statistics** introduces commonly used statistical inference methods for numerical and categorical data. Students learn how to set up and perform hypothesis tests and construct confidence intervals; interpret p-values and confidence bounds; and communicate these results correctly, effectively, and in context without relying on statistical jargon. Building on computing skills they acquired in the previous course, learners conduct these analyses in R. The data analysis project in this courses focuses on inference on data from the Behavioral Risk Factor Surveillance System.
	**Course 3: Linear Regression and Modeling** introduces simple and multiple linear regression. Students learn the fundamental theory behind linear regression and, through data examples, learn to fit, examine, and use regression models to examine relationships between multiple variables. Model fitting and assessment uses R, and a substantial number of examples focus on interpretation and diagnostics for model checking. For the data analysis project in this course, learners conduct exploratory data analysis as well as single and multiple regression on data about movies.
	**Course 4: Bayesian Statistics** introduces learners to the underlying theory and perspective of the Bayesian paradigm and shows end-to-end Bayesian analyses that move from framing the question to building models to eliciting prior probabilities to implementing in R. The course also introduces credible regions, Bayesian comparisons of means and proportions, Bayesian regression and inference using multiple models, and discussion of Bayesian prediction. The data analysis project is Bayesian inference and regression for movies data.
	**Course 5: Statistics with R Capstone** aims to remind learners of goals of earlier courses and expand on these ever so slightly. Learners receive a large and complex data set and the analysis requires them to apply a variety of methods and techniques introduced in the previous courses, including exploratory data analysis through data visualization and numerical summaries, statistical inference, and modeling, as well as interpretations of these results in the context of the data and the research question. Learners are encouraged to implement both frequentist and Bayesian techniques; discuss how these two approaches are similar and different in context of the data; and explain what these differences mean for conclusions that can be drawn from the data.

Components of the Courses

Almost all MOOCs feature videos, and a large majority also have automatically graded, mostly multiple choice, assessment components. I believe a successful and engaging MOOC also must offer more than just videos and assessments: a structure and resources for students needing more guidance (e.g., learning objectives, practice problems) and for those who want to go deeper (e.g., textbook readings, data analysis projects).

Videos

Each module includes seven to 10 videos of roughly four to seven minutes in length. Most of these videos introduce new concepts; the remaining provide additional examples and worked-out problems. The slides that serve as the background in the videos are created in Keynote (Apple's presentation software application) and feature a substantial number of animations such that text, visualizations, and calculations shown on the slides follow the pace of speech in the videos. Many learners have expressed in their course feedback that these features make the videos more engaging and easier to follow compared to videos in many other MOOCs. Sample videos from the Inferential Statistics course are hosted on YouTube at *bit.ly/2LrO6KZ*.

Learning Objectives

Each module also features a set of learning objectives, such as these from the Inferential Statistics course:

- *Explain how the hypothesis testing framework resembles a court trial.*
- *Recognize that in hypothesis testing, we evaluate two competing claims: the null hypothesis, which represents a skeptical perspective or the status quo, and the alternative hypothesis, which represents an alternative under consideration and is often represented by a range of possible parameter values.*
- *Define a* p*-value as the conditional probability of obtaining a sample statistic at least as extreme as the one observed given that the null hypothesis is true:* p*-value = P(observed or more extreme sample statistic* | H_0 *true).*

These learning objectives are constructed using verbs from the revised Bloom's Taxonomy and aim to keep learners organized and focused while watching the videos. The learners are told to have the learning objectives handy while watching the videos and revisit sections of the videos and/or suggested readings for any objectives that they think they have not mastered at the end of the module.

The learning objectives are provided as separate standalone documents, with a few simple conceptual questions after each batch of related learning objectives for learners to check their understanding before moving on.

Suggested Readings and Practice

Suggested readings for the first three courses come from *OpenIntro Statistics*. This book is free and open-source—learners enrolled in the MOOC do not have to purchase another textbook. The readings are optional, since the videos explicitly introduce and cover all required topics for the course, but many learners have reported that they like having a reference book that closely follows the course material. Practice problems are also suggested from the end of chapter exercises in this book.

For the fourth course on Bayesian statistics, readings are suggested from *An Introduction to Bayesian Thinkin*g. This textbook is by the Bayesian Statistics course development team (faculty and PhD students) specifically as a companion to this course and is also freely available on the web.

Computing Labs

Each module in the course features a computing lab in R to give learners hands-on experience with data analysis using modern statistical software, specifically R, as well as provide tools they will need to complete the data analysis projects successfully.

The statistical content of the labs matches the learning objectives of the respective modules it appears in, and the application examples (i.e., data sets and research questions) are primarily from social and life sciences. The labs also make heavy use of an R package, **statsr**, which was designed specifically as a companion for the specialization.

Two other important aspects of the labs are that (1) they use the **tidyverse** syntax and (2) they are completed as reproducible R Markdown reports.

The tidyverse is a language for solving data science challenges with R code. It is an opinionated collection of R packages where the grammar used in each of the packages is optimized for working with data—specifically for wrangling, cleaning, visualizing, and modeling data. The tidyverse syntax for beginners helps learners explore real and interesting data, build informative and appealing visualizations, and draw useful conclusions as much as possible.

Figure 1. Learners of course in a country, divided by the population of the country.

R Markdown provides an easy-to-use authoring framework for combining statistical computing and written analysis in one document. It builds on the idea of literate programming, which emphasizes the use of detailed comments embedded in code to explain exactly what the code was doing. The primary benefit of R Markdown is that it restores the logical connection between statistical computing and statistical analysis by synchronizing these two parts in a single reproducible report. From an instructional perspective, this approach has many advantages: Reports produced using R Markdown present the code and the output in one place, making it easier for learners to learn R and locate the cause of an error, and learners keep their code organized and workspace clean, which is difficult for new learners to achieve if primarily using their R console to run code.

Learners receive each lab in an R Markdown template that they can use as a starting point for their lab reports. Earlier labs in the specialization include a lot of scaffolding, and almost have a fill-in-the-blanks feel to them. As the course progresses, the scaffolding in the templates is removed, and by the end of the first course, learners are able to produce a fully reproducible data analysis project that is much more extensive than any of their labs.

All labs in the specialization are hosted in a publicly available GitHub repository at *github.com/StatsWithR/labs*.

Quizzes

Each module also features two sets of multiple choice quizzes, one formative and one summative. Each question is encoded with feedback that points learners back to relevant learning objectives. Learners can attempt the summative quizzes multiple times with slightly modified versions of the questions.

Data Analysis Projects

Each course ends with a data analysis project (see Table 1), and the specialization wraps up with an extended capstone project. Each student who turns in a project evaluates three other students' work using a peer evaluation rubric. Learners are also strongly encouraged to seek informal feedback on their projects in the course discussion forums. All data analysis projects appearing in the courses in this specialization are hosted in a publicly available GitHub repository at *github.com/StatsWithR/projects*.

Learner Profile

The learners in this specialization come from virtually all over the world. Figure 1 shows the geographic distribution of learners of the course in a country divided by the population of the country. The data only reflect learners enrolled in January 2019, but the distribution has been roughly the same throughout the lifetime of the specialization. We can see that while learners from developed nations have a proportionally larger presence in the course, the course has learners from majority of the emerging nations as well.

Figure 2 shows educational and employment demographics of the learners. Note that these data also only reflect learners enrolled in January 2019. We can see that the majority of learners have an MS or BS degree and a little less than half of the learners are employed full-time. This specialization has a higher

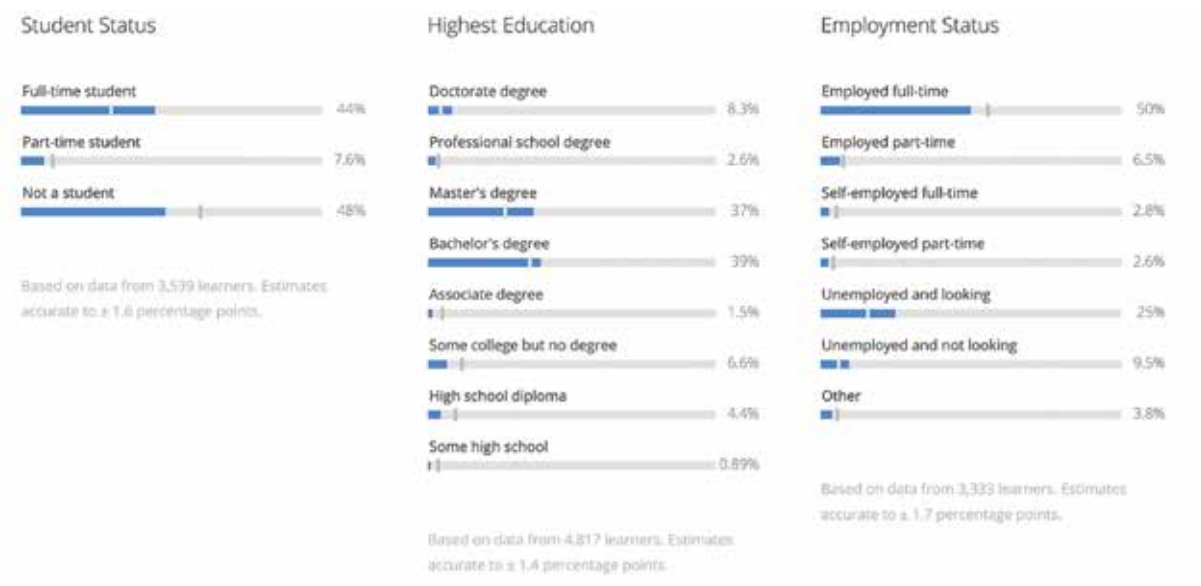


Figure 2. Educational and employment demographics of learners. The gray ticks on the bars show Coursera averages.

proportion of learners who are unemployed and looking for employment than the Coursera average.

Learner Feedback

This specialization routinely gets high ratings from learners and is regularly featured on lists such as 50 Most Popular MOOCs of All Time (*onlinecoursereport.com/the-50-most-popular-moocs-of-all-time*). This is a sampling of stories the learners shared as part of their course feedback.

> *"…I'm a qualitative social science researcher for the Education Development Center in New York City. I joined this class because I wanted to gain quantitative data literacy skills in order to participate more closely in new projects. I'm confident that what I learned in this course has given me a competitive edge in today's market, and I look forward to the following courses in the Statistics with R specialization. Thank you for creating such a thoughtful syllabus and course. I felt supported while engaging in new challenges, and your open source book was an incredibly useful resource(!). It allowed me to explore specific topics of interest in more depth. Thank you! I'll keep coming back to the book throughout my professional career."*

> *"I am an international student from Vietnam. I am working as a business analyst in a tech company. My job requires me to conduct statistical analysis. Learning about statistics throughout this course really helps me in understanding relevant concepts that I can apply into my work. The course instructor approaches every concept with an intuitive manner, making [it easy for students] to understand. The examples are brilliant. Thank you very much for such a high quality course."*

> *"As a scientist in the medical device industry I am always looking for new tools and ways to analyze data. This course was a great way to simultaneously obtain an introduction to R with largely a review of basic statistics. I particularly enjoyed the Bayesian section but really enjoyed all the statistical videos and the follow up with the textbook and problems to reinforce the learning. The course has provided a gateway into R for me to apply to many of my projects in my career. I look forward to the rest of the courses in the specialization, and applying R and the statistics I learned from this course and the future courses."*

Some common themes that emerge in the learner stories are interest in learning R; enjoyment of real data examples; and appreciation of learning resources beyond the videos, such as a free online textbook.

Challenges

There are three main challenges with offering this content on an online platform; two are associated with

access to the labs and the other is associated with the data analysis projects.

Assessment

Given that this is a course with thousands of learners enrolled at any given point, human-grading is simply not feasible. The lab assessments are set up as multiple-choice questions. Learners complete the lab exercises by generating R Markdown reports in which they analyze a dataset. Then, they answer a series of multiple-choice questions about the data analysis results. The challenge is that the multiple-choice questions do not assess the full spectrum of the skills we want learners to acquire via these labs—they assess whether learners can obtain the correct results using R, but they do not assess mastery of R syntax, reproducibility of their analysis, or related skills.

In addition, autograding is not feasible for open-ended data analysis projects, so peer evaluation is the only solution for grading these projects. Even with a very detailed rubric, consistency in grading is difficult to attain, and it is challenging for learners who are just learning the material themselves to evaluate others' work. Variability in quality and depth of feedback provided also can leave learners frustrated. The option to share their projects on the discussion forums and get feedback can be helpful for some learners, but others are not so keen about publicly sharing their projects.

Computing Infrastructure

Our preferred method for getting students with no computing background started with R is a cloud-based access to RStudio to avoid challenges in local installation and to provide a uniform computing environment for all learners. However, it is not feasible to offer a centralized cloud-based solution to all learners enrolled in a MOOC, so students have to install R and RStudio locally, along with the correct versions of all packages they use in the labs.

As a partial solution to this challenge, we offer students the option to complete the labs in the first course of the specialization on DataCamp (*datacamp.com*), an online learning platform that provides in-browser access to RStudio. This helps students struggling with software installation issues early on in the course to get started with data analysis; they can go back to tackling software challenges once they feel more confident with R.

Community-building

Student interaction, or the lack thereof, is often a major challenge in online courses, but in MOOCs, the discussion forums can serve as a major strength of the course. Thousands of students are enrolled in the course at any point; if even a small percentage chooses to browse the discussion forums, and an even smaller percentage interacts with other learners on the course discussion forums, this still results in a large number of learners interacting with each other.

Over the years of the MOOC being offered, a handful of very knowledgeable and helpful course mentors emerged from the discussion forums. These are learners who took the courses at some point and now volunteer their time to answer student questions and provide direction for new learners. ◘

Further Reading

Pappano, Laura. 2012. The Year of the MOOC. *New York Times* February 12, 2012.

Kolowich, Steve. 2014. The year the media stopped caring about MOOCs. *Chronicle of Higher Education*.

Anderson, Lorin W., et al. 2001. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*, abridged. White Plains, NY: Longman.

Diez, David, et al. 2015. OpenIntro Statistics. OpenIntro. *openintro.org*.

Clyde, Merlise, et al. 2018. An Introduction to Bayesian Thinking, 1st ed. GitHub. *statswithr.github.io/book*.

Robinson, D. 2017. Teach the tidyverse to beginners, July 5, 2017. *varianceexplained.org/r/teach-tidyverse*.

Xie, Yihui, Allaire, J.J., and Grolemund, Garrett. 2018. R Markdown: *The Definitive Guide*. CRC Press.

Knuth, Donald Ervin. 1984. Literate programming. *The Computer Journal* 27:2(97–111).

Çetinkaya-Rundel, Mine, and Rundel, Colin. 2018. Infrastructure and tools for teaching computing throughout the statistical curriculum. *The American Statistician* 72:1(58–65).

About the Author

Mine Çetinkaya-Rundel, a statistician from Turkey, is an associate professor of the practice at Duke University and a professional educator at RStudio.

Nicole Lazar
Column Editor

Crowdsourcing Your Way to Big Data

"Crowdsourcing" is one of those buzzwords that seem to be all around us in this Big Data (another buzzword!) era. A formal definition of the word, according to an online dictionary, is "the practice of obtaining information or input into a task or project by enlisting the services of a large number of people, typically via the Internet."

Crowdsourcing in the data realm has three primary goals: data collection, data cleaning, and data analysis or processing. The first uses the Internet as a tool to gather data on or from many people, typically widely dispersed, in a short period of time. The second and third involve people around the world helping scientists clean and analyze their data by each taking on small parts of the work; cumulatively, data can be cleaned quickly, for example, if many people around the world, in their spare time, are each working on one small part of the data set. There can also be an aspect of citizen science at play here.

Perhaps one of the most-familiar examples of crowdsourcing for data collection and processing is Amazon's Mechanical Turk, known as MTurk for short. The original MTurk was a chess-playing machine built in the late 18th century. The Turk was displayed across Europe and played many humans to defeat. It was later revealed that the Mechanical Turk was no automaton, but rather a fake; a human player in fact made the moves. It's only in more-recent years that computers have actually been able to defeat humans in games such as chess and Go.

Amazon's Mechanical Turk takes its inspiration and its name from that 250-year-old story. According to the FAQ at the MTurk website, the main idea is that humans are better at many tasks than are computers. Pooling together the time, resources, and abilities of many people both relieves the burden on each individual worker and provides researchers and businesses access to a large, diffuse "workforce" for particular projects.

The MTurk project is not without controversy and detractors. When I typed "is amazon mechanical turk" into Google, the first few auto-complete suggestions were: "safe," "worth it," "free," and "legitimate." Its value as a research tool—particularly for conducting surveys—has been questioned. A few academic papers have been written about the use of Mechanical Turk for behavioral research, and UC Berkeley's Committee for Protection of Human Subjects has issued guidelines for using MTurk in online research.

Confidentiality and anonymity of MTurk Workers (as the human participants are called) is one area of research and concern. As statisticians, we also can easily recognize that there will be problems with bias in responses. People who choose to be MTurk Workers, whether in their spare time for extra cash or on a more-regular basis, are not a random sample of (just about) any population of interest. This fact alone will skew survey or other results obtained via this crowdsourcing tool. On the other hand, since Workers can come from any country and may have a wide range of backgrounds, they are not the "typical" undergraduate student subjects, who come with their own biases.

The website *www.mturktracker.com* makes it easy to form a snapshot image of MTurk Workers. Browsing statistics for December 2018, I found that overall, about half are male and half female, although intriguingly, the balance shifts over the course of the day and over the day of the week. The site gives demographics separately for the USA and India, where the vast majority of Workers live, and it is possible to dig down into these differences by country as well. Most workers are relatively young, born after 1980, although a fair proportion were born between 1960 and 1979 as well, and a few were even born after 2000.

Workers tend to live either alone or with one other person—almost 50% of December

Workers fall into those two categories combined, although this is more the case in the USA than in India. (There is much more you can find here—if you are interested, follow the link provided to explore on your own.)

I was fascinated to discover, in researching this column, that there is a flourishing academic field of study focused on Amazon's MTurk, analyzing the Workers, the Requesters (those who have work that they wish to crowdsource), their interactions, and more. The Pew Research Center, for example, put out a report in July 2016, titled "Research in the Crowdsourcing Age, A Case Study." It examines characteristics of the Workers—an expanded version of the brief "data dive" I carried out using *www.mturk-tracker.com*; considers the Requesters, who are mostly from business and academia; and looks at the types of tasks being requested—not all of which involve data collection as we might typically think of it.

While MTurk is a major player in data crowdsourcing, it is important to keep in mind that there are many other facets as well, such as the citizen science movement mentioned at the start of this column. Organizations such as SciStarter (*https://scistarter.com*), the Citizen Science Alliance (*https://www.citizensciencealliance.org*), and the Citizen Science Association (*https://www.citizenscience.org*) all help to pair researchers who have projects with private individuals worldwide who want to be involved in the process of collecting and analyzing their data, or simply with participating in the scientific endeavor in some way.

This is, again, a crowdsourcing model, because researchers pull on the time and efforts of a physically diffused "workforce" rather than having all team members centrally located in a lab. Unlike with MTurk, however, citizen scientists are most-often volunteers.

The nature of the projects may differ somewhat as well. MTurk appears to be widely used in the social and behavioral sciences (surveys are one of the popular requests for Workers), while more-general citizen science projects may cover a wider range of disciplines, including the hard sciences. Some of the same concerns about quality of data arise, as is inevitable.

It is also important to have common protocols that all citizen scientists involved in a project will follow, to ensure compatibility and enhance reproducibility. This is made easier by modern technology; for example, the Marine Debris Tracker is a mobile app that allows users to report on different types of debris they encounter in the water and on beaches. Collectively, there are more data than any one researcher (or group of researchers) would be able to assemble in a reasonable amount of time.

Astronomy is another area where citizen scientists have made and continue to make useful contributions. For most, the barriers to participate in citizen science are low, since we are now connected around the world via phones that can operate as data-gathering, and data-creating, machines.

Questions of ethics arise, and these have been addressed both by the Citizen Science Association and European Citizen Science Association, the latter of which developed guidelines for best practices in citizen science (for the complete list, see *https://ecsa.citizen-science.net/sites/default/files/ecsa_ten_principles_of_citizen_science.pdf*). Pertinent to our discussion: Several of the guidelines mention the collection, use, and analysis of data.

Once I started exploring the topic of crowdsourcing as a tool of—and for—Big Data, I realized that there is so much out there; more than I can cover in the time and space I have available to write this month, so I do plan to revisit the theme in a future column. For now, I will touch on a few more-interesting tidbits.

At the University of California, Berkeley, the AMPLab does research on a wide range of interesting Big Data applications. "AMP" stands for "Algorithms" (in the form of Machine Learning), "Machines" (cloud computing), and "People" (crowdsourcing) coming together to untangle the intricacies of big data. A project called CrowdER uses crowdsourcing for data cleaning. It's a well-worn truism (or is it a cliché?) that the majority of a data scientist's time is spent on cleaning the data (by some counts I've heard, up to 80%), and only a small portion in the actual analysis. CrowdER aims to streamline one part of the cleaning step, thereby freeing up resources, particularly time, for analysis.

The specific problem that CrowdER addresses is "entity resolution." This is the question of deciding whether two different records in a data set, or across data sets, refer to the same observational unit or not. For humans, this is not a difficult task, although it may be time-consuming for either humans or computers.

For example, if I have a data record for "New Orleans, LA" and another for "New Orleans" and yet a third for "NOLA," I will quickly conclude that these all refer to the same place; a computer might not be able to do that. Performing all pairwise comparisons of records is inefficient; however, in most scenarios, it is reasonable to assume that the majority of records are, in fact, unique. Intervention is needed only where there is ambiguity or question.

That means one should prune out the clear cases using, for example, machine learning techniques and crowdsource the rest. This machine-human hybrid takes advantage of the computer's ability to sort through the records quickly and the human's real-world knowledge and ability to spot patterns.

How about data analysis? I turn now to MIT, where researchers have developed a tool called "FeatureHub." FeatureHub crowdsources feature extraction for prediction models in large data sets. Researchers—statisticians or domain area experts, for instance—can log on to FeatureHub, review a data set of interest, and propose features to be used in the analysis. Machine learning and statistical model-building techniques are then used for the actual training and testing of the models based on the proposed features. Researchers can work independently, but there is also an opportunity to collaborate with others.

In initial evaluations of the FeatureHub framework, the developers compared results to those obtained by modelers working on Kaggle. The predictive abilities were quite similar: FeatureHub wouldn't win the Kaggle competitions, but does succeed in extracting much of the useful information. However, as the developers note, while the feature extraction in the typical analysis of a Kaggle data set may take weeks, FeatureHub accomplished the same in a matter of days.

Crowdsourcing, in short, is showing its potential as a useful tool. There are, of course, pitfalls to be aware of, particularly biases inherent in crowdsourced data collection via surveys. It's intriguing, to me at least, to think about the ways that we can enrich our data landscape by harnessing our human strengths and skills together, and pairing those with tasks that are better left to computers as too tedious for us.

Do you know of interesting uses of crowdsourcing? If you do, I'd love to hear from you. C

Further Reading

www.mturk-tracker.com. 2019. *https://pewrsr.ch/29xWLPo*.

Mason, W., and Suri, S. 2012. Conducting behavioral research on Amazon's Mechanical Turk. Behavior Research Methods 44:1–23; available at *https://link.springer.com/article/10.3758%2Fs13428-011-0124-6*.

Smith, M.J., Wedge, R., and Veeramachaneni, K. 2017. FeatureHub: Towards collaborative data science. *IEEE International Conference on Data Science and Advanced Analytics*. DOI: 10.1109/DSAA.2017.66.

About the Author

Nicole Lazar, who writes *The Big Picture*, earned her PhD from the University of Chicago. She is a professor in the Department of Statistics at the University of Georgia, and her research interests include the statistical analysis of neuroimaging data, empirical and other likelihood methods, and data visualization. She also is an associate editor of *The American Statistician* and *The Annals of Applied Statistics* and author of *The Statistical Analysis of Functional MRI Data*.

STATtr@k

A WEBSITE FOR NEW STATISTICS PROFESSIONALS

STAT*tr@k* offers tips on:

- how to apply for a job
- how to be a successful graduate student
- how to make the transition from coursework to research

STAT*tr@k* also offers information about career and mentorship sites, upcoming conferences, and awards and competitions. New articles will appear monthly.

Check it out *at STATtrak.amstat.org*!

Christian Robert
Column Editor

The Beauty of Mathematics in Computer Science

Jun Wu

Softcover: 284 pages

Year: 2018

Publisher: Chapman and Hall/CRC Press

ISBN-13: 978-1138049673

This book by Jun Wu, "staff research scientist in Google who invented Google's Chinese, Japanese, and Korean web search algorithms," was translated from the Chinese originals from Google blog entries (meaning most references are pre-2010). A large part of the book is about word processing and web navigation, which is the author's research specialty. It is not so much about mathematics (when rereading the first chapters to write this review, I realized why the part about language processing in *AIQ*, reviewed below, sounded familiar: I had read it in *The Beauty of Mathematics in Computer Science*).

> *"One of my main objectives for writing the book is to introduce some mathematical knowledge related to the IT industry to people who do not work in the industry."*

In the first chapter, which is about the history of languages, I found out, among other things, that ancient Jewish copyists of the Bible had an error-correcting algorithm consisting of giving each character a numerical equivalent; summing up each row, then all rows; and checking that the sum at the end of the page was the original one. The second chapter explains why the early attempts at language computer processing, based on grammar rules, were unsuccessful and how a statistical approach broke the blockade, explained via Markov chains in the following chapter, along with the Good-Turing estimate of the transition probabilities. Next comes a short and low-tech chapter on word segmentation, and then an introduction to hidden Markov models, mentioning the Baum-Welch algorithm as a special case of EM, which makes a return by Chapter 26. There is also a chapter about entropies and Kullback-Leibler divergence.

A first intermission is provided by a chapter dedicated to the late Frederick Jelinek, the author's mentor (including what I find to be a rather unfortunate and unnecessary equivalent drawn between the Nazi and Communist eras in Czechoslovakia, p. 64). This chapter sounds a wee bit too much like an extended obituary.

The next block of chapters is about search engines, with a few pages about Boolean logic, dynamic programming, graph theory, Google's PageRank, and TF-IDF (term frequency/inverse document frequency). Unsurprisingly, given that the entries were originally written for Google's blog, Google's tools and concepts keep popping up throughout the entire book.

Another intermede is about Amit Singhal, designer of Google's internal search ranking system, Ascorer (with another unfortunate equivalent drawn, this time with the AK-47 Kalashnikov rifle as "elegantly simple," "effective, reliable, uncomplicated, and easy to implement or operate," p. 105). Even though I do get the reason for the analogy, using an equivalent tool whose purpose is not to kill other people would have been just decent…

Chapters follow about measuring proximity between news articles by vectors in a 64,000 dimension vocabulary space and their angle; singular value decomposition; and turning URLs as long integers into 16 byte random numbers by the Mersenne Twister (making me wonder why random, except

Preliminary versions of these reviews were posted on *xianblog.wordpress.com.*

for encryption?), missing both the square in von Neumann's first PRNG (p. 124) and the opportunity to link the probability of overlap with the birthday problem (p. 129). These are followed by another chapter about cryptography, always a favorite in math vulgarization books (but with no mention made of the originators of public key cryptography, like James Hellis or the RSA trio, or of the impact of quantum computers on the reliability of these methods). Then there is a mathematic chapter about spam detection.

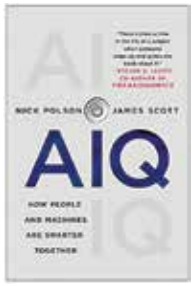
Another group of chapters covers maximum entropy models (in a rather incomprehensible way, I think; see p. 159), and continues with an interesting argument about how Shannon's first theorem predicts that it should be faster to type Chinese characters than roman characters. This is followed by the Bloom filter, which operates as an approximate Poisson variate. In Bayesian networks, the "probability of any node [are said to be] computed by Bayes' formula" (not really), with a slightly more-advanced discussion on providing the highest posterior probability network. In a chapter about conditional random fields, the conditioning is not clearly discussed (p. 192). Next are chapters about Viterbi's algorithm (and successful career) and the EM algorithm, nicknamed "God's algorithm" in the book (Chapter 26), although I never heard of this (should I repeat? unfortunate) nickname previously.

The final two chapters are about neural networks and Big Data, and were clearly written later than the rest of the book, with the predictable illustration of AlphaGo (but without technical details). The 20-page chapter about Big Data does not contain a larger amount of mathematics, with no equation apart from Chebyshev's inequality and a frequency estimate for a conditional probability. But I learned about 23&me running genetic tests at a loss to build a huge (if biased) genetic database. (The bias in "Big Data" issues is not covered by this chapter.)

To conclude, I found the book to provide a fairly interesting insight by a senior scientist at Google into his vision of the field and his job experience, with loads of anecdotes and some historical background, but very Google-centric with what I felt was an excessive amount of name-dropping and of "I did, I solved, I…," etc. The title is rather misleading in my opinion, since the amount of math is incredibly limited and rarely sufficient to connect with the subject at hand. Although this is quite a relative concept, I did not spot beauty therein but rather technical advances and tricks that allowed the author and Google to beat the competition.

AIQ

Nick Polson and James Scott



Hardcover: 272 pages

Year: 2018

Publisher: St. Martin's Press

ISBN-13: 978-1250182159

AIQ was my Christmas Day read, which I mostly perused while the rest of the household was still sleeping. The book, written by two (mainstream) Bayesians, Nick Polson and James Scott, was published before the ISBA meeting last year in Edinburgh, but I only bought it recently as a holiday present. Now that the holidays are over, I recall this as a pleasant book to read, especially while drinking spiced tea by the wood fire. It is well-written and full of facts and anecdotes I did not know or had forgotten. Intended for a general audience, it is also quite light from a technical side, rather obviously, but also from a philosophical side. While strongly positivist about the potential of AIs for the general good, it cannot be seen as an antidote to the doomlike and philosophical *Superintelligence* by Nick Bostrom or the more-factual *Weapons of Math Destruction* by Cathy O'Neal (both previously reviewed in this *CHANCE* column).

Indeed, I find the book quite benevolent and maybe a wee bit too rosy in its assessment of AIs. The discussion of how Facebook allowed hidden Russian intervention in the U.S. presidential election of 2017, which may have significantly contributed to turning the White House orange, is missing, in my opinion, the viral nature of the game and its more-frightening impact, when endless loops of highly targeted posts can cut people off from the most basic common sense.

While the authors are "optimistic that, given the chance, people can be smart enough," I worry deeply about this, from the sheer fact that the hoax that Hillary Clinton was involved in a child sex ring was ever considered seriously by real people, to the point of someone shooting a real weapon at the pizza restaurant supposedly involved, if thankfully hurting no one. Hence, I am much less optimistic than the authors about the ability of a large-enough portion of the population, not even the majority, to keep a

critical distance from the extremist messages planted by AI-driven media.

Similarly, while Nick and James point out (rather late in the book) that Big Data (meaning large data) are not necessarily good data for being unrepresentative of the population at large, i.e., biased, they do not propose any highly convincing solutions for battling this bias in existing or incoming AIs. This leads to a global pessimism that AIs may do well (enough) for a majority of the population and still discriminate against a minority by the same reasoning. As described in Cathy O'Neal's book and elsewhere, proprietary software does not even have to explain why it discriminates.

More globally, the business school environment of the authors may have prevented them from expressing worry about the massive power grab operated by AI-based companies, which grow genetically with little interest in democracy and states, as shown (again) by the recent U.S. presidential election or by their systematic fiscal optimization policies. Or by the massive recourse to machine learning by Chinese authorities toward a dreadfully real social credit system grade for all citizens.

> *"La rage de vouloir conclure est une des manies les plus funestes et les plus stériles qui appartiennent à l'humanité…Chaque religion et chaque philosophie a prétendu avoir Dieu à elle, toiser l'infini et connaître la recette du bonheur."*
> —Gustave Flaubert

As for choice morsels, I did not know about Henrietta Leavitt's prediction rule for pulsating stars being behind Hubble's discovery, which makes the story sound like an astronomy duel with Rosalind Franklin's (ignored) DNA contribution. The use of Bayes' rule for locating lost vessels is also found in *The Theorem that Would Not Die*, while I would have also discussed its failure in locating the more-recent Malaysia Airlines Flight 370 disappearance. Similarly, I had never heard the great expression of "model rust." Nor, even more surprisingly, the quote above from the great novelist Flaubert.

As suggested above, there was a sense of "déja vu" in the story about how a 180º switch in perspective on language understanding by machines brought the massive improvement that we witness today. I could not remember where until I wrote the review of *The Beauty of Mathematics in Computer Science*. I have also read about Newton missing the boat on the precision of the coinage accuracy (was it in Bryson's book on the Royal Society, *Seeing Further*?), but with less-neutral views of the role of Newton in the matter, since the Laplace of England would have benefited from keeping the lax measures of assessment.

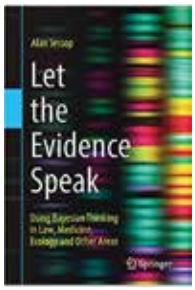
It was great to see familiar figures like Luke Bornn and Katherine Heller appearing in some pages—Luke for his work on the statistical analysis of basketball games, Katherine for her work on predictive analytics in medicine—and reflecting on the missed opportunities represented by the accumulation of data on any patient throughout their life that is as grossly ignored nowadays as it was in Nightingale's time.

The message of the chapter "The Lady with the Lamp" may again be somewhat over-optimistic: While AI and health companies see clear incentives for developing more encompassing prediction and diagnostic techniques, this will only benefit patients who can afford the ensuing care. Given the state of healthcare systems in the most-developed countries, this is a decreasing proportion; not to mention the less-developed countries.

Overall, a great read for the general public, dedramatizing the rise of the machines and mixing statistics and machine learning to explain the (human) intelligence behind AIs. Nothing on the technical side, to be sure, but this was never the intention of the authors. ■

Let the Evidence Speak

Alan Jessop



Softcover: 232 pages

Year: 2018

Publisher: Springer

ISBN-13: 978-3319713915

This book by Alan Jessop, a professor at the Durham University Business School in Great Britain, aims at presenting Bayesian ideas and methods for decision-making "without formula because they are not necessary; the ability to add and multiply is all that is needed." The trick employed by the author is in using a Bayes grid; in other words, a two-by-two table. (A few formulas there survived the slaughter, such as the

formula for the entropy on p. 91, in a chapter about information that I find definitely unclear). When leaving the 2x2 world, things become more complicated and the construction of a prior belief as a probability density gets heroic without the availability of math formulas.

The first part of the book is about Likelihood, albeit not the likelihood function, despite having the general rule that (p. 73)

> belief is proportional to base rate x likelihood,

which is the book's version of Bayes's (base?!) theorem. It then goes on to discuss the less-structured nature of prior (or prior beliefs) against likelihood by describing Tony O'Hagan's way of scaling experts' beliefs in terms of a Beta distribution and mentioning Jaynes's maximum entropy prior again, without admitting the help of a single formula. What is hard to fathom from the text is how can one derive the likelihood outside surveys. (Using the illustration of 1963 Oswald's murder by Ruby in the likelihood chapter does not particularly help.)

A bit of nitpicking at this stage: the sentence

> *"The ancient Greeks, and before them the Chinese and the Aztecs…"*

is historically incorrect since, while the Chinese empire dates back before the Greek dark ages, the Aztecs only ruled Mexico from the 14th century (AD) until the Spaniard invasion.

While most of the book sticks with unidimensional parameters, it also discusses more-complex structures, for which it relies on Monte Carlo, although the description is rather cryptic (more like "Use your spreadsheet!"; p. 133). The book at this stage turns to a storytelling mode, by considering, for instance, the Federalist papers analysis by Mosteller and Wallace. The reader can only follow the general process of assessing a document authorship for a single word, since multidimensional cases (for either data or parameters) are clearly out of reach. The same comment applies to the ecology, archeology, and psychology chapters that follow.

The intermediary chapter on the "grossly misleading" (court wording) of the statistical evidence in the Sally Clark prosecution is more accessible in that, again, it relies on a single number.

Returning to the ban of the Bayes rule by British courts:

> *"In the light of the strong criticism by this court in the 1990s of using Bayes theorem before the jury in cases where there was no reliable statistical evidence, the practice of using a Bayesian approach and likelihood ratios to formulate opinions placed before a jury without that process being disclosed and debated in court is contrary to principles of open justice."*

the discussion found in the book is quite moderate and inclusive, in that a Bayesian analysis helps in gathering evidence about a case, but may be misunderstood or misused at the [non-Bayesian] decision level.

Let the Evidence Speak is an interesting introduction to Bayesian thinking through a simplifying device, the Bayes grid, which seems to come from management, with a large number of examples, if not necessarily all realistic, and some side stories. I doubt this exposure can produce expert practitioners, but it makes for a worthwhile awakening of someone "likely to have read this book because [one] had heard of Bayes but [was] uncertain what it was" (p. 222). With commendable caution and warnings along the way. ◘

Is that a big number?

Andrew Elliot

Hardcover: 352 pages

Year: 2018

Publisher: Oxford University Press

The overall aim of this book by Andrew Elliott is to encourage numeracy (or fight innumeracy) by making (better) sense of absolute quantities by putting them in perspective, teaching about log scales, visualization, and divide-and-conquer techniques—and providing a massive list of examples and comparisons, sometimes for page after page…The book is associated with a

fairly rich website, itself linked to the many blogs by the author and a myriad other links and items of information (among which I learned of the recent and absurd launch of Elon Musk's Tesla car in space. A premiere in garbage dumping…). From what I can gather from these sites, some (most?) of the material in the book seems to have emerged from the various blog entries.

> *"Length of River Thames (386 km) is 2 x length of the Suez Canal (193.3 km)"*

Maybe I was too exhausted by English heat and a very busy week in Warwick for our computational statistics week, the football 2018 World Cup having nothing to do with this, but I could not keep reading the chapters of the book in a continuous manner. I was suffering from massive information over-dump! Being given thousands of entries kills, for me, the appeal of putting weight or sense to large, very large, and humongous quantities.

The final vignette in each chapter, of pairing of numbers like the one above or this one below—

> *"Time since earliest writing (5200 y) is 25 x time since birth of Darwin (208 y)"*

—only evokes the remote memory of some children's journal I read from time to time (as a kid) with this type of entries. Maybe it was a journal I would browse while waiting at the hairdresser's (which brings back memories of endless waits…)

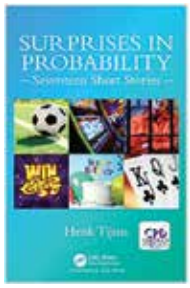
Some of the background about measurement and other curios carry a sense of Wikipediesque absolute in their absurdly minute details.

A last item of disappointment about the book is its poor graphical design or support. While the author insists on the importance of visualization for grasping the scales of large quantities, and the webpage is full of such entries, there is very little backup or great graphs to be found in Is that a big number? Some of the pictures seem taken from an anonymous databank (where are the towers of San Geminiano?) and there are not enough graphics. For instance, the fantastic graphics of xkcd conveying the meaning of different orders of magnitude, such as the superb xkcd money chart poster, would have been an excellent addition. Or the one about the future. Or many, many others…

While the style is sometimes light and funny, an overall impression of dryness remains. I much more preferred Kaiser Fung's *Numbers Rule Your World: The Hidden Influence of Probabilities and Statistics on Everything You Do* and even more, both *Guesstimation* books.

Surprises in Probability–Seventeen Short Stories

Henk Tijms



Softcover: 144 pages

Year: 2018

Publisher: Chapman and Hall/CRC Press

ISBN-13: 978-0367000820

A very short book (128 pages, but at a very high price!) is Henk Tijms' *Surprises in Probability–Seventeen Short Stories*. Tijms is an emeritus professor of econometrics at the Vrije University in Amsterdam and he wrote these 17 pieces for either the Dutch Statistical Society magazine or a blog he ran for the *New York Times*.

The author mentions that the book can be useful for teachers and, indeed, this is a collection of surprising probability results—surprising in the sense that the numerical probabilities are not necessarily intuitive. Most illustrations involve betting of one sort or another, with only basic (combinatorial) probability distributions involved. Readers should not even worry about this basic probability background, since most statements are exposed without a proof.

Most examples are classical, from the prisoner's problem to the Monty Hall paradox, to the birthday problem, Benford's distribution of digits, the gambler's ruin, the gambler's fallacy, the St. Petersbourg paradox, the secretary's problem, and stopping rules. The most-advanced notion is the one of (finite state) Markov chains, since martingales are only mentioned in connection with pseudo-probabilist schemes for winning the lottery—for which (our very own!) Jeff Rosenthal makes an appearance, thanks to his uncovering of the Ontario Lottery scam.

> *"In no other branch of mathematics is it so easy for experts to blunder as in probability theory."*
> —Martin Gardner

A few stories have entries about Bayesian statistics, with mentions made of the O.J. Simpson, Sally Clark, and Lucia de Berk miscarriages of justice, although these mentions make the connection most tenuous. Simulation is also mentioned as a manner of achieving approximations to more-complex probabilities,

but not to the point of discussing surprises about simulation, which could have been the case with the simulation of rare events.

Story 10 features 10 beautiful probability formulas and reminded me of Ian Stewart's *17 Equations that Changed the World*; obviously on another scale and in a much less-convincing way—the Normal (or Gauss) density, Bayes' formula, gambler's ruin formula, squared-root formula (meaning standard deviation decreases as $\sqrt{n}$), Kelly's betting formula (?), asymptotic law of distribution of prime numbers (?), another squared-root formula for the one-dimensional random walk, newsboy formula (?), Pollaczek-Khintchine formula (?), and waiting-time formula. I am not sure I would have included any of these, even those I was aware of…

All in all, this small book is a nice if unsurprising database for illustrations and possibly for exercises in elementary probability courses, although it will require some work from the instructor to link the statements to their proof, as one would expect from blog entries. It makes for nice reading, especially while traveling. I hope some fellow traveler will pick up the book from where I left it in the Mexico City airport.

About the Author

Christian Robert is a professor of statistics at the universities of Paris-Dauphine, France, and Warwick, UK. He has written extensively about Bayesian statistics and computational methods, including the books *The Bayesian Choice* and *Monte Carlo Statistical Methods*. He has served as president of the International Society for Bayesian Analysis and editor in chief of the *Journal of the Royal Statistical Society (Series B)*, and currently is deputy editor of *Biometrika*.

Howard Wainer
Column Editor

A Statistician Reads the Obituaries: On the Relative Effects of Race and Crime on Employment

I was saddened to read an obituary in the November 9, 2018, *New York Times* for Devah Pager, the Malkin Professor of Public Policy at Harvard University. Professor Pager was but 46 years of age; an accomplished and well-known sociologist, whose interests included the effects of race and prior felony convictions on finding a job. Her 2007 book, *Marked*, was about this topic and, among its many honors, received the Book of the Year Award from the Association for Humanist Sociology.

Katherine Q. Seelye, author of the *Times* obituary, described one field experiment that Professor Pager carried out in which she recruited two teams of "young, well-groomed, well-spoken college men of the same height—one team black, and the other white—and gave them identical résumés as they applied for 350 entry-level jobs in Milwaukee. The applicants took turns saying that they had served an 18-month sentence for cocaine possession."

What she found was that blacks who said they had a criminal record had a callback rate of 5%, while those who did not were called back 14% of the time. For the white students, the callback rates were 17% for those who said they had a criminal record and 34% for those who said they did not.

At this point in the narrative, the difference between an obituary writer and a statistician manifests itself. The statistician (or at least this statistician) constructs a two-way table:

	Proportions	
	Black	White
Felon	0.05	0.17
Not Felon	0.14	0.34

It is almost surely impossible for any statistician. when faced with a table like this, not to compute row and column means to get some sense of the relative size of the effects of the two orthogonal factors. Obviously, I succumbed.

	Black	White	Mean	**Felon Effects**
Felon	0.05	0.17	0.11	**–0.07**
Not Felon	0.14	0.34	0.24	**0.07**
Mean	0.10	0.26	**0.18**	
Race Effects	**–0.08**	**0.08**		

I centered the row and column means by subtracting out 0.18, the grand mean. This yielded the row and column effects and I noticed that the race effects were a little larger than the Felony effects!

This sort of decomposition reflects a simple additive model:

callback likelihood = race effect + felony effect + grand mean (1)

But is it this simple? Is there an interaction? Does being both black and a felon yield a worse outcome than their simple sum would suggest? To examine this, I calculated the fit of the model in (1):

Fit

	Black	White
Felon	0.03	0.19
Not Felon	0.16	0.32

… and then subtracted the fit from the original data, to yield the residuals:

Residuals

	Black	White
Felon	0.02	– 0.02
Not Felon	– 0.02	0.02

…showing that there are smallish residuals but not in the direction one might have feared (e.g., being a white felon lowers the chance of a callback), but the big story is that being black was (marginally) worse than being a felon in terms of the likelihood of getting a callback.

Pager emphasized these results (see, for example, her 2009 paper). Her genius was in using prior felony convictions to provide a context for the effects of racism. This paralleled the approach taken by David Strayer and his colleagues in their widely cited 2006 paper, in which they made the evocative comparison between driving while intoxicated and driving while on the phone.

My goals in this note were two-fold: first, to illustrate how little statistical technology is required to decipher two-way tables of important and interesting data, and second, how much one can learn from a thoughtful obituary of an eminent scientist.

Further Reading

Pager, Devah. 2007. *Marked: Race, Crime, and Finding Work in an Era of Mass Incarceration*. Chicago, IL: University of Chicago Press.

Pager, D., Western, B., and Bonikowski, B. (2009). Discrimination in a low-wage labor market: A field experiment. *American Sociological Review* 74:777–799.

Strayer, D.L., Drews, F.A., and Crouch, D.J. (2006). A comparison of the cell phone driver and the drunk driver. *Human Factors* 48(2):381–391.

About the Author

Howard Wainer is a statistician and author who lives in Pennington, NJ. He has written this column since 1990.

Women in Statistics and Data Science conference

October 3-5, 2019 | Bellevue, WA

WELCOMING. **INSPIRING.** EMPOWERING. **MOTIVATING.** EYE-OPENING. **YOU.**

Although #WSDS2018 was my first WSDS conference, the genuineness & camaraderie... to say the least...was simply mindblowing. I am so thankful for the conference's existence and I will attend every chance I get!"

- Lakisha Armstrong, *President, ASA Maryland Chapter*

Formal sessions and informal networking opportunities will empower and challenge women statisticians and data scientists to:

Share knowledge

Build community

Grow influence

REGISTER

Early Registration Closes
August 15, 2019

Hotel Reservations Close
September 9, 2019

Learn more and get involved in WSDS 2019 at *ww2.amstat.org/wsds*.

DENVER, COLORADO
JSM2019

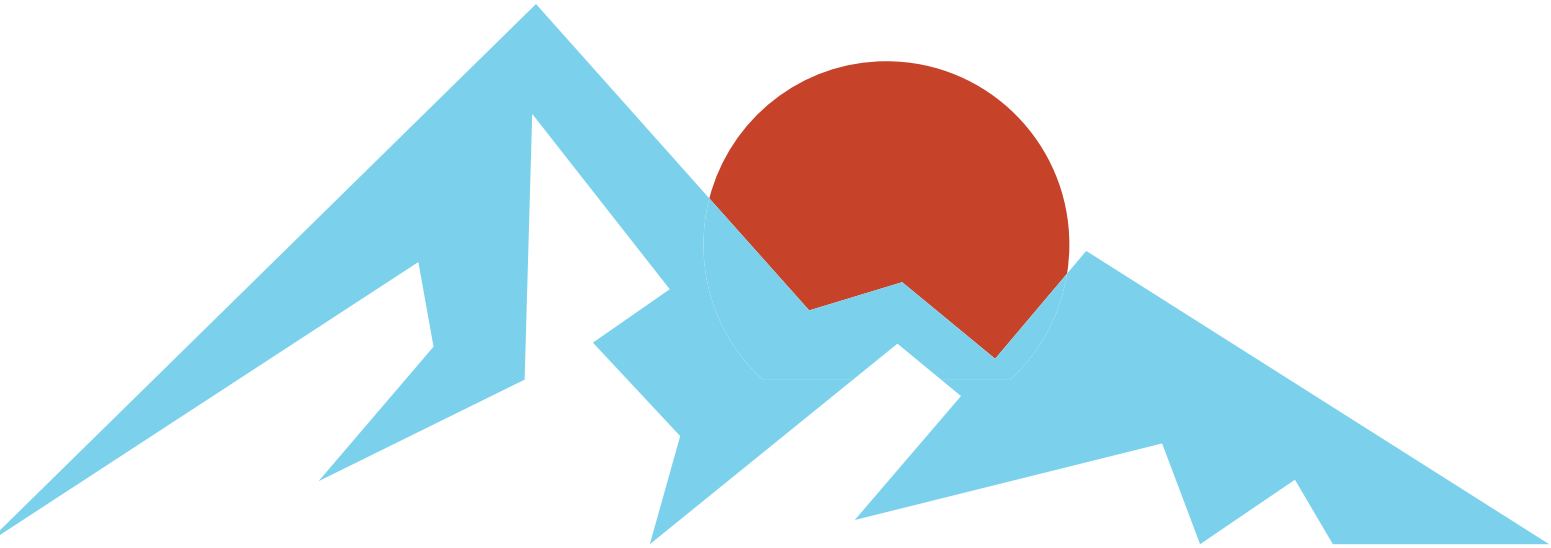
KEY DATES

June 29, 2019
Regular Registration Closes

June 30 – August 1
Late Registration

July 3
Housing Deadline

July 27–August 1
2019 JOINT STATISTICAL MEETINGS



ww2.amstat.org/meetings/jsm/2019

THIS IS STATISTICS

HELP US RECRUIT THE NEXT GENERATION OF STATISTICIANS

The field of statistics is growing fast. Jobs are plentiful, opportunities are exciting, and salaries are high. So what's keeping more kids from entering the field?

Many just don't know about statistics. But the ASA is working to change that, and here's how you can help:

- Send your students to ***www.ThisIsStatistics.org*** and use its resources in your classroom. It's all about the profession of statistics.
- Download a handout for your students about careers in statistics at ***www.ThisIsStatistics.org/educators***.

If you're on social media, connect with us at ***www.Facebook.com/ThisIsStats*** and ***www.Twitter.com/ThisIsStats***. Encourage your students to connect with us, as well.

Site features:

- Videos of young statisticians passionate about their work
- A myth-busting quiz about statistics
- Photos of cool careers in statistics, like a NASA biostatistician and a wildlife statistician
- Colorful graphics displaying salary and job growth data
- A blog about jobs in statistics and data science
- An interactive map of places that employ statisticians in the U.S.